Optimisation et analyse convexe pour la dynamique non-régulière

Florent Cadoux

UNIVERSITÉ JOSEPH FOURIER

THÈSE

pour obtenir le grade de **DOCTEUR DE L'UNIVERSITE JOSEPH FOURIER**

Spécialité : « Mathématiques Appliquées »

préparée au laboratoire Inria Rhône-Alpes, équipe Bipop dans le cadre de **l'École Doctorale « Mathématiques, Sciences et Technologies de l'Information, Informatique »**

préparée et soutenue publiquement par

Florent Cadoux

le 26 Novembre 2009

Titre :

Optimisation et analyse convexe pour la dynamique non-régulière

---

sous la direction de Claude Lemaréchal et Vincent Acary

JURY

Dr Emmanuel Maître	Président
Dr Jaroslav Haslinger	Rapporteur
Dr Jean-Baptiste Hiriart-Urruty	Rapporteur
Dr Patrick Ballard	Examinateur
Dr Claude Lemaréchal	Directeur de thèse
Dr Vincent Acary	Co-directeur de thèse

Remerciements

Je n'aurais sans doute pas survécu à l'alternance de périodes d'euphorie et de désespoir caractéristique de la thèse[1] sans l'aide inestimable de mon directeur de thèse, Claude : j'ai apprécié sa très grande disponibilité et son honnêteté intellectuelle incorruptible, autant que son anti-conformisme, sa culture et son goût pour la langue française. Il m'a convaincu par son exemple que les *recherches* en Mathématiques aboutissent parfois à des *trouvailles*, qu'il arrive que celles-ci recèlent une certaine utilité dans le monde réel, et qu'il est même parfois possible de les exploiter. Mathématicien intransigeant, programmeur autarcique adepte de technologies préhistoriques mais efficaces, linguiste capable de comprendre le langage cryptique des ingénieurs, il est un scientifique complet dont l'oeuvre couvre de façon remarquable tous les niveaux d'abstraction, des sphères idéales de l'analyse convexe au cambouis de la mise en oeuvre industrielle ; puissent les miracles de l'hérédité scientifique me permettre de suivre son exemple. Je souhaite également remercier mon co-directeur de thèse, Vincent, qui m'a initié aux joies de la mécanique du contact, avec qui j'ai eu l'occasion de m'exercer au débat d'idées au cours de nombreuses discussions scientifiques, économiques et politiques, et a assuré la défense centrale lorsque je cafouillais au milieu de terrain lors du foot du mardi midi. Tous deux m'ont laissé une grande liberté de pensée et d'action.

Je suis également redevable envers les deux rapporteurs de ma thèse, Jaroslav Haslinger et Jean-Baptiste Hiriart-Urruty, ainsi que les deux examinateurs, Patrick Ballard et Emmanuel Maître, qui ont accepté de donner de leur temps pour conférer quelque crédibilité à mon travail.

Je dois aussi beaucoup aux membres de l'équipe Bipop, passés et présents : Jérôme – alias Dr J–, mon grand frère de thèse qui a joué avec constance son rôle de mentor, m'a transmis ses théories sur la vie et a supporté stoïquement mes sarcasmes sur les fonctions sous-$C^{1,\alpha}$; Florence, dont le travail sur les tiges à constitué un cas d'application original – et, pour tout dire, inespéré – pour mes algorithmes, et qui a apporté à l'équipe la touche de féminité qui lui faisait cruellement défaut ; Pierre-Brice, même depuis le Japon ; Olivier, Marc et Arnaud, qui m'ont initié à l'escalade avec une patience mal récompensée (je bloque toujours au 5b !), pour leur décontraction communicative ; et Rémou, Fabien et Estelle pour les parties de volley à l'INRIA et au parc Mistral. Je

[1] Pratiquement tous les candidats au doctorat semblent victimes de ce cycle ambition-désillusion-consolation ; lui résister constitue la difficulté principale pour accéder au diplôme, et on pourrait même soutenir que l'obtention du diplôme en question démontre moins les aptitudes scientifiques du lauréat que sa capacité de résistance à la démotivation.

remercie également Constantin, co-bureau parfait par sa discrétion et sa bonne humeur, Franck et ses smart pointers, Maurice et sa version compilable de Siconos, et Roger, sa clé du vestiaire, ses ballons et sa ponctualité légendaire...

Je remercie aussi l'Ecole Polytechnique qui a assuré ma formation antérieure[2] et ma subsistance pendant la thèse en m'octroyant une bourse AMX, l'Ensimag, où j'ai exercé mon monitorat, et ses étudiants dont les progrès m'ont donné l'impression rassurante que je servais à quelque chose dans les moments où je ne pouvais pas décemment chercher ce réconfort dans mon travail de recherche, le CIES pour ses activités distrayantes et souvent intéressantes, le LJK et en particulier Eric Bonnnetier qui a accepté d'être mon tuteur de monitorat malgré ses nombreuses autres fonctions, et enfin l'INRIA pour les conditions de travail exceptionnelles dont j'ai bénéficié au sein de l'équipe Bipop ; je remercie en particulier notre chef d'équipe Bernard, notre assistante Barta et le service de documentation qui m'a plus d'une fois déniché des articles antiques et apparemment introuvables.

Je présente mes excuses à tous ceux à qui j'ai été incapable d'expliquer simplement ce que je faisais pendant ma thèse, et pourquoi on fait de la recherche en maths alors que le théorème de Pythagore a déjà été découvert depuis longtemps. Je pense en particulier à mes parents qui se demandent encore ce qu'est le calcul scolastique (sto-chas-tique !), si un quaternion est une série de quatre coups de poing, et de quel droit les mathématiciens écrivent "différentier" avec un 't' alors que le reste de la population le fait avec un 'c'. J'espère avoir un jour l'occasion de leur démontrer sans mauvaise foi que tout ceci a servi à quelque chose.

Enfin, je remercie mon vélo sur lequel j'ai parcouru environ 1600 fois le trajet Grenoble-Montbonnot (15000 km sans la moindre crevaison !), l'arbre fièrement isolé au sommet de la colline de Venon en face de ma fenêtre, et la bouteille de Coca-Cola en décomposition qui m'a été confiée par Fabien et que j'ai laissé moisir consciencieusement au soleil pendant trois ans. J'en suis déjà le 4ème consignataire ; espérons qu'une nouvelle victime du doctorat se présentera avant mon départ pour que je puisse lui transmettre ce noble flambeau.

[2]Pour être complet, je suis aussi très reconnaissant envers mon professeur de Mathématiques Spéciales au lycée Champollion, Michel Quercia, qui a conforté mon goût pour les mathématiques et m'a initié à leur pratique avec une pédagogie et une rigueur absolument exemplaires.

Résumé / Summary

Version française

L'objectif de ce travail est de proposer une nouvelle approche pour la résolution du problème de contact unilatéral avec frottement de Coulomb tridimensionnel en mécanique des solides. On s'intéresse à des systèmes dynamiques composés de plusieurs corps possédant un nombre fini de degrés de liberté : rigides, ou déformables qui sont des approximations spatiales de modèles continus. Le frottement entre les corps est modélisé en utilisant une formulation classique de la loi de Coulomb. Après discrétisation en temps (ou approximation quasi-statique), on obtient à chaque pas de temps un problème contenant des équations de complémentarité sur un produit de cônes du second ordre, et d'autres équations. Plusieurs méthodes de résolution ont été proposées pour différentes formulations équivalentes de ce problème, en particulier par Moreau, Alart et Curnier, et De Saxcé. En considérant les équations de complémentarité comme celles des conditions d'optimalité (KKT) d'un problème d'optimisation, on propose une reformulation équivalente nouvelle sous forme d'un problème de minimisation paramétrique convexe couplé avec un problème de point fixe. Grâce à ce point de vue, on démontre l'existence de solutions sous une hypothèse assez faible, et vérifiable en pratique. De plus, on peut souvent calculer effectivement l'une de ces solutions en résolvant numériquement l'équation de point fixe. Les performances de cette approche sont comparées à celles des méthodes existantes.

English version

The aim of this work is to propose a new approach to the solution of 3D unilateral contact problems with Coulomb friction in solid mechanics. We consider dynamical systems composed of several bodies with a finite number of degrees of freedom : rigid bodies, or deformable bodies which are spatial approximations of continuous models. Friction between bodies is modelled using a classical formulation of Coulomb's law. After time discretization (or quasi-static approximation), we get at each time step a problem containing complementarity equations posed on a product of second order cones, plus other equations. Several methods have been proposed in the literature for different equivalent formulations of this problem, in particular by Moreau, Alart and Curnier, and De Saxcé. Considering the complementarity equations as optimality conditions (KKT) of an optimization problem, we propose a new equivalent reformulation as a parametric convex

minimization problem coupled with a fixed point problem. Thanks to this viewpoint, we prove the existence of solutions under a quite mild assumption which can be checked in practice. Moreover, we can practically compute one of these solutions by solving numerically the fixed point equation. The performance of this approach is compared with existing methods.

Index des notations et abréviations

Notations

E_Y	module d'Young
σ	tenseur des contraintes
t	variable de temps
d	dimension de l'espace physique ($d = 2$ ou 3)
n	nombre de points de contact
m	nombre de degrés de liberté du système
A, B	objets impliqués dans un contacts
e ou e^i	vecteur unitaire normal orienté de B vers A (au i-ème contact)
x_{N} et x_{T}	composantes normale et tangentielle (par rapport à e) du vecteur $x \in \mathbb{R}^d$
P_{N} et P_{T}	projection sur la droite normale et le plan tangent (par rapport à e)
q	coordonnées généralisées du système ($q \in \mathbb{R}^m$)
v	vitesses généralisées ($v = \dot{q} \in \mathbb{R}^m$)
u ou u^i	vitesses relatives de A par rapport à B (au i-ème point de contact)
r ou r^i	force de contact exercée par B sur A (au i-ème point de contact)
μ ou μ^i	coefficient de frottement (au i-ème point de contact)
K_μ ou K	cône du second ordre de paramètre $\mu \geq 0$ ($K \subset \mathbb{R}^d$) orienté par e ; la notation allégée K ou K^i (au lieu de K_μ ou K_{μ_i}) est utilisée lorsque la valeur de μ est claire d'après le contexte
C ou $C(e, \mu)$	ensemble des couples (u, r) satisfaisant une certaine loi de frottement
$\mathcal{C}(e, \mu)$	ensemble des couples (u, r) compatibles avec la loi de Coulomb
$B(0, \alpha)$	boule euclidienne de centre 0 et de rayon α
$f_{AC,\mathrm{N}}$ et $f_{AC,\mathrm{T}}$	parties normale et tangentielle de la fonction d'Alart et Curnier
ρ_{N} et ρ_{T}	paramètres normal et tangentiel de la fonction d'Alart et Curnier
f_{AC}	fonction d'Alart et Curnier
$s \in \mathbb{R}^n_+$	paramètre des problèmes d'optimisation paramètrique
$T(s)$	produit des n cylindres de frottement de Tresca aux n points de contact
$\mathrm{P}_C(\cdot)$	opérateur de projection sur l'ensemble convexe C
$N_C(x)$	cône normal à l'ensemble convexe C au point x
$\widetilde{u}$	vitesse modifiée associée à u par le changement de variable de De Saxcé
K°	cône polaire d'un cône convexe fermé K
K^*	cône dual (opposé du cône polaire) d'un cône convexe fermé K

∂	(1) frontière d'un ensemble ($\partial S := \bar{S} \setminus \mathrm{int}(S)$)
	(2) sous-différentiel d'une fonction convexe
H, w	données de l'équation cinématique $u = Hv + w$
M, f	données de l'équation dynamique $Mv + f = H^\top r$
$\mathbf{S}_m^{++}$	cône des matrices symétriques définies positives
W	opérateur de Delassus ($W = HM^{-1}H^\top$)
q	terme constant ($q = w - HM^{-1}f$) dans l'équation $u = Wr + q$
E_c	énergie cinétique
Jac[$\cdot$]	matrice jacobienne
Diag(x_1, x_n)	matrice diag. dont les éléments diag. sont $x_1, \ldots, x_n$
E	matrice définie par $E := \mathrm{Diag}(\mu^i e^i)$
L	produit cartésien des n cônes de frottement K_i
J	fonction « énergie » primale
$C(s)$ et $\bar{C}(s)$	ensemble réalisable strict (resp. large) du problème primal
$q(s)$ et $p(s)$	valeur du problème primal (resp. dual)
$F(\cdot)$	fonction de point fixe, dans l'approche proposée
ri	intérieur relatif (d'un ensemble convexe)
dom	domaine d'une fonction
$J^*(\cdot)$	conjuguée de la fonction convexe J
i_C	fonction indicatrice de l'ensemble (convexe) C
$\nabla[\cdot]$	gradient
(H)	hypothèse faible du résultat d'existence
$(\bar{H})$	hypothèse forte du résultat d'existence
e_H	fonction *excès* de Hausdorff
Δ_H	distance de Hausdorff
Λ	ensemble de sous-niveau de J
Δ	vecteur de taille n défini par $\Delta := (\delta, \ldots, \delta)$
I ou I_k	matrice identité (de taille k)
$0_{m\times n}$	matrice nulle de taille $m \times n$
f_{DS}	fonction de De Saxcé
Φ	notation générique pour la fonction que l'on cherche à annuler
$H[\cdot]$	matrice Hessienne
C	matrice telle que $CC^\top = M$ (ex : décomposition de Choleski)
X	vecteur qui rassemble les inconnues $(v, \widetilde{u}, r)$
ρ_{DS}	paramètre de la fonction de De Saxcé

Abréviations

LCP	*linear complementarity problem*
QP	*quadratic programming*
LP	*linear programming*
SOCP	*second order cone programming*
SOQP	*second order quadratic programming*
SeDuMi	solveur SOCP disponible publiquement
AC	Alart et Curnier
DS	De Saxcé
GS	Gauss-Seidel
FB	Fisher-Burmeister
H-pf	méthode de Haslinger + point fixe
NA-pf	« notre approche » + itérations de point fixe
NA-Newton	« notre approche » + méthode de Newton
NA-Broyden	« notre approche » + méthode de Broyden (quasi-Newton)

Table des matières

Introduction

Cette thèse traite des rapports entre l'optimisation non-régulière et la dynamique non-régulière à travers deux exemples : le contact frottant de Coulomb en mécanique et les méthodes de coupes en optimisation combinatoire.

0.1 Simulation en mécanique du contact

Dans de nombreux domaines scientifiques, et en particulier en mécanique, la simulation est devenue un outil aussi important et étudié que la théorie et l'expérience pour comprendre les phénomènes physiques. La confrontation souvent réussie entre les résultats d'expériences réelles et celles réalisées *in silico* ont donné suffisamment confiance aux chercheurs et aux ingénieurs pour que la simulation soit largement utilisée dans le milieu industriel, donnant naissance à de nombreux codes de calcul scientifique. Ces logiciels utilisent des technologies variées (méthode des différences finies, éléments finis, volumes finis, méthodes spectrales... pour ne citer que les plus célèbres) et permettent de reproduire la statique et la dynamique de nombreux systèmes mécaniques comportant des fluides et des solides rigides ou déformables, pour une large gamme de lois de comportement. Cependant, il subsiste bien des phénomènes physiques difficiles à modéliser ou, même lorsqu'un modèle acceptable est disponible, à simuler. La mécanique du contact, dont les aspects numériques sont l'objet de cette thèse, en fait partie : il s'agit de simuler la dynamique d'un ensemble de solides susceptibles de se toucher, rebondir les uns sur les autres, glisser les uns à la surface des autres, etc. Il est donc nécessaire de faire des hypothèses physiques pour modéliser les impacts et le frottement. Plus la description retenue est fine, plus les calculs à effectuer sont lourds et plus on est limité dans la taille des systèmes physiques que l'on peut simuler. Il faut donc faire un compromis, qui dépend de l'application visée, entre le réalisme de la simulation (que l'on peut améliorer en choisissant un modèle plus fin, des schémas d'intégration en temps d'ordre plus élevé, des pas de temps plus petits, une arithmétique plus précise dans les calculs sur ordinateur, etc) et le coût de la simulation (il ne faut pas arriver à des équations que l'on ne sait pas du tout résoudre, ou dont le temps de résolution est trop important pour l'application voulue).

0.1.1 Domaine de validité et applications

Dans la première partie cette thèse, nous étudions la simulation numérique d'un système de solides rigides ou déformables dont les impacts sont supposés inélastiques ou régis par le modèle de Moreau [Mor88], et le frottement par le modèle de Coulomb (proposé vers 1780 par Charles Augustin Coulomb, 1736-1806). Ceci restreint la généralité des travaux aux situations où ces modèles sont suffisamment fins pour capturer les phénomènes physiques que l'on désire étudier, et suffisamment grossiers pour que les calculs prennent un temps raisonnable. L'expérience de Moreau [Mor88, Mor94, Mor03, Mor06] et d'autres mécaniciens suggère cependant que la mécanique des matériaux granulaires, de la maçonnerie non-cohésive, et celle des robots par exemple, font partie des domaines d'application possibles du modèle étudié. On peut utiliser ce modèle pour prédire ou contrôler le mouvement, étudier les positions d'équilibre, et même pour extraire des informations inaccessibles à l'expérience, comme la répartition des forces dans un tas de pierres ou de sable. Dans d'autres cas, le modèle proposé est moins bien adapté : par exemple, la communauté de l'informatique graphique est également très active dans le domaine de la simulation du contact unilatéral avec frottement, mais ses exigences sont relativement différentes. Pour une application donnée, comme par exemple un jeu vidéo, on peut considérer que la stabilité et la vitesse d'exécution de l'algorithme sont des critères plus importants que l'adéquation à un modèle physique.

0.1.2 Non-régularité

Le problème peut être qualifié de non-régulier (ou non-lisse) pour différentes raisons, ce qui le rend plus difficile, à la fois en théorie et en pratique [Bro99], que la mécanique lagrangienne classique où le mouvement n'est contraint que par des liaisons bilatérales sans frottement. Quand le contact unilatéral est pris en compte, des collisions ont lieu durant lesquelles le mouvement peut changer très rapidement. Si les corps sont déformables, on peut envisager de calculer finement les déformations au niveau du point de contact et se passer d'une loi d'impact. Mais cette approche nécessite au minimum de raffiner beaucoup la discrétisation temporelle, et n'est pas envisageable pour de gros systèmes, ni quand les corps sont modélisés par des solides rigides. Pour cette raison, lorsqu'on étudie le mouvement continu, on modélise souvent les chocs par des sauts *instantanés* de la vitesse. Par conséquent, la notion d'accélération doit être généralisée et étudiée dans des espaces de distributions, et non pas de fonctions. Ceci peut-être qualifié de non-régularité temporelle. Le fait que la force de frottement et la vitesse du point de contact soient reliées non pas par une fonction régulière, mais par une multifonction (loi de Coulomb) est appelée la non-régularité en loi.

Dans ce travail, on se concentre sur les aspects numériques et la simulation. De ce point de vue, nous adopterons les méthodes suivantes pour gérer les différentes sources de non-régularité.

– Non régularité spatiale : les contraintes unilatérales (deux objets ne peuvent pas s'interpénétrer) sont modélisées directement dans la loi de contact, au niveau des vitesses : au niveau d'un point de contact entre deux objets, la vitesse relative doit

appartenir à un demi-espace imposé.
- Non-régularité en loi : les valeurs des forces et des vitesses à chaque pas de temps sont calculées, conformément à la loi de Coulomb, par des algorithmes variés dont l'étude est l'objet principal de la thèse.
- La non-régularité temporelle – impacts, sauts dans les vitesses – est prise en compte par un algorithme de type *time stepping* (ce qui signifie que la durée d'un pas de temps est choisie à l'avance). Contrairement aux algorithmes dits *event driven* (où l'on ajuste la durée des pas de temps), ceux-ci ne s'arrêtent pas à chaque évènement. Autrement dit, il est possible avec ces méthodes de traiter plusieurs collisions en un seul pas de temps.

0.1.3 Exemple introductif

L'exemple simple de mécanique statique qui suit permet d'illustrer la notion de non-régularité spatiale (mais pas la non-régularité en loi, ni temporelle), d'exhiber le rapport étroit entre la mécanique du contact et l'optimisation sous contraintes, et de montrer la différence entre les techniques qui traitent les contraintes de manière exacte, comme celles que nous allons utiliser, et les techniques approximatives qui autorisent un certain degré de violation, comme la méthode de pénalisation. Le terme « exact » dans la phrase précédente est abusif, puisqu'un certain degré de pénétration est inévitable (même avec les techniques dites exactes), mais il est couramment employé dans ce cas. On parle aussi de méthodes « par contraintes », par opposition aux méthodes par pénalisation ou par impulsions [MC95]. Considérons donc le problème statique suivant en dimension 1. Une masse m est suspendue à un ressort de raideur k, dans le champ de gravitation g comme sur la figure 1. La masse ne peut pas descendre en-dessous du sol, éloigné de l'ancrage du ressort d'une distance h.

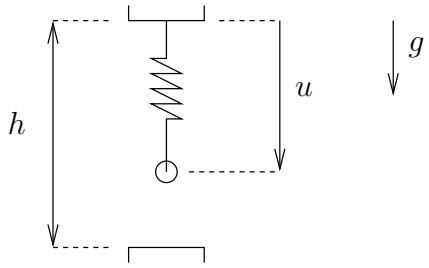


FIG. 1 – Système masse-ressort avec contrainte unilatérale

En l'absence de contrainte, la solution est $u = \frac{mg}{k}$. Supposons donc que $h \leq \frac{mg}{k}$. On cherche les points d'équilibre statique, c'est-à-dire les minimums de l'énergie potentielle $\Pi(u)$ suivante

$$\Pi(u) = \frac{1}{2}ku^2 - mgu$$

sous la contrainte unilatérale $c(u) := u - h \leq 0$. Autrement dit, il s'agit de résoudre un problème d'optimisation quadratique convexe sous contrainte d'inégalité linéaire. Ceci

est un premier lien entre la mécanique et l'optimisation. Il existe des manières exactes et approchées pour résoudre ce problème. L'approche proposée par Lagrange consiste à introduire un *multiplicateur* λ pour chaque contrainte, et à résoudre, si possible, le problème suivant :

$$\Pi'(u) + \lambda c'(u) = 0 \tag{1a}$$
$$\lambda c(u) = 0 \tag{1b}$$

sous les contraintes $\lambda \geq 0$ et $c(u) \leq 0$. Elle fournit les solutions exactes du problème statique. Ici on peut résoudre ce problème facilement : l'équation (1b) impose que soit λ, soit $c(u)$ soit nul. Si $\lambda = 0$, on retrouve $u = \frac{mg}{k}$ qui viole la contrainte $c(u) \leq 0$ et n'est donc pas solution. Si $c(u) = 0$, alors par définition de $c(\cdot)$, on a $u = h$. Alors l'équation (1a) donne $\lambda = mg - kh \geq 0$, donc la contrainte $\lambda \geq 0$ est satisfaite et on constate que ce problème possède une solution unique $u = h, \lambda = mg - kh$. On peut aussi voir cette méthode comme la recherche d'un point-selle du *lagrangien* suivant

$$L(u, \lambda) := \Pi(u) + \lambda\, c(u).$$

Illustrons maintenant l'approche par pénalisation. Pour simplifier, nous nous limitons à une contrainte bilatérale $c(u) = 0$. Cette approche consiste à modifier la fonction énergie potentielle de la manière suivante (ou d'une autre, il existe de nombreuses fonctions de pénalisation)

$$\tilde{\Pi}(u) := \Pi(u) + \frac{\mu}{2} c(u)^2$$

et à minimiser $\tilde{\Pi}$. Physiquement, cette modification revient à ajouter un ressort de raideur μ qui relie la masse au sol. On peut résoudre le problème de minimisation, on trouve $u = \frac{mg+\mu h}{k+\mu}$. Si $\mu = 0$, on retrouve évidemment la solution sans contrainte (puisque $\tilde{\Pi} = \Pi$), et dans la limite $\mu \longrightarrow +\infty$ on retrouve la solution exacte $u = h$. En pratique, on utilise une valeur finie pour μ (le choix de sa valeur n'étant pas du tout évident) et la fonction $c(\cdot)$ prend la valeur

$$c(u) = c(\frac{mg + \mu h}{k + \mu}) = \frac{mg - hk}{k + \mu} < 0$$

donc la contrainte est violée : quand on utilise la méthode de pénalisation, les contraintes unilatérales peuvent être (légèrement, si on s'y prend bien) violées. On peut aussi combiner la méthode de Lagrange et celle par pénalisation, on obtient alors la technique dite du « lagrangien augmenté » qui consiste à chercher un point-selle de la fonction

$$L_{\text{aug}}(u, \lambda) = \Pi(u) + \frac{\mu}{2} c(u)^2 + \lambda\, c(u).$$

Ces méthodes (multiplicateurs de Lagrange, pénalisation, ou lagrangien augmenté) sont très différentes du point de vue théorique et numérique. La méthode des multiplicateurs de Lagrange présente l'avantage d'être exacte et l'inconvénient de nécessiter l'introduction de variables supplémentaires, tandis que la méthode par pénalisation a

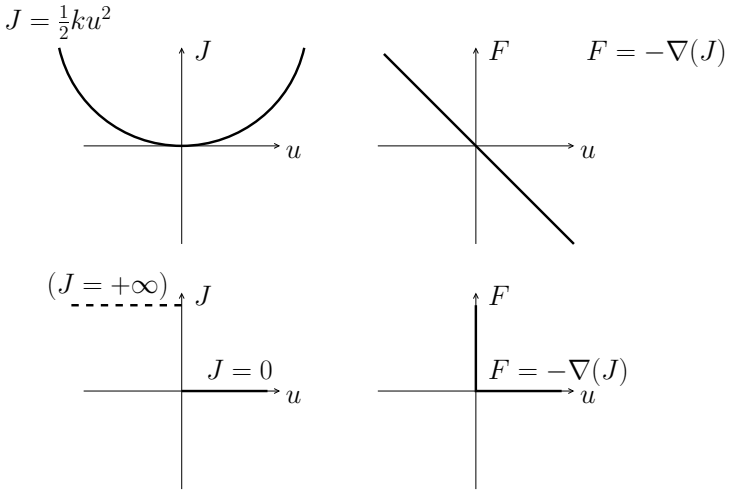


FIG. 2 – Analogie entre le potentiel associé à un ressort et celui associé à un mur

l'avantage de ne pas requérir de variable supplémentaire, mais n'est qu'une méthode approchée ; elle nécessite un ajustement délicat du paramètre de pénalisation μ, et provoque dans les simulations dynamiques des oscillations parasites (non physiques) dues à l'introduction par les « ressorts » de pénalisation de modes de vibration rapides et artificiels. Lorsque des multiplicateurs sont utilisés, on parle parfois de méthodes *par contraintes*, par opposition aux méthodes de pénalisation. Dans cette thèse on s'intéresse uniquement aux méthodes par contraintes.

Les méthodes par pénalisation diffèrent des méthodes par contraintes par la régularisation qu'elles entraînent sur le problème. Par conséquent, les outils mathématiques utiles à l'une ou l'autre approche ne seront pas les mêmes. Tandis que l'approche par pénalisation ramène le problème dans le domaine de la mécanique régulière, pour laquelle les outils de l'analyse classique suffisent, la méthode des multiplicateurs de Lagrange conserve au problème mécanique son caractère non-régulier et il devient nécessaire d'utiliser les outils de l'analyse non-lisse, en particulier de l'analyse convexe. Ceci peut être illustré par la généralisation de la notion de potentiel en mécanique classique. Puisque la méthode de pénalisation consiste essentiellement à remplacer une contrainte d'impénétrabilité par une force très répulsive qui agit comme un ressort contre la pénétration, poursuivons cette analogie : de la même manière que la force $-ku$ exercée par un ressort linéaire de raideur k en réponse à une extension u dérive du potentiel $\frac{1}{2}ku^2$, il est tentant de décrire la force exercée par un mur impénétrable comme dérivant d'un potentiel. Ceci motive la notion de *potentiel convexe* et de *sous-différentiel*, des notions qui généralisent la notion de potentiel et de gradient, comme l'illustre la figure 2.

FIG. 3 – Léonard de Vinci

0.2 Note historique

L'étude expérimentale du phénomène de frottement, ses interprétations microscopiques, et la formulation des lois et des modèles ont été réalisées essentiellement par Léonard de Vinci, Guillaume Amontons, Charles-Augustin Coulomb et Leonhard Euler.

Léonard de Vinci (1452-1519, figure 3) fut le premier à rapporter des expériences quantitatives sur le frottement. Il réalisa des expériences de frottement statique sur un plan incliné, dont il faisait varier l'angle, et affirma que pour un objet donné, la force de frottement était indépendante de la surface de contact. Il constata la proportionnalité entre le poids de l'objet et la force de frottement, et mesura pour ce coefficient une valeur de 0.25 qu'il croyait apparemment universelle (indépendante des matériaux en contact). Guillaume Amontons (1663-1705) répéta l'expérience sur un plan horizontal ; il observa le phénomène de frottement dynamique en exerçant une force tangentielle avec un ressort (figure 4, où le ressort D mesure la force de frottement pendant le glissement relatif de A et B, et le ressort C permet de faire varier la force normale). Il affirma, comme Léonard de Vinci, que la force de frottement était indépendante de la surface de contact, et ajouta qu'elle était proportionelle à la force normale. Il appela le coefficient de proportionnalité « coefficient de frottement » . Amontons étudia aussi la lubrification, observant que le coefficient de frottement était le même pour toute paire de matériaux (il utilisait du fer, du plomb, du cuivre et du bois) si l'interface était préalablement enduite de graisse, avançant la valeur 0.3 pour le coefficient en question.

C'est Leonhard Euler (1707-1783, figure 5) qui affirma que les coefficients de frottement statique et dynamique étaient en général différents. Il reprit les expériences de Léonard de Vinci, et proposa une interprétation microscopique du phénomène de frottement en considérant que des aspérités triangulaires en étaient la cause. Enfin, Coulomb (1736-1806, figure 6) réalisa un dispositif expérimental (figure 7) capable de mesurer le coefficient de frottement dynamique à différentes vitesses. Il observa que la force de

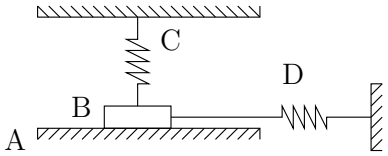


FIG. 4 – Expérience d'Amontons

FIG. 5 – Leonhard Euler

FIG. 6 – Charles-Augustin Coulomb

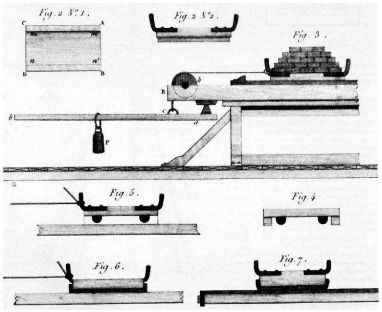


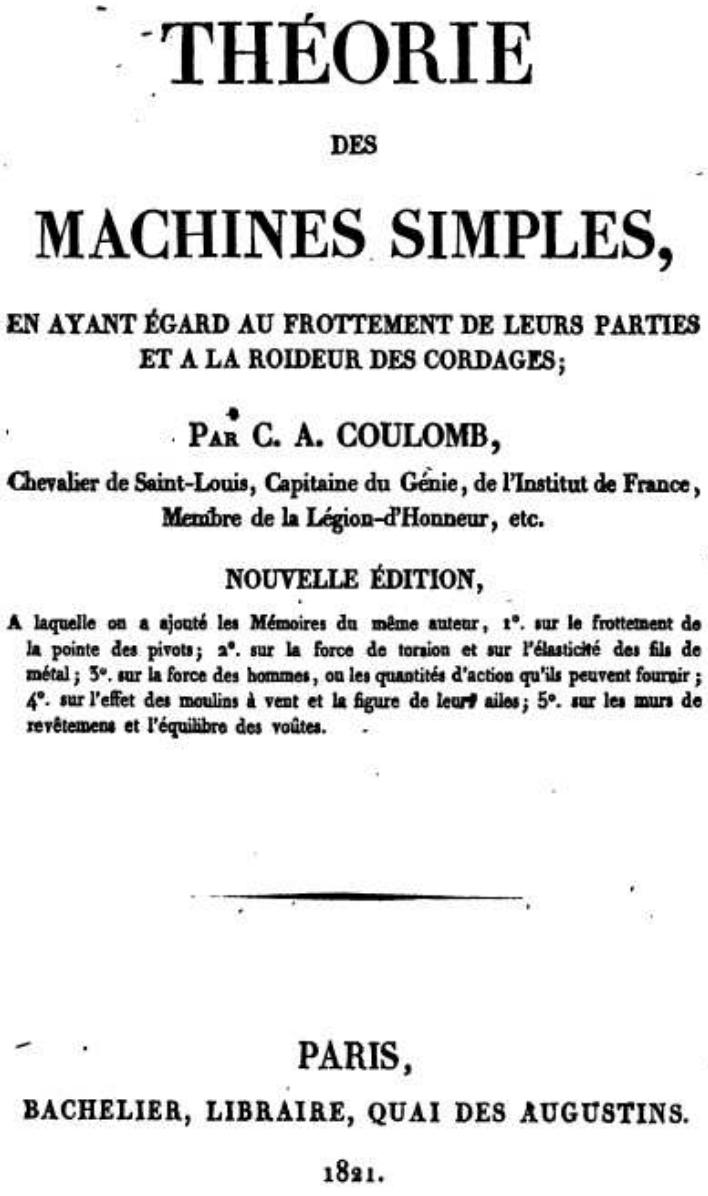
FIG. 7 – Expérience de Coulomb

frottement était indépendante de la vitesse de glissement. Il s'intéressa aussi à la variation du coefficient de frottement statique en fonction du temps de repos entre les deux surfaces en contact, constatant qu'il augmentait avec la durée du repos. Dans son ouvrage « Théorie des machines simples, en ayant égard au frottement de leurs parties et à la roideur des cordages » (figure 8), ses idées sont appliquées à la mécanique des machines[3]. Voici le début de l'introduction de cet ouvrage, où l'on voit que Coulomb distinguait nettement entre les coefficients de frottement statique et dynamique. L'exemple final du vaisseau lancé à l'eau sur un plan incliné prouve aussi qu'il avait conscience du caractère approximatif et des limites de validité de sa théorie du frottement, observant que le coefficient de frottement (considéré comme constant par la théorie) pouvait en réalité varier de 1/3 à moins de 1/14 pour une paire de matériaux donnée, en fonction de l'échelle de l'expérience.

« M. Amontons, dans les Mémoires de l'Académie des Sciences pour 1699, paraît être le premier auteur qui ait cherché à évaluer le frottement et la roideur des cordes dans le calcul des machines. Il crut trouver, par ses expériences, que l'étendue des surfaces n'entrait pour rien dans les frottements, dont la mesure dépendait uniquement de la pression des parties en contact : il en conclut que, dans tous les cas, le frottement est proportionnel aux pressions.

La plupart des mécaniciens suivirent les résultats de M. Amontons ; cependant M. Muschembreek trouva, dans plusieurs expériences, que les frottements ne dépendaient pas uniquement de la pression, et que l'étendue des surfaces y influait. MM. de Camus, dans son Traité des forces mouvantes, et Désaguilliers, dans son Cours de Physique, s'aperçurent que le frottement d'un corps ébranlé était moins considérable que celui d'un corps que l'on voulait sortir de l'état de repos : mais ni l'un ni l'autre ne cherchèrent à déterminer le rapport qui pouvait exister entre ces deux espèces de frottement. M. l'abbé Bossut, dans son excellent Traité de Mécanique, penche pour le système de M. Amontons, qui donne une plus grande facilité dans les calculs, et qui suffit dans la plupart des cas de pratique ; pourvu que l'on ait soin de distinguer le frottement dans les surfaces en mouvement, d'avec la force qu'il faut employer pour détacher ces mêmes surfaces après un certain temps de repos. L'on voit, de plus, par les réflexions qui précèdent le calcul

[3]Ce livre, tombé dans le domaine public, a été numérisé et est disponible sur internet gratuitement.



THÉORIE

DES

MACHINES SIMPLES,

EN AYANT ÉGARD AU FROTTEMENT DE LEURS PARTIES
ET A LA ROIDEUR DES CORDAGES;

PAR C. A. COULOMB,

Chevalier de Saint-Louis, Capitaine du Génie, de l'Institut de France,
Membre de la Légion-d'Honneur, etc.

NOUVELLE ÉDITION,

A laquelle on a ajouté les Mémoires du même auteur, 1°. sur le frottement de la pointe des pivots; 2°. sur la force de torsion et sur l'élasticité des fils de métal; 3°. sur la force des hommes, ou les quantités d'action qu'ils peuvent fournir; 4°. sur l'effet des moulins à vent et la figure de leurs ailes; 5°. sur les murs de revêtemens et l'équilibre des voûtes.

PARIS,

BACHELIER, LIBRAIRE, QUAI DES AUGUSTINS.

1821.

FIG. 8 – Couverture de la réédition d'un ouvrage de Coulomb

du frottement des machines dans la mécanique de M. l'abbé Bossut, que ce célèbre auteur a prévu, comme l'on pourra s'en convaincre par nos expériences, ce qui arriverait relativement à l'étendue des surfaces, aux pressions et aux vitesses dans les expériences qui restaient à faire.

Des essais faits en petit, dans un cabinet de physique, ne peuvent pas suffire pour nous diriger dans le calcul des machines destinées à soulever plusieurs milliers [sic] ; parce que la moindre inégalité, le plus faible obstacle placé entre les surfaces, la cohérence de quelques parties plus ou moins homogènes, jettent la plus grande irrégularité dans les résultats. L'on exécute tous les jours dans nos ports une manoeuvre qui montre combien peuvent être fautives des conclusions sur le frottement, tirées des expériences faites en petit ; c'est celle de lancer les vaisseaux à l'eau, sur un plan incliné de 10 ou 12 lignes par pied, ce qui indique que le frottement n'est pas, dans cette opération, le quatorzième de la pression ; tandis qu'en faisant glisser, sous de petites pressions, un madrier de chêne sur un autre madrier du même bois, on avait cru qu'il était le tiers de la pression. »

Coulomb avait donc compris que de manière générale, les phénomènes à l'interface – qui font l'objet de la science appelée aujourd'hui tribologie – sont complexes, et que sa loi de frottement à un paramètre ne pouvait être utilisée que dans un certain domaine de validité qui dépend entre autre de l'échelle de l'expérience.

0.3 Problème de séparation et méthodes de coupes

La seconde partie de la thèse est consacrée à une étude plus théorique du problème de séparation en analyse convexe. Il consiste à déterminer un hyperplan qui divise l'espace en deux demi-espaces contenant respectivement un point et un ensemble convexe donnés à l'avance (le point n'étant pas contenu dans le convexe). Ce problème apparaît en optimisation combinatoire lorsqu'on veut séparer une solution approximative, et non réalisable, d'un problème, de l'ensemble des points réalisables. La solution approchée est typiquement obtenue par relaxation, comme dans le cas de la résolution d'un programme linéaire en nombres entiers à l'aide d'un algorithme de type « séparer et couper » (*branch and cut* en anglais) dans lequel les solutions approchées sont obtenues en résolvant la relaxation linéaire du programme en nombres entiers. En général, il existe une infinité d'hyperplans séparateurs (ou coupes) possibles, et le problème est de générer les bons, en un certain sens. En pratique, les bonnes coupes sont celles qui réalisent un compromis favorable entre le temps de calcul nécessaire pour les obtenir, et leur qualité que l'on mesure de différentes manières (capacité à améliorer la valeur du problème relaxé, caractère creux, efficacité sur une bibliothèque de tests après intégration dans un algorithme de branch-and-cut, etc). On propose une méthode pour générer de bonnes coupes au sens suivant : profondes, au sens de la distance euclidienne, et qui exposent des facettes de l'ensemble convexe dont on cherche à séparer un point (lorsque cet ensemble est polyédral). Les outils utilisés ici sont ceux de l'analyse convexe et de l'optimisation continue. Ces méthodes ne s'appliquent pas seulement au cas typique de la programmation entière, elles peuvent aussi être utilisées pour résoudre les problèmes de complémentarité linéaires (LCP) tels que ceux qui apparaissent dans le cas du problème du frottement bidimen-

sionnel et plus généralement dans d'autres problèmes de dynamique non-régulière. Par exemple, la situation de la figure 1 illustre bien le caractère combinatoire des problèmes de contact unilatéral : soit $\lambda = 0$, soit $u = h$. Cette situation de *complémentarité* est en fait assez générale, et apparaît également dans d'autres systèmes dynamiques comme ceux qui gouvernent l'évolution de certains circuits électroniques [AB08].

Organisation des chapitres

Au premier chapitre, intitulé « Modèles d'impact et de frottement », on introduit la loi de Coulomb qui est utilisée dans toute la suite. Le second chapitre (« Dynamique non-régulière ») rappelle les méthodes de Lagrange et d'Euler pour établir des équations du mouvement pour un système mécanique régulier (sans contact ni frottement) ; on intègre ensuite le modèle de contact et frottement à ces équations, puis on les discrétise pour aboutir au problème élémentaire (dit *problème incrémental*) dont la résolution permet de passer d'un pas de temps au suivant. Après ces deux chapitres d'inspiration plutôt mécanique, on consacre le troisième (« Résultat d'existence ») à une étude mathématique du problème incrémental qui aboutit à un résultat d'existence de solution. Cette démonstration d'existence nouvelle suggère de plus une méthode pratique pour résoudre le problème incrémental en résolvant une certaine équation de point fixe. Le quatrième chapitre (« Résolution pratique du problème incrémental ») est donc essentiellement algorithmique, on y passe en revue les méthodes disponibles pour résoudre le problème élémentaire. En particulier, une nouvelle méthode de résolution – qui s'appuie sur les développements théoriques du chapitre 3 – est décrite. Le cinquième chapitre (« Expériences numériques ») contient les résultats des tests effectués avec les différents algorithmes, et compare l'approche proposée aux méthodes existantes.

La seconde partie, qui aborde l'optimisation combinatoire et l'analyse convexe, commence au chapitre 6 (« Problème de séparation ») par la description du problème de génération de coupes. Enfin dans le chapitre suivant (« Méthode par décomposition en facettes »), on présente une nouvelle approche pour attaquer ce problème, qui utilise des notions d'analyse convexe comme le polaire et le polaire inverse, et des notions d'optimisation continue comme un algorithme dû à Wolfe.

Première partie

Mécanique du contact

Chapitre 1

Modèles d'impact et de frottement

Introduction

La première étape pour simuler le comportement d'un système mécanique est le choix d'un modèle adapté à la nature et à l'échelle des phénomènes physiques que l'on cherche à reproduire. Pour de nombreuses applications, la loi de Coulomb est un modèle de frottement acceptable : après avoir présenté quelques-unes des lois tribologiques qui lui font concurrence, nous l'adopterons pour toute la suite de la thèse. Différentes formulations équivalentes de cette loi seront aussi présentées.

En ce qui concerne les impacts, dont il est difficile de se passer lorsqu'on simule des corps rigides, le problème est plus délicat. On doit choisir une loi d'impact : quelles seront les vitesses après le choc, connaissant les vitesses avant l'évènement ? La section 1.1 montre que la réponse dépend beaucoup de la nature des objets considérés, en particulier de leur géométrie, et qu'il n'y a pas de loi simple et possédant un large domaine de validité. Pour cette raison, nous adopterons l'approche pragmatique suivante : tous les chocs seront supposés parfaitement inélastiques. Du point de vue numérique, ceci ne restreint pas trop la généralité du propos : on peut toujours introduire facilement une loi d'impact de Newton à un paramètre (en utilisant la règle de Moreau) ou imposer au début de chaque pas de temps des impacts calculés d'une manière ou d'une autre avant de reprendre le cours normal de la simulation.

1.1 Modèles d'impact

Dans cette sous-section, on étudie l'impact de deux barres déformables en une dimension d'espace, afin de comparer les prédictions du modèle élastique linéaire et celles du modèle des solides indéformables (considéré comme plus grossier). Cet exemple est adapté de [Wri06].

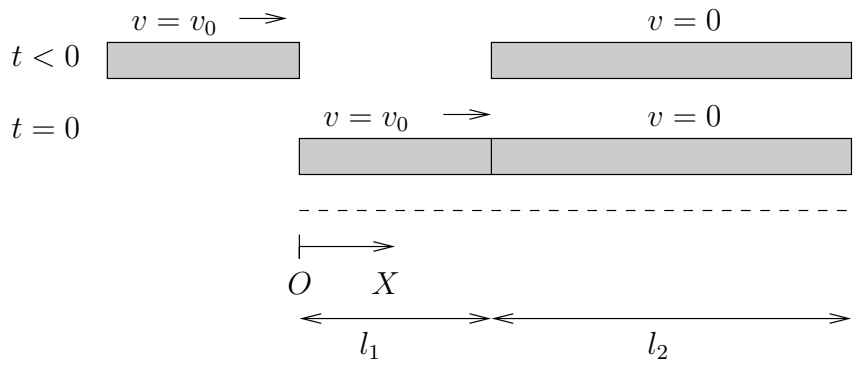


FIG. 1.1 – Deux barres élastiques entrent en collision

1.1.1 Modèle élastique linéaire

Les deux barres sont représentées sur la figure 1.1. La barre de gauche, de longueur l_1, se déplace à vitesse v_0 vers la barre de droite, de longueur l_2, immobile. A l'instant initial $t = 0$, les barres entrent en collision. On s'intéresse au mouvement ultérieur, jusqu'à leur séparation au temps T. Les barres ont le même module d'Young E_Y. On note $l := l_1 + l_2$ la longueur totale, et $X \in [0, l]$ la coordonnée d'espace dans la configuration de référence (c'est-à-dire à $t = 0$), comme sur la figure 1.1. On note $x(X)$ la position actuelle et $u(X) := x(X) - X$ le déplacement. Le tenseur des contraintes est $\sigma = E_Y \frac{\partial u}{\partial X}$. L'équation du mouvement est

$$E_Y \frac{\partial^2 u}{\partial x^2} = \rho \frac{\partial^2 u}{\partial t^2}$$

c'est-à-dire l'équation des ondes avec une vitesse de propagation $c := \sqrt{\frac{E_Y}{\rho}}$. Ses solutions sont de la forme

$$u(X, t) = f(X - ct) + g(X + ct)$$

où les fonctions f et g sont à déterminer en fonction des conditions initiales et au bord. Ici, ces conditions sont les suivantes.

$$\forall X \in [0, l] \qquad u(X, 0) = 0 \tag{1.1}$$

$$\forall X \in [0, l_1] \qquad \frac{\partial}{\partial t} u(X, 0) = v_0 \tag{1.2}$$

$$\forall X \in [l_1, l] \qquad \frac{\partial}{\partial t} u(X, 0) = 0 \tag{1.3}$$

$$\forall t \in [0, T] \qquad \sigma(0, t) = 0 \tag{1.4}$$

$$\forall t \in [0, T] \qquad \sigma(l, t) = 0 \tag{1.5}$$

$$\forall t \in [0, T] \qquad \sigma(l_1, t) \leq 0. \tag{1.6}$$

L'équation (1.1) impose un déplacement nul à $t = 0$, c'est-à-dire que les barres ne sont initialement pas déformées. L'équation (1.2) exprime que la barre de gauche a une vitesse initiale v_0, et l'équation (1.3) que la barre de droite a une vitesse initiale nulle.

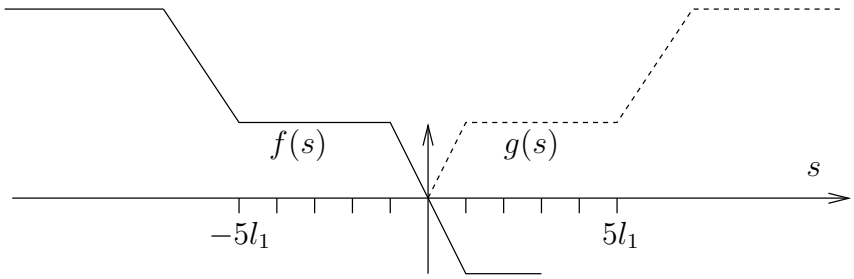


FIG. 1.2 – Les fonctions f et g

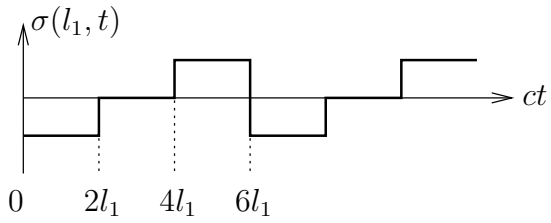


FIG. 1.3 – Contrainte $\sigma(l_1,t)$ au point de contact des deux barres

Les équations (1.4) et (1.5) imposent l'absence de contraintes sur les deux extrémités libres. Enfin l'équation (1.6) est la condition de contact, qui dit que la contrainte en $X = l_1$ doit être une contrainte de compression. Dès que cette condition est violée, les barres se séparent.

On peut exprimer ces six contraintes en termes des fonctions f et g. Il vient

$$\forall X \in [0,l] \qquad f(X) + g(X) = 0 \tag{1.7}$$
$$\forall X \in [0,l_1] \qquad c(-f'(X) + g'(X)) = v_0 \tag{1.8}$$
$$\forall X \in [l_1,l] \qquad c(-f'(X) + g'(X)) = 0 \tag{1.9}$$
$$\forall t \in [0,T] \qquad f'(-ct) + g'(+ct) = 0 \tag{1.10}$$
$$\forall t \in [0,T] \qquad f'(l-ct) + g'(l+ct) = 0 \tag{1.11}$$
$$\forall t \in [0,T] \qquad f'(l_1-ct) + g'(l_1+ct) \leq 0 \tag{1.12}$$

Si un couple de fonctions (f,g) est solution, alors pour tout $\alpha \in \mathbb{R}$, le couple $(f+\alpha, g-\alpha)$ est également solution. Les fonctions f et g sont donc définies à une constante près, et on peut supposer que $f(0) = 0$. Sous cette hypothèse, on voit facilement que les fonctions f et g sont nécessairement celles représentées sur la figure 1.2. Elles sont continues, affines par morceaux, avec une pente alternativement de 0 et $\frac{-v_0}{2c}$ ou $\frac{v_0}{2c}$. On a tracé les fonctions f et g pour $l_2 = 2l_1$ sur la figure 1.2, et sur la figure 1.3 la contrainte $\sigma(l_1,t)$ au point l_1 donnée par la formule

$$\sigma(l_1,t) = f'(l_1 - ct) + g'(l_1 + ct).$$

On voit sur la figure 1.3 que σ devient strictement positif en l_1 au temps $t = \frac{4l_1}{c}$, qui est donc le temps T de fin d'impact (à cet instant, les deux barres se séparent). Quelques

Matériau	Module d'Young E_Y (MPa)	Densité ρ (kg/m3)	Vitesse prop. c (m/s)	Temps impact T (s)
acier	210000	8	162000	$1e^{-5}$
béton	30000	2.5	110000	$2e^{-5}$
bois de chêne	12000	0.9	115000	$2e^{-5}$

TAB. 1.1 – Quelques ordres de grandeur sur le phénomène d'impact

ordres de grandeurs pour c et T (calculé pour $l_1 = 1$, en mètres) sont donnés dans le tableau 1.1. On constate que les vitesses de propagation sont de l'ordre de 10^5 mètres par seconde, ce qui signifie que pour une taille caractéristique de l'ordre du mètre, la simulation de ce système nécessite des pas de temps significativement inférieurs à 10^{-5} seconde. Ceci signifie aussi que dans les simulations de modèles rigides, le temps d'impact peut être négligé lorsque le temps caractéristique du reste de la dynamique est significativement supérieur à 10^{-5} seconde.

Le déplacement u dans les deux barres est représenté en $t = 0, \frac{l_1}{c}, \ldots, \frac{5l_1}{c}$ (même si la séparation intervient à $t = \frac{4l_1}{c}$) sur la figure 1.4, toujours dans le cas où $l_2 = 2l_1$. Le tableau 1.2 montre que juste après l'impact, la barre 1 est immobile et la barre 2 se déplace à vitesse uniforme $\frac{v_0}{2}$. On peut aussi calculer le déplacement du point de contact au cours de l'impact, qui vaut

$$u(l_1, 4\frac{l_1}{c}) = f(-3l_1) + g(5l_1) = \frac{v_0}{c} l_1.$$

Si on prend de nouveau l_1 de l'ordre du mètre, et v_0 de l'ordre du mètre par seconde, on trouve un déplacement du point d'impact de l'ordre de 10^{-5} mètre. Dans les simulations, il est donc souvent légitime de négliger le déplacement dû aux impacts.

X	$0 \to l_1$	$l_1 \to 2l_1$	$l_1 \to 3l_1$
$f'(X - 4l_1)$	0	0	0
$g'(X + 4l_1)$	0	$\frac{v_0}{2}$	$\frac{v_0}{2}$
$\frac{\partial u}{\partial t}(X, 4\frac{l_1}{c})$	0	$\frac{v_0}{2}$	$\frac{v_0}{2}$

TAB. 1.2 – Vitesse avant et après impact

1.1.2 Modèle rigide

On peut comparer ces résultats à ceux d'un modèle de solides rigides. Si on considère le choc comme parfaitement élastique (c'est-à-dire qu'il ne dissipe pas d'énergie), les équations suivantes suffisent à déterminer les vitesses v_1 et v_2 des deux barres après le

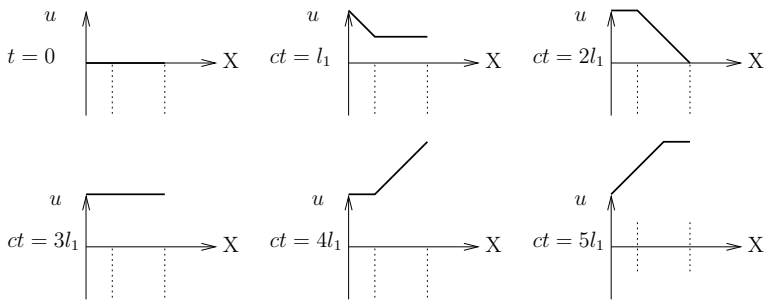


FIG. 1.4 – Déplacement induit par la propagation des ondes élastiques

choc

$$\rho\, l_1 v_1 + \rho\, l_2 v_2 = \rho\, l_1 v_0 \tag{1.13}$$

$$\frac{1}{2}\rho\, l_1 v_1^2 + \frac{1}{2}\rho\, l_2 v_2^2 = \frac{1}{2}\rho\, l_1 v_0^2. \tag{1.14}$$

L'équation (1.13) exprime la conservation de la quantité de mouvement, et (1.14) celle de l'énergie. Après quelques calculs, on trouve que les vitesses après le choc sont $v_1 = \frac{l_1 - l_2}{l_1 + l_2} v_0$ et $v_2 = \frac{2l_1}{l_1 + l_2} v_0$, ce qui est très différent de la prédiction du modèle élastique selon lequel la barre 1 est immobile et la barre 2 se déplace à vitesse uniforme $\frac{v_0}{2}$.

1.1.3 Modèle de Newton

Une troisième possibilité consiste à utiliser une loi d'impact empirique connue sous le nom de « loi de Newton » , selon laquelle la vitesse normale relative après un impact est égale à l'opposé de la vitesse normale relative avant impact, multipliée par un coefficient d'amortissement $\alpha \in [0, 1]$. Les équations suivantes suffisent à déterminer les vitesses après le choc

$$\rho\, l_1 v_1 + \rho\, l_2 v_2 = \rho\, l_1 v_0 \tag{1.15}$$

$$v_2 - v_1 = \alpha v_0. \tag{1.16}$$

Après quelques calculs, il vient cette fois $v_1 = \frac{l_1 - \alpha l_2}{l_1 + l_2} v_0$ et $v_2 = \frac{(1+\alpha) l_1}{l_1 + l_2} v_0$.

1.1.4 Valeur du coefficient de restitution de Newton

En utilisant le modèle élastique, on peut essayer de calculer la valeur du coefficient de restitution α dans le modèle des barres élastiques. Autrement dit, en admettant que le modèle élastique soit plus fin que les modèles rigides, on peut l'utiliser pour calibrer la valeur de α de la loi de Newton. Supposons, pour réduire le nombre de cas à étudier, que $l_2 \geq 2l_1$. Alors on peut montrer que les fonctions f', g' et la vitesse v au moment de la séparation au temps T ($v(X, T) = -c(f'(X - cT) + g'(X + cT))$) prennent les valeurs suivantes.

	$0 \leq ct \leq l_1$	$l_1 \leq ct \leq 3l_1$	$3l_1 \leq ct \leq 2l_2 - l_1$
f'	0	0	0
g'	0	$v_0/(2c)$	0
v	0	$v_0/2$	0

Ainsi, la vitesse de la barre 2 après le choc est $\frac{v_0}{2}$ sur une longueur $2l_1$ et 0 sur le reste de sa longueur. Pour relier cette distribution de vitesse à la vitesse v_2 dans le modèle rigide, on considère que v_2 est la vitesse du centre de gravité de la barre 2, ce qui donne

$$v_2 = \frac{1}{\rho\, l_2} \int_{l_1}^{3l_1} \rho \frac{v_0}{2} dx = \frac{l_1}{l_2} v_0$$

et donc le coefficient de restitution α vaut

$$\alpha := \frac{v_2 - v_1}{v_0} = \frac{l_1}{l_2}. \tag{1.17}$$

Ce coefficient *dépend donc de la géométrie* du système étudié, ici des longueurs l_1 et l_2, et le modèle élastique ne fournit pas de valeur « universelle » pour α. Ce résultat est décourageant, au sens où il semble impossible de modéliser la collision entre deux solides sans passer par un modèle déformable avec une discrétisation spatiale fine des objets, non seulement au niveau de la zone de contact, mais aussi dans le reste du solide puisque le phénomène d'impact met en jeu des ondes qui se propagent dans tout le solide. Autrement dit, l'élaboration d'un modèle général d'impact pour les solides indéformables, dépendant d'un petit nombre de paramètres, semble irréaliste.

D'autre part, dans le modèle retenu, la « perte » d'énergie à l'impact vient uniquement de la propagation d'ondes élastiques dans les solides, qui ne la dissipent pas. Si on raffinait la loi de comportement du matériau pour intégrer de la plasticité et de la viscosité, le résultat serait certainement encore différent mais dépendrait sans doute moins de la géométrie globale, puisque les déformations seraient alors locales et se propageraient moins loin dans le solide à cause de l'amortissement. En revanche, on découvrirait peut-être une dépendance de α par rapport à la vitesse d'impact v_0, dont la formule (1.17) est, de façon remarquable, indépendante.

1.1.5 Loi d'impact retenue

Dans cette thèse, on s'intéresse au phénomène du frottement, pour lequel on retiendra le modèle (1.22) ci-dessous, qui suffit à empêcher l'interpénétration. En ce sens, il est capable de produire des impacts, mais il ne permet pas de modéliser les rebonds. Il s'agit en quelque sorte de la loi d'impact la plus « paresseuse », qui ne produit que des chocs inélastiques, et on aimerait enrichir le modèle avec une loi d'impact plus élaborée de manière à autoriser des rebonds. Cependant, les sous-sections qui précèdent montrent qu'il n'est guère possible de modéliser les impacts de façon très générale, sans considérer la géométrie du système. Avec pragmatisme, on adoptera donc pour nos simulations soit des chocs inélastiques, soit la règle de Moreau qui consiste à

– fixer un coefficient de restitution $\alpha \in [0, 1]$,

– retenir à chaque pas de temps la vitesse relative au point de contact dans la direction normale, notée u_{N}^{old}
– et à remplacer pour chaque point de contact la vitesse u par la vitesse modifiée $u + \alpha\, u_{\mathrm{N}}^{old}$ dans la loi de frottement.

Ainsi, lorsqu'une impulsion de contact a lieu, on a $u + \alpha u_{\mathrm{N}}^{old} = 0$ (au lieu de $u = 0$), c'est-à-dire $u = -\alpha u_{\mathrm{N}}^{old}$. Une autre manière de voir cette technique est la suivante : si un objet entre en contact avec un mur immobile, on « ment » au modèle en faisant comme si le mur se déplaçait vers l'objet à la vitesse $\alpha u_{\mathrm{N}}^{old}$, puis on modélise le choc comme inélastique. On retrouve donc de cette manière la loi d'impact de Newton à un paramètre pour les impacts normaux. Au besoin, on peut ajouter de la même manière des impacts tangentiels. Cette loi phénoménologique n'est valide que pour des objets dont la géométrie est proche de celle de la sphère ; pour d'autres systèmes, elle peut provoquer la création d'énergie dans le système par les forces de contact et frottement [Str91].

Une autre approche possible est de traiter séparément les impacts et le frottement, en appliquant au début de chaque pas de temps des impacts calculés d'une manière ou d'une autre selon le système étudié, puis en avançant d'un pas de temps sans considérer les impacts. En tout état de cause, nous ne nous préoccuperons plus des impacts dans ce document.

1.2 Modèles de frottement

Les modèles utilisés pour décrire le phénomène de frottement dépendent de la nature des matériaux en contact. La *tribologie*, la science du frottement, a pour objet de comprendre ce phénomène et de proposer des lois de comportement adaptées aux différents matériaux et à leurs différents états de surface.

Pour formaliser, identifions l'espace qui nous entoure à $\mathbb{R}^d$ (avec $d = 2$ ou 3) en supposant qu'une base orthonormale y a été fixée, de sorte que le produit scalaire usuel peut être identifié au produit scalaire canonique de $\mathbb{R}^d$. Les éléments de $\mathbb{R}^d$ sont notés $x = (x_1, \ldots, x_d)$ et le d-uplet x est à son tour identifié au vecteur colonne $x = [x_1, \ldots x_d]^\top$. En un point de contact donné, on considère la vitesse relative $u \in \mathbb{R}^d$ des deux objets en contact, et la force de contact $r \in \mathbb{R}^d$. L'orientation est arbitraire, on appelle l'un des solides A, l'autre B, et on définit u comme la vitesse de B par rapport à A ; de même, r est la force exercée par A sur B. Un modèle de frottement est la donnée d'un ensemble $C \subset \mathbb{R}^d \times \mathbb{R}^d$: on dit que le couple (u, r) satisfait la loi de frottement lorsque $(u, r) \in C$. C'est la structure plus ou moins compliquée de l'ensemble C qui détermine la difficulté posée par la non-régularité en loi.

La variable (u, r) est de dimension $2d$ et la dynamique du système impose a priori une contrainte de dimension d. Pour que le problème soit raisonnable (autant d'équations que d'inconnues), il faut donc que la loi de frottement impose une contrainte « de dimension d » : plus précisément, on aimerait que C soit une sous-variété de dimension d de $\mathbb{R}^{2d}$. Malheureusement, pour la loi de Coulomb, l'ensemble C n'est qu'une union de sous-variétés de dimension d. C'est ce qu'on appelle la non-régularité en loi, qui provoque le

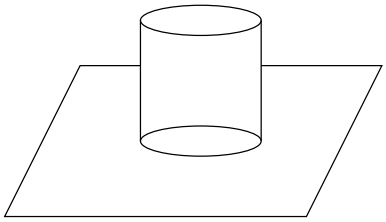


FIG. 1.5 – Une zone de contact avec une infinité de points, non polyédrale

caractère non-régulier du problème.

1.2.1 Forces et vitesses locales

Considérons un système mécanique comportant un certain nombre fini n de contacts ponctuels à un instant donné t.

L'hypothèse du nombre fini de points de contact est très restrictive, puisqu'elle interdit les situations où deux objets se touchent par une face ou une arête. Lorsque la zone de contact est polyédrale (un segment, un triangle, etc) et que les corps sont indéformables, on peut considérer que le contact a lieu seulement aux points extrêmes du polytope de contact, et l'on se ramène facilement à un nombre fini de points de contacts [Bar89]. Dans les autres cas, pour une surface de contact non polyédrale comme sur la figure 1.5 ou pour des objets déformables, l'approche qui va être présentée ne s'applique pas directement. La solution la plus immédiate est de ne considérer qu'un nombre élevé, mais fini, de points de contacts et de traiter le problème de manière approximative en ignorant les autres.

Dans cette section seulement, on considère que le temps est fixé et la dépendance en t n'est pas indiquée dans les notations. Pour chaque contact, étiqueté par $i \in 1, \dots n$, on note A^i et B^i les deux corps impliqués, et on suppose qu'ils sont suffisamment lisses au voisinage de leur point de contact pour qu'un vecteur normal unitaire e et un plan tangent commun soit définis sans ambiguïté, de sorte que $(e, e_{\mathrm{T}_1}, \dots e_{\mathrm{T}_{d-1}})$ forme une base orthonormale, appelée la base locale (figure 1.6). On note alors x_{N} et x_{T} les composantes normale et tangentielle d'un vecteur x, définies par

$$x_{\mathrm{N}} := x \cdot e \quad \text{et} \quad x_{\mathrm{T}} := (x \cdot e_{\mathrm{T}_1}, \dots, x \cdot e_{\mathrm{T}_{d-1}}). \tag{1.18}$$

Noter que la définition de x_{N} et x_{T} dépend de la base locale choisie, et que lorsque les corps ne sont pas lisses au point de contact la direction normale peut ne pas être unique (figure 1.7). On définit aussi les matrices de projection sur la direction normale $P_{\mathrm{N}} \in \mathbb{R}^{1\times d}$ et sur le plan tangent $P_{\mathrm{T}} \in \mathbb{R}^{d-1\times d}$ par

$$P_{\mathrm{N}} := [e]^\top \quad \text{et} \quad P_{\mathrm{T}} := [e_{\mathrm{T}_1}, \dots, e_{\mathrm{T}_{d-1}}]^\top. \tag{1.19}$$

Enfin, en prenant le corps A^i comme référence, on considère la vitesse relative locale u^i de B^i par rapport à A^i et la force r^i appliquées par A^i sur B^i.

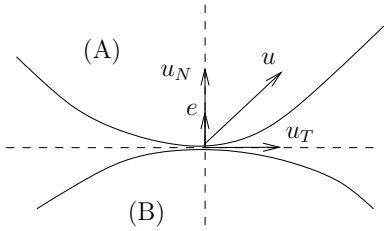


FIG. 1.6 – Les corps A^i et B^i avec le plan tangent et la direction normale

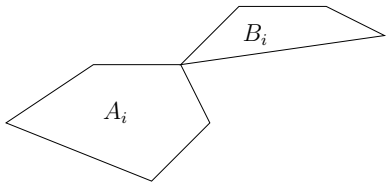


FIG. 1.7 – La normale et la tangente au point de contact sont mal définies

1.2.2 Modèle visqueux

Ce modèle est paramétré par une constante physique $\alpha \in \mathbb{R}_+$. La loi s'écrit simplement

$$r = -\alpha u.$$

Ce modèle a le mérite d'être simple et facile à utiliser. Il est parfois acceptable pour le contact entre un solide et un fluide, mais pas pour le contact entre deux solides. En effet, d'une part il n'empêche pas l'interpénétration qui doit donc être prévenue séparément (par exemple par des forces de pénalité), et d'autre part il ne capture pas le phénomène de seuil : une brique posée sur un plan incliné ne peut pas rester à l'équilibre sous l'effet de son poids et de la force de frottement visqueux. Par contre, la structure extrêmement simple de C (un hyperplan !) permet d'envisager des simulations rapides et de très gros problèmes.

1.2.3 Modèle de Tresca

Ce modèle dépend aussi d'un paramètre physique constant $\theta \in \mathbb{R}_+$, et se formule ainsi

- soit (adhérence) : $\|r_{\mathrm{T}}\| \leq \theta$ et $u_{\mathrm{T}} = 0$
- soit (glissement) : $\|r_{\mathrm{T}}\| = \theta$ et u_{T} est opposé à r_{T}

$$\exists \alpha \geq 0 \ : \ r_{\mathrm{T}} = -\alpha\, u_{\mathrm{T}}.$$

Seule la partie tangente $(r_{\mathrm{T}}, u_{\mathrm{T}})$ est prise en compte dans ce modèle, l'interpénétration doit donc encore être empêchée à l'aide d'un modèle supplémentaire. Le phénomène de seuil est bien modélisé, il est donc possible d'obtenir l'équilibre statique d'une brique sur un plan incliné (figure 1.9). Par contre, ce modèle ne capture pas un autre phéno-

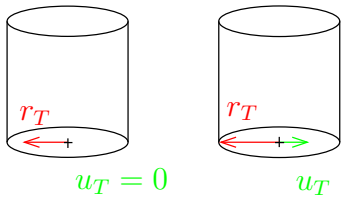


FIG. 1.8 – Loi de frottement de Tresca

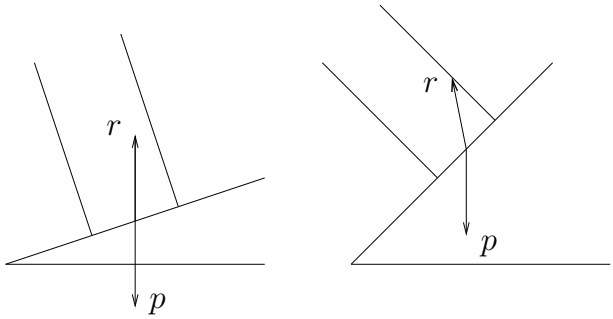


FIG. 1.9 – Le frottement de Tresca capture le phénomène de seuil

mène important du frottement solide : l'augmentation de la force tangentielle lorsque la force normale augmente. Autrement dit, deux objets possédant la même géométrie mais des masses différentes pourraient être mis en glissement sur un sol horizontal par la même force de poussée. L'expérience indique plutôt qu'en général, la force de frottement qui s'oppose au mouvement des objets lourds est plus grande que celle qui s'oppose au mouvement des objets légers. La loi de Tresca ne peut pas reproduire ce comportement, contrairement à la loi de Coulomb illustrée à la section suivante.

1.2.4 Modèle de Coulomb, formulation disjonctive

Le bon modèle pour une application donnée est celui qui réalise le meilleur compromis entre la précision du modèle (il doit ressembler autant que possible à la réalité) et sa nécessaire simplicité (pour pouvoir envisager des calculs effectifs sur ordinateur pour l'application visée). Nous supposerons que la loi de frottement de Coulomb est celle qui réalise ce compromis : elle est physiquement acceptable dans de nombreuses situations tout en restant suffisamment simple à formuler pour qu'il soit envisageable de mener des simulations, même de grande taille. Elle permet de capturer le phénomène de seuil. On définit le *cône du second ordre*, ou *cône de frottement*, de coefficient μ ($0 \leq \mu \leq \infty$) et de direction e par

$$K := \{x \ : \ \|x_{\mathrm{T}}\| \leq \mu\, x_{\mathrm{N}}\}. \tag{1.20}$$

Noter que cette définition dépend de la direction normale e choisie et du coefficient μ, mais pas de la base choisie dans le plan tangent. Le modèle historique de Coulomb est alors le suivant : aux points de contact actifs, que l'on a supposé présents en nombre fini

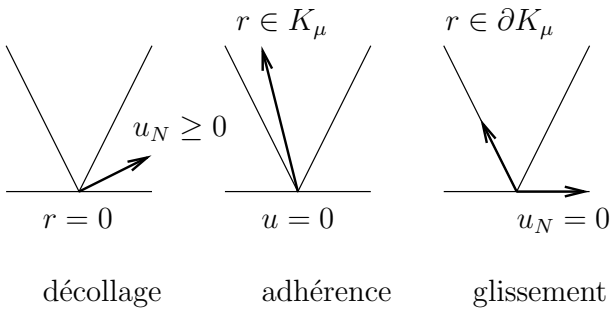


FIG. 1.10 – Les trois cas de la loi de Coulomb

dans le système, la force r exercée par un corps A sur un corps B et la vitesse relative u de B par rapport à A au niveau du point de contact doivent vérifier (au moins) l'une des contraintes suivantes

- décollage : $r = 0$ et $u_{\mathrm{N}} \geq 0$,
- adhérence : $r \in K$ et $u = 0$,
- glissement : $r \in \partial K \setminus 0$, $u_{\mathrm{N}} = 0$, $u_{\mathrm{T}} \neq 0$ opposé à r_{T} :

$$\exists \alpha > 0,\ r_{\mathrm{T}} = -\alpha\, u_{\mathrm{T}} = -\alpha\, u.$$

Ces trois cas sont représentés sur la figure 1.10. On remarque que ce modèle n'autorise pas de force attractive entre les objets car r_{N} est toujours positive (ceci modélise le contact sec, sans adhérence, non cohésif). De plus, cette force répulsive ne peut qu'empêcher l'interpénétration ($u_{\mathrm{N}} \geq 0$) et non pas faire décoller les objets ($u_{\mathrm{N}} > 0 \Rightarrow r_{\mathrm{N}} = 0$). Le problème de la loi d'impact, qui permet de modéliser le rebond lors d'un choc, est ignoré conformément à la sous-section 1.1.5.

On pourrait encore enrichir le modèle : en réalité, la valeur de μ n'est pas constante en général, mais elle diminue avec la force normale. C'est ce qu'avait observé Coulomb pour le contact bois-bois, en mesurant une valeur de μ d'environ 0.3 en laboratoire (avec une petite force normale) et d'environ $\frac{1}{14}$ sur les chantiers navals (avec une force normale beaucoup plus importante). D'autre part, on distingue souvent entre le coefficient de frottement statique et le coefficient de frottement dynamique, supposé plus faible. Plus finement encore, on peut considérer que le coefficient de frottement est une fonction (décroissante) de la vitesse de glissement. De nombreuses améliorations tribologiques supplémentaires sont envisageables, au prix d'une complexité croissante du modèle. Nous adopterons la loi de Coulomb à un seul paramètre μ identique pour le frottement statique et le frottement dynamique. La table 1.3 tirée de [Wri06] donne une idée de la valeur de μ pour quelques paires de matériaux.

On remarque aussi que la loi de Coulomb assure l'impénétrabilité au niveau des vitesses, en imposant une vitesse relative positive dans la direction normale ($u_{\mathrm{N}} \geq 0$). On peut voir ceci comme la dérivée par rapport au temps de la contrainte de non-pénétration, et le lemme de viabilité de Moreau [Mor88] montre que la contrainte $u_{\mathrm{N}} \geq 0$ suffit effectivement à empêcher l'interpénétration. Certains auteurs utilisent même

Matériaux	Coefficient de frottement μ
béton - béton	0.5 - 1
béton - sable	0.35 - 0.6
béton - acier	0.2 - 0.4
métal - bois	0.3 - 0.65
bois - bois	0.4 - 1
caoutchouc - acier	0.15 - 0.65
acier - acier	0.2 - 0.8
acier - téflon	0.04 - 0.06
acier - glace	0.015 - 0.03

TAB. 1.3 – Quelques valeurs typiques de μ

une dérivation supplémentaire et imposent une accélération relative normale positive en chaque point de contact.

1.2.5 Discrétisation de la loi de Coulomb

Dans cette thèse, on s'intéresse uniquement au problème de contact dynamique discrétisé, et non pas au problème continu. Les inconnues seront donc des vitesses u et des forces r discrétisées, identifiées à des vecteurs de $\mathbb{R}^{nd}$ si l'on simule un système à n points de contact en dimension d. On écrira ainsi

$$u = (u^1, \dots, u^n) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$$

où les $u^i \in \mathbb{R}^d$ seront les vitesses relatives discrétisées (censées approcher les vitesses relatives du modèle continu) aux points de contact, et

$$r = (r^1, \dots, r^n) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$$

où les $r^i \in \mathbb{R}^d$ seront les impulsions discrétisées (censées approcher l'intégrale de la force de contact sur un pas de temps). Le rôle de la loi de frottement est de restreindre l'ensemble des couples $(u^i, r^i) \in \mathbb{R}^d \times \mathbb{R}^d$ à un ensemble C, pour l'instant abstrait

$$\forall i \in 1, \dots n \ : \ (u^i, r^i) \in C(e^i, \mu^i) \subset \mathbb{R}^d \times \mathbb{R}^d. \tag{1.21}$$

L'ensemble $\mathcal{C}(e, \mu)$ contient les couples (u, r) compatibles avec la loi de frottement de Coulomb de coefficient μ en un point de contact où la normale est e. Pour le définir précisément, on discrétise directement le modèle de Coulomb continu de la manière suivante. Soit $(u, r) \in \mathbb{R}^d \times \mathbb{R}^d$; les vecteurs normaux et tangents r_{N}, r_{T}, u_{N} et u_{T} sont définis par (1.18), K par (1.20), et on définit $C(e, \mu)$ par

$$(u, r) \in \mathcal{C}(e, \mu) \iff \begin{cases} \text{soit} : r = 0 \text{ et } u_{\mathrm{N}} \geq 0 \\ \text{soit} : r \in K \text{ et } u = 0 \\ \text{soit} : r \in \partial K \setminus 0, \ u_{\mathrm{N}} = 0, \exists \alpha > 0, \ r_{\mathrm{T}} = -\alpha u. \end{cases} \tag{1.22}$$

Sous cette forme disjonctive, la contrainte $(u, r) \in \mathcal{C}(e, \mu)$ semble difficilement exploitable. Cependant, on verra dans la section suivante que cette contrainte n'est pas une disjonction de trois cas sans rapport les uns avec les autres, mais qu'elle possède au contraire beaucoup de structure grâce à laquelle on parvient souvent à calculer effectivement les forces de frottement lors des simulations.

1.3 Reformulations de la loi de Coulomb

Le problème complet auquel nous allons nous intéresser – (1.33) ci-dessous – consiste à trouver un couple (u, r) (plus éventuellement d'autres inconnues) qui satisfait d'une part la loi de Coulomb et d'autre part un autre jeu d'équations (cinématiques et dynamiques). Dans cette thèse, ces équations additionnelles seront supposées linéaires par rapport à u, r et aux autres variables éventuelles, et c'est dans la satisfaction de la loi de frottement (1.22) que se concentreront les difficultés. Puisque la loi de frottement est formulée comme une contrainte disjonctive à trois cas pour chacun des n contacts, on peut envisager de tester chacun des 3^n cas. Ceci devient impraticable dès que n dépasse quelques unités. Heureusement, il est possible de reformuler la contrainte (1.22) de manière plus pratique,

- soit comme l'ensemble des zéros ou des points fixes d'une fonction (approche fonctionnelle),
- soit, en faisant un changement de variables, comme une contrainte de complémentarité conique (approche par complémentarité) ; de plus, en dimension 2, le problème complet (1.33) ci-dessous, constitué par la loi de frottement (1.22) et les équations linéaires additionnelles, se met sous la forme problème de complémentarité linéaire (LCP).

Les deux ingrédients de base pour effectuer ces reformulations sont la fonction d'Alart et Curnier [AC91] (sous-section 1.3.1) et le changement de variables de De Saxcé [DSF98] (sous-section 1.3.3).

1.3.1 Formulation d'Alart et Curnier

On fixe deux constantes $\rho_{\mathrm{N}}, \rho_{\mathrm{T}} > 0$ (disons $\rho_{\mathrm{N}} = \rho_{\mathrm{T}} = 1$). On note P_S la projection sur un ensemble convexe fermé S, et $B(0, \alpha)$ la boule euclidienne de rayon α, ici en dimension $d-1$, centrée en 0. Elle est vide lorsque $\alpha < 0$. On définit ensuite la partie normale de la fonction d'Alart et Curnier

$$f_{AC,\mathrm{N}} : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ (u, r) \longmapsto \mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}) - r_{\mathrm{N}} \end{cases} \tag{1.23}$$

et sa partie tangentielle

$$f_{AC,\mathrm{T}} : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^{d-1} \\ (u, r) \longmapsto \mathrm{P}_{B(0,\, \mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) - r_{\mathrm{T}} \end{cases} \tag{1.24}$$

avec par convention, $\forall y \in \mathbb{R}^{d-1},\ \mathrm{P}_{\emptyset}(y) = 0_{d-1}$.

Propriété 1.1. On peut reformuler (1.22) de la manière suivante

$$(u,r) \in \mathcal{C}(e,\mu) \iff [f_{AC,\mathrm{N}}, f_{AC,\mathrm{T}}](u,r) = 0. \tag{1.25}$$

Démonstration. La démonstration est élémentaire dans chacun des cas, il suffit de les considérer tous. Montrons d'abord que (1.22) $\Rightarrow [f_{AC,\mathrm{N}}, f_{AC,\mathrm{T}}](u,r) = 0$. Dans le cas du décollage, $r = 0$ et $u_{\mathrm{N}} \geq 0$ donc $\mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}) = \mathrm{P}_{\mathbb{R}^+}(-\rho_{\mathrm{N}}\, u_{\mathrm{N}}) = 0 = r_{\mathrm{N}}$ et $\mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = \mathrm{P}_{B(0,0)}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = 0 = r_{\mathrm{T}}$. Dans le cas de l'adhérence, $u = 0$ et $r \in K$ donc $\mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}) = \mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}}) = r_{\mathrm{N}}$ et $\mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = \mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}}) = r_{\mathrm{T}}$. Enfin dans le cas du glissement, $\mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}) = \mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}}) = r_{\mathrm{N}}$ (car $u_{\mathrm{N}} = 0$ et $r_{\mathrm{N}} \geq 0$) et $\mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = r_{\mathrm{T}}$ (car $r_{\mathrm{T}} \in \partial K$ et $-\rho_{\mathrm{T}}\, u_{\mathrm{T}} \in N_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}})$.

Réciproquement, si $[f_{AC,\mathrm{N}}, f_{AC,\mathrm{T}}](u,r) = 0$, montrons que (1.22) est vérifiée. Comme $f_{AC,\mathrm{T}}(u,r) = 0$ on a $r_{\mathrm{T}} \in B(0,\, \mu\, r_{\mathrm{N}})$ donc $r \in K$. Comme $f_{AC,\mathrm{N}}(u,r) = 0$, on a $r_{\mathrm{N}} \geq 0$ et on distingue les cas $r_{\mathrm{N}} = 0$ (*) et $r_{\mathrm{N}} > 0$ (**). Si $r_{\mathrm{N}} = 0$ (*) alors $r_{\mathrm{T}} = 0$ (car $r_{\mathrm{T}} = \mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = \mathrm{P}_{B(0,0)}(\dots) = 0$) et $u_{\mathrm{N}} \geq 0$ (car $0 = r_{\mathrm{N}} = \mathrm{P}_{\mathbb{R}^+}(-\rho_{\mathrm{N}}\, u_{\mathrm{N}})$) et on est dans le cas du décollage. Si $r_{\mathrm{N}} > 0$ (**) alors $r_{\mathrm{N}} = \mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}) = r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}}$ donc $u_{\mathrm{N}} = 0$. On distingue de nouveau deux cas : $r_{\mathrm{T}} \in \operatorname{int} B(0,\, \mu\, r_{\mathrm{N}})$ et $r_{\mathrm{T}} \in \partial B(0,\, \mu\, r_{\mathrm{N}})$. Dans le premier cas, $r_{\mathrm{T}} = \mathrm{P}_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}) = r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}}$ donc $u_{\mathrm{T}} = 0$, c'est le cas de l'adhérence. Dans le second cas, $r \in \partial K$ et $u_{\mathrm{T}} \in N_{B(0,\,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}})$ donc $\exists \alpha > 0$: $r_{\mathrm{T}} = -\alpha\, u_{\mathrm{T}}$, c'est le glissement. Ceci achève la démonstration. ∎

Les formules pour $f_{AC,\mathrm{N}}$ et $f_{AC,\mathrm{T}}$ sont explicites et peuvent être différentiées sur une partie de leur domaine (pas partout car ce sont des fonctions non-régulières), ce qui permet d'utiliser l'algorithme de Newton pour rechercher leurs zéros. La justification théorique de cette technique est délicate (aux points de non-régularité, l'itéré de Newton n'est pas défini) mais elle fonctionne souvent bien en pratique (voir [AC91, CKPS98] et le chapitre 5).

1.3.2 Formulation de Haslinger

A partir de la formulation d'Alart et Curnier, on passe facilement à celle proposée par Haslinger. L'idée essentielle consiste à considérer μr_{N} comme un paramètre $s \in \mathbb{R}^+$; on introduit ensuite le demi-cylindre $T(s) \subset \mathbb{R}^d$ par

$$T(s) := \{r \in \mathbb{R}^d \colon r_{\mathrm{N}} \geq 0,\ \|r_{\mathrm{T}}\| \leq s\}. \tag{1.26}$$

On a alors la caractérisation suivante de la loi de Coulomb.

Propriété 1.2. On note $N_{T(s)}(r)$ le cône normal à $T(s)$ en r. Alors

$$(u,r) \in \mathcal{C}(e,\mu) \iff [r \in T(\mu r_{\mathrm{N}}) \text{ et } u \in -N_{T(\mu r_{\mathrm{N}})}(r)]. \tag{1.27}$$

Démonstration. Considérons de nouveau $f_{AC,\mathrm{N}}$ et $f_{AC,\mathrm{T}}$ définies par (1.23) et (1.24). L'équation $f_{AC,\mathrm{N}}(u,r) = 0$ exprime exactement que $r_{\mathrm{N}} \in \mathbb{R}^+$ et $u_{\mathrm{N}} \in -N_{\mathbb{R}^+}(r_{\mathrm{N}})$; l'équation $f_{AC,\mathrm{T}}(u,r) = 0$ signifie que $r \in B(0, \mu r_{\mathrm{N}})$ et que $u \in -N_{B(0,\mu r_{\mathrm{N}})}(r)$. Enfin

comme $T(\mu r_{\mathrm{N}}) = \mathbb{R}^+ \times B(0, \mu r_{\mathrm{N}})$, $[f_{AC,\mathrm{N}}, f_{AC,\mathrm{T}}](u, r) = 0$ équivaut à $r \in T(\mu r_{\mathrm{N}})$ et $u \in -N_{T(\mu r_{\mathrm{N}})}(r)$. ■

L'intérêt de cette reformulation est de faire apparaître le cône normal à l'ensemble convexe $T(s)$, ce qui favorise l'introduction des techniques d'optimisation (algorithme d'Uzawa [HT83], itérations de point fixe [HKD04]). Comme la structure de $T(s)$ est très simple (en particulier, le calcul de la projection sur $T(s)$ est trivial), la résolution de problèmes d'optimisation sur cet ensemble est envisageable en pratique.

1.3.3 Formulation de De Saxcé

Dans cette sous-section, on présente une approche proposée par De Saxcé [DSF98] qui consiste à effectuer un changement de variables après lequel la contrainte (1.22) devient une contrainte de complémentarité conique du second ordre. Introduisons le changement de variable $u \to \widetilde{u}$ (voir la figure 1.11) défini par

$$\widetilde{u} := u + \mu \|u_{\mathrm{T}}\| e. \tag{1.28}$$

Propriété 1.3. Soit

$$K^* := -K^\circ = \{y \colon \mu \|y_{\mathrm{T}}\| \leq y_{\mathrm{N}}\}$$

le cône dual de K (où K° est le cône polaire de K, [HUL01]). La formulation (1.22) de la loi de Coulomb se récrit

$$K^* \ni \tilde{u} \perp r \in K. \tag{1.29}$$

Démonstration. De nouveau, on considère tous les cas. Si (1.22) est vérifiée, alors $r \in K$ et $u_{\mathrm{N}} \geq 0$ donc $\widetilde{u} \in K^*$, il suffit de vérifier que $\widetilde{u} \perp r$. Dans les cas du décollage ($r = 0$) et de l'adhérence ($u = 0$ donc $\widetilde{u} = 0$) c'est évident, et dans le cas du glissement on a

$$\widetilde{u} \cdot r = u_{\mathrm{T}} \cdot r_{\mathrm{T}} + \mu r_{\mathrm{N}} \|u_{\mathrm{T}}\| = -\|r_{\mathrm{T}}\| \|u_{\mathrm{T}}\| + \mu r_{\mathrm{N}} \|u_{\mathrm{T}}\| = -\mu \|r_{\mathrm{N}}\| \|u_{\mathrm{T}}\| + \mu r_{\mathrm{N}} \|u_{\mathrm{T}}\| = 0.$$

Dans tous les cas, (1.29) est vérifiée.

Réciproquement, si (1.29) est vérifiée, on voit que $u_{\mathrm{N}} + \mu \|u_{\mathrm{T}}\| = \widetilde{u}_{\mathrm{N}} \geq \mu \|\widetilde{u}_{\mathrm{T}}\| = \mu\, u_{\mathrm{T}}$ (car $\widetilde{u} \in K^*$) donc $u_{\mathrm{N}} \geq 0$. On distingue trois cas (*), (**) et (***). Si $r = 0$ (*) alors, comme $u_{\mathrm{N}} \geq 0$, (1.22) est vérifiée (décollage). Si $\widetilde{u} = 0$ (**) alors $u = 0$ et comme $r \in K$ par hypothèse, (1.22) est vérifiée (adhérence). Enfin si $\widetilde{u} \neq 0$ et $r \neq 0$ (***), alors $\widetilde{u} \in \partial K^*$ donc $u_{\mathrm{N}} = 0$, et $r \in \partial K \setminus \{0\}$. De plus, $0 = \widetilde{u} \cdot r = u_{\mathrm{T}} \cdot r_{\mathrm{T}} + \mu\, r_{\mathrm{N}} \|u_{\mathrm{T}}\|$ donc $\mu\, r_{\mathrm{N}} \|u_{\mathrm{T}}\| = |u_{\mathrm{T}} \cdot r_{\mathrm{T}}| \leq \|u_{\mathrm{T}}\| \|r_{\mathrm{T}}\| = \|u_{\mathrm{T}}\| \mu\, r_{\mathrm{N}}$. L'inégalité de Cauchy-Schwarz est une égalité, ce qui prouve que u_{T} et r_{T} sont colinéaires. Comme enfin $u_{\mathrm{T}} \cdot r_{\mathrm{T}} = -\mu\, r_{\mathrm{N}} \|u_{\mathrm{T}}\| < 0$, ils sont en sens inverse et (1.22) est satisfaite (glissement). Ceci achève la démonstration. ■

Dans la nouvelle « puissance »

$$\widetilde{u} \cdot r = u \cdot r + \mu \|u_{\mathrm{T}}\| r_{\mathrm{N}},$$

le terme supplémentaire $\mu \|u_{\mathrm{T}}\| r_{\mathrm{N}}$ est appelé le *bipotentiel* de De Saxcé.

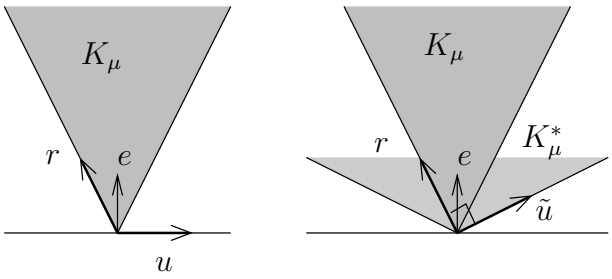


FIG. 1.11 – Changement de variable $u \to \widetilde{u}$

1.3.4 Formulation par complémentarité linéaire en dimension 2

En dimension 2, on peut reformuler (1.22) comme une contrainte de complémentarité linéaire sans passer par le changement de variable non-linéaire $u \to \widetilde{u}$ défini par (1.28). De plus, la contrainte de complémentarité est posée directement sur l'orthant positif. En contrepartie, le nombre de variables est doublé : de $(u, r) \in \mathbb{R}^4$, on passe à $\mathbb{R}^8$. L'idée est de considérer le problème suivant : trouver $(u_{\mathrm{N}}, u_{\mathrm{T}}, r_{\mathrm{N}}, r_{\mathrm{T}}, \eta^+, \eta^-, \lambda^+, \lambda^-) \in \mathbb{R}^8$ tel que

$$\left\{\begin{array}{lcll} r_{\mathrm{T}} & = & \lambda^+ - \lambda^- & (a) \\ \mu\, r_{\mathrm{N}} & = & \lambda^+ + \lambda^- & (b) \\ u_{\mathrm{T}} & = & \eta^+ - \eta^- & (c) \end{array}\right. \quad \text{et} \quad \left\{\begin{array}{lcll} 0 \le u_{\mathrm{N}} & \perp & r_{\mathrm{N}} \ge 0 & (d) \\ 0 \le \lambda^- & \perp & \eta^- \ge 0 & (e) \\ 0 \le \lambda^+ & \perp & \eta^+ \ge 0. & (f) \end{array}\right. \tag{1.30}$$

Propriété 1.4. Pour tout $(u, r) \in \mathbb{R}^4$, $(u, r) \in \mathcal{C}(e, \mu)$ si et seulement s'il existe quatre réels η^+, η^-, λ^+, λ^- tels que (1.30) soit satisfaite.

Démonstration. Soit (u, r) vérifiant (1.22). On pose $\lambda^+ = (r_{\mathrm{T}} + \mu\, r_{\mathrm{N}})/2$ et $\lambda^- = (-r_{\mathrm{T}} + \mu\, r_{\mathrm{N}})/2$, ainsi que $\eta^+ = \max(0, u_{\mathrm{T}})$ et $\eta^- = -\min(0, u_{\mathrm{T}})$. Les trois égalités (1.30)-(a,b,c) et les six inégalités dans (1.30)-(d,e,f) sont satisfaites. Il reste à vérifier les trois conditions d'orthogonalité dans (1.30)-(d,e,f). Dans le cas du décollage, $r = 0$ donc $r_{\mathrm{N}} = \lambda^+ = \lambda^- = 0$. Dans le cas de l'adhérence, $u = 0$ donc $u_{\mathrm{N}} = \eta^+ = \eta^- = 0$. Dans le cas du glissement, $u_{\mathrm{N}} = 0$ et $r_{\mathrm{T}} = \pm\mu\, r_{\mathrm{N}}$ donc $\lambda^+ = 0$ (si $r_{\mathrm{T}} = -\mu\, r_{\mathrm{N}}$, et alors $u_{\mathrm{T}} \ge 0$ donc $\eta^- = 0$) ou $\lambda^- = 0$ (si $r_{\mathrm{T}} = +\mu\, r_{\mathrm{N}}$, et alors $u_{\mathrm{T}} \le 0$ donc $\eta^+ = 0$). Dans les trois cas, les conditions d'orthogonalité dans (1.30)-(d,e,f) sont satisfaites.

Réciproquement, s'il existe $(u_{\mathrm{N}}, u_{\mathrm{T}}, r_{\mathrm{N}}, r_{\mathrm{T}}, \eta^+, \eta^-, \lambda^+, \lambda^-)$ satisfaisant (1.30), alors $u_{\mathrm{N}} \ge 0$ et $r \in K$ car $r_{\mathrm{N}} \ge 0$ et $r_{\mathrm{T}} \in [-\mu\, r_{\mathrm{N}}, \mu\, r_{\mathrm{N}}]$ d'après (1.30)-(a,b) et le fait que $\lambda^+, \lambda^- \ge 0$. De plus, si $r_{\mathrm{N}} = 0$ alors $\lambda^+ = \lambda^- = 0$ d'après (1.30)-(b) donc $r_{\mathrm{T}} = 0$ d'après (1.30)-(a). Ainsi (1.22) est satisfaite (décollage). Si $r_{\mathrm{N}} > 0$ alors $u_{\mathrm{N}} = 0$ d'après (1.30)-(d). De plus, $(\lambda^+, \lambda^-) \neq (0, 0)$ d'après (1.30)-(b) et on distingue deux cas selon que (*) $\lambda^+ \neq 0$ et $\lambda^- \neq 0$ ou (**) l'une des deux variables λ^+ et λ^- est nulle. Dans le premier cas (*), $r_{\mathrm{T}} = \lambda^+ - \lambda^- \in]-\mu\, r_{\mathrm{N}}, \mu\, r_{\mathrm{N}}, [$ donc $r \in \mathrm{int}(K)$ et $\eta^+ = \eta^- = 0$ d'après (1.30)-(e,f), donc $u = 0$. Ainsi (1.22) est satisfaite (adhérence). Enfin si (**) l'une des deux variables λ^+ et λ^- est nulle, par exemple $\lambda^+ = 0$, alors $\lambda^- = \mu\, r_{\mathrm{N}}$ d'après (1.30)-(b) et $r_{\mathrm{T}} = -\mu\, r_{\mathrm{N}}$ d'après (1.30)-(a). D'autre part $\eta^- = 0$ d'après (1.30)-(e) donc $u_{\mathrm{T}} \ge 0$ et (1.22) est encore satisfaite (glissement). Ceci achève la démonstration. ■

1.4 Difficultés du modèle

Dans cette section, on présente le problème incrémental complet (équation (1.33) ci-dessous) dont la résolution est au coeur de cette thèse, et on tente d'évaluer sa difficulté en le comparant à des problèmes bien connus comme la programmation quadratique. La manière de construire effectivement les données du problème incrémental pour un problème mécanique particulier sera expliquée rapidement au chapitre suivant.

1.4.1 Problème incrémental

A chaque pas de temps on cherche les valeurs de trois vecteurs inconnus : les vitesses généralisées $v \in \mathbb{R}^m$ (m étant le nombre de degrés de liberté), les forces aux n points de contact que l'on rassemble dans un seul vecteur $r \in \mathbb{R}^{nd}$, et les vitesses relatives $u \in \mathbb{R}^{nd}$ en ces n points. Dans le cadre d'une cinématique et d'une dynamique discrétisée linéaires, les vitesses généralisées sont reliées aux vitesses relatives par une relation « cinématique »

$$u = H\,v + w \tag{1.31}$$

où H et w sont connus, et aux forces par une relation « dynamique discrétisée »

$$M\,v + f = H^\top r \tag{1.32}$$

où la matrice de masse M et le vecteur f sont connus. Enfin, on ajoute la loi de frottement en imposant la contrainte (1.22). Le problème final, appelé *problème incrémental*, est donc : trouver (v, u, r) dans $\mathbb{R}^{m+nd+nd}$ tel que

$$\left\{\begin{array}{l} M\,v + f = H^\top r \\ u = H\,v + w \\ (u^i, r^i) \in \mathcal{C}(e^i, \mu^i) \quad \forall i \in 1, \dots n. \end{array}\right. \tag{1.33}$$

Tout au long de la thèse, on fera l'hypothèse que la matrice de masse est symétrique définie positive.

Hypothèse 1.1.

$$M \in S_n^{++}.$$

En éliminant $v = M^{-1}(H^\top r - f)$ de la deuxième équation, et en posant $W := HM^{-1}H^\top$ et $q := w - HM^{-1}f$, on voit que (1.33) est équivalent à trouver (u, r) dans $\mathbb{R}^{nd+nd}$ tel que

$$\left\{\begin{array}{l} u = Wr + q \\ (u^i, r^i) \in \mathcal{C}(e^i, \mu^i) \quad \forall i \in 1, \dots n \end{array}\right. \tag{1.34}$$

assorti de l'équation $v = M^{-1}(H^\top r - f)$. On utilisera dans la suite les deux formulations (1.33) et (1.34). La deuxième version (1.34) peut sembler plus attrayante dans la mesure où le problème est de taille plus réduite grâce à l'élimination de v, mais en pratique on dispose rarement de W directement et son calcul par sa définition $W := HM^{-1}H^\top$ est coûteux. De plus, si M est souvent creuse en pratique, il est plus rare que W le soit.

1.4.2 Cas sans contact

Lorsqu'il n'y a aucun contact actif dans le système ($n = 0$), le problème revient à résoudre un système linéaire $M\,v + f = 0$ dans lequel la matrice est définie positive. C'est la situation habituelle en mécanique lagrangienne régulière ; en pratique, la résolution peut être effectuée par exemple par l'algorithme du gradient conjugué.

1.4.3 Cas sans frottement

Lorsque tous les coefficients de frottement sont nuls dans le système, (1.33) est un problème de complémentarité linéaire (LCP) qui constitue des conditions nécessaires et suffisantes d'optimalité pour un problème d'optimisation quadratique sous contraintes linéaires (QP). On peut donc utiliser n'importe quel solveur de QP ou de LCP pour le résoudre.

Plus précisément, lorsque tous les coefficients de frottement μ^i sont nuls, les variables r_{T} et u_{T} disparaissent de la loi de frottement (1.22) car r_{T} est fixé à zéro et u_{T} n'est plus contrainte. Quitte à effectuer un changement de base, on peut supposer que $r = (r_{\mathrm{N}}, r_{\mathrm{T}}^\top)^\top$ (autrement dit que la composante normale de r est égale à sa première coordonnée et sa composante tangentielle est égale à ses $d-1$ dernières coordonnées). On définit alors H_{N} comme la matrice extraite de H en conservant seulement les lignes $1, d+1, 2d+1\ldots$ correspondant aux composantes normales, et on extrait de la même manière w_{N} de w. Le problème (1.33) devient : trouver (v, u, r) tel que

$$\begin{cases} M\,v + f = H_{\mathrm{N}}^\top r_{\mathrm{N}} \\ u_{\mathrm{N}} = H_{\mathrm{N}}\,v + w_{\mathrm{N}} \\ 0 \le r_{\mathrm{N}} \perp u_{\mathrm{N}} \ge 0. \end{cases} \tag{1.35}$$

On reconnaît alors les conditions d'optimalité de Karush-Kuhn-Tucker du problème d'optimisation quadratique suivant en la variable v seulement :

$$\begin{cases} \min \frac{1}{2} v^\top M v + f^\top v \\ H_{\mathrm{N}} v + w_{\mathrm{N}} \ge 0. \end{cases} \tag{1.36}$$

Dans ce cas, on peut utiliser directement un solveur quadratique pour résoudre le problème (1.36) ou son problème dual, et obtenir la solution de (1.33). Comme les données M, f, H_{N} et w_{N} sont quelconques, le problème incrémental (1.33) est donc plus général que la programmation quadratique.

1.4.4 Non-unicité, non-existence

En s'inspirant de l'exemple couramment appelé « paradoxe » de Painlevé, on peut construire un exemple de problème très simple (un seul degré de liberté, un seul contact, en dimension 2) et instructif quant à l'existence et la nature des solutions de (1.33). On considère le système de la figure 1.12. Le point x_A se déplace sur l'axe Ox à vitesse imposée u_0. Une barre rigide homogène de masse ν et de longueur l est articulée au point x_A par une liaison pivot sans frottement. L'extrémité x_B de la barre peut être en

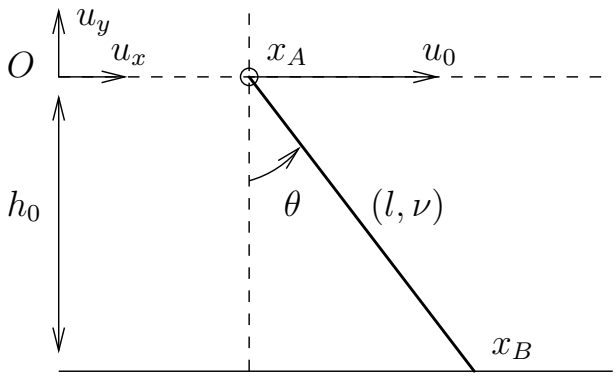


FIG. 1.12 – Un exemple très simple de problème de contact

contact avec le sol comme sur la figure 1.12, ou décoller du sol si θ augmente. Autrement dit, le contact entre le sol et la barre est unilatéral. En cas de contact, le sol exerce sur la barre une force r, et on note u la vitesse de x_B. Ce système possède un seul degré de liberté, paramétré par l'angle θ, et il est soumis au champ de gravitation g selon l'axe Oy.

L'évolution du système est gouvernée par l'équation

$$\frac{1}{3}\nu l^2\ddot{\theta} = \frac{l}{2}\nu g \sin(\theta) + l\,(r_x\,\cos\theta + r_y\,\sin\theta). \tag{1.37}$$

On discrétise cette équation en notant v la vitesse généralisée (approximation de $\dot{\theta}$) et v_0 sa valeur au pas de temps précédent. En approchant $\ddot{\theta}$ par $(v - v_0)/h$ où h est la durée d'un pas de temps, on obtient le problème (1.33) avec

$$M = \frac{1}{3h}\nu l^2,\ f = -\frac{l}{2}\nu g \sin\theta - \frac{1}{3h}\nu l^2 v_0,\ H = l\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix},\ w = \begin{bmatrix}u_0\\ 0\end{bmatrix}. \tag{1.38}$$

En faisant varier la valeur des paramètres physiques et des conditions initiales, on peut produire différentes situations intéressantes : plusieurs solutions distinctes, une solution unique, ou pas de solution du tout.

Non-unicité

On prend les valeurs suivantes : $\nu = 2$, $l = 1$, $g = 1$ (noter que $g > 0$ donc la gravité est orientée vers le haut), $h = 2/3$, $u_0 = 1$ et $v_0 = 0$. On ne fixe pas les valeurs de μ et θ, et on suppose que $h_0 < l$ de sorte que lors du contact $\theta \in]0, \pi/2[$. Les données (1.38) deviennent

$$M = 1,\ f = -\sin(\theta),\ H = \begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix},\ w = \begin{bmatrix}1\\ 0\end{bmatrix}. \tag{1.39}$$

et le problème incrémental (1.33) est

$$\left\{\begin{array}{lcll} v & = & \cos(\theta)r_x + \sin(\theta)r_y + \sin(\theta) & \\ u_x & = & \cos(\theta)v + 1 & \text{et}\quad (u, r) \in \mathcal{C}(e_y, \mu). \\ u_y & = & \sin(\theta)v & \end{array}\right. \tag{1.40}$$

Lemme 1.2. *Le problème (1.40) possède deux solutions si* $\tan(\theta) < \mu$*, et une solution unique sinon.*

Démonstration. Cherchons les solutions éventuelles du problème incrémental (1.40) en effectuant à la main une recherche énumérative ($n = 1$ donc il n'y a que trois cas). (1) Décollage : $r = 0$, donc $v = \sin(\theta)$ et $u_{\mathrm{N}} = u_y = \sin(\theta) \geq 0$ donc ceci est une première solution acceptable. (2) Adhérence : $u = 0$, ceci est impossible. (3) Glissement : $u_{\mathrm{N}} = u_y = 0$ donc $v = 0$, donc $u = (1, 0)$ et r doit vérifier les deux équations linéaires suivantes (*i.e.*, la première équation de (1.40) et la condition $r \in \partial K$ avec r_{T} opposé à u_{T})

$$\left\{ \begin{array}{lcl} \cos(\theta) r_x + \sin(\theta) r_y & = & -\sin(\theta) \\ r_x + \mu r_y & = & 0 \end{array} \right. \tag{1.41}$$

qui possède une solution avec $r_y \geq 0$ si et seulement si $\tan(\theta) < \mu$. ∎

On peut donc avoir deux solutions au problème incrémental (1.33) avec les données (1.39). Intuitivement, les deux solutions sont raisonnables. La première consiste à laisser décoller le point x_B sans exercer de force sur lui ; puisque sa trajectoire naturelle (sans force de contact r) ne viole pas la contrainte unilatérale, il est « inutile » qu'une force de contact agisse. Dans la deuxième solution, la force r exerce un couple sur la barre qui lui permet de ne pas décoller malgré sa vitesse de rotation initiale v_0. Le contact est maintenu et la barre se met à glisser. On remarque que la force de frottement exerce un couple sur la barre dans le sens horaire (anti-trigonométrique) parce que l'angle $\theta = \pi/4 \approx 0.78$ est plus petit que l'ouverture du cône de Coulomb K (ici $\mu = 2$ donc l'ouverture du cône est $\arctan(2) \approx 1.11$). Dans le cas contraire, lorsque l'angle θ est plus grand que l'ouverture du cône K, la force de frottement exerce un couple dans le sens anti-horaire (trigonométrique) et cette force a tendance à faire décoller la barre et non pas à la maintenir en contact.

Cette situation de non-unicité est gênante du point de vue théorique : lorsqu'elle se produit, on peut soupçonner que notre modèle (hypothèse de corps parfaitement rigides, loi de frottement, et toutes les hypothèses faites) atteint ses limites de validité puisqu'il ne permet pas de prédire de manière déterministe l'évolution du système. Lors des simulations, on ne détecte généralement pas l'existence de solutions multiples (l'algorithme s'arrête quand il en trouve une) ; faute de mieux, on suppose alors que l'algorithme va « naturellement » choisir parmi les multiples solutions celle qui est « la plus proche » de la solution précédente, utilisée pour initialiser l'algorithme.

Toutefois, la non-unicité avec deux solutions isolées est inquiétante aussi du point de vue numérique ; les approches par optimisation que nous nous proposons d'étudier dans cette thèse ne bénéficieront pas de bonnes propriétés comme la convexité puisque l'ensemble des solutions peut ne pas être convexe. On peut donc craindre que même le calcul de l'une des solutions, lorsqu'il en existe, soit un problème difficile. Au sujet des difficultés provoquées par la multiplicité de solutions dans les problèmes de mécanique non-régulière, on pourra consulter [Mor06].

Non-existence

On prend maintenant les valeurs suivantes dans (1.38) : $\nu = 2$, $l = 1$, $g = -1$ (cette fois la gravité est dirigée vers le bas), $h = 2/3$, et $v_0 = 0$. On ne fixe pas les valeurs de u_0, μ et θ, et on suppose encore que $h_0 < l$. Les données (1.38) deviennent

$$M = 1,\ \ f = \sin\theta,\ \ H = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix},\ \ w = \begin{bmatrix} u_0 \\ 0 \end{bmatrix}. \tag{1.42}$$

et le problème incrémental (1.33) devient

$$\left\{ \begin{array}{lcll} v & = & \cos(\theta)\, r_x + \sin(\theta)\, r_y - \sin\theta & \\ u_x & = & \cos(\theta)\, v + u_0 & \text{et} \quad (u, r) \in \mathcal{C}(e_y, \mu). \\ u_y & = & \sin(\theta)\, v & \end{array} \right. \tag{1.43}$$

Lemme 1.3. *Le problème (1.43) possède une solution unique lorsque*

$$u_0 \leq 0 \quad ou \quad [u_0 > 0 \quad et \quad \tan\theta > \mu]$$

et aucune solution sinon.

Démonstration. Cherchons de nouveau les solutions par inspection des trois cas possibles dans la loi de Coulomb (1.22). (1) Décollage : $r = 0$ donc $v = -\sin\theta$, donc $u_{\mathrm{N}} = -\sin(\theta)^2 < 0$, impossible. (2) Adhérence : $u = 0$ est impossible. (3) Glissement : $u_{\mathrm{N}} = 0$ donc $v = 0$ et $u_{\mathrm{T}} = u_0$. Si $u_0 \leq 0$ alors $r_{\mathrm{T}} \geq 0$ doit être au bord du cône.

$$\left\{ \begin{array}{lcl} \cos(\theta)\, r_x + \sin(\theta)\, r_y & = & \sin(\theta) \\ r_x - \mu\, r_y & = & 0 \end{array} \right. \tag{1.44}$$

dont la solution unique vérifie bien $r_y = \tan(\theta)/(\tan(\theta) + \mu) > 0$; c'est une solution acceptable. Sinon, si $u_0 > 0$, alors $r_{\mathrm{T}} \leq 0$ doit être au bord du cône. On obtient le système linéaire

$$\left\{ \begin{array}{lcl} \cos(\theta)\, r_x + \sin(\theta)\, r_y & = & \sin(\theta) \\ r_x + \mu\, r_y & = & 0. \end{array} \right. \tag{1.45}$$

La solution de ce dernier système est $r_y = \frac{\tan(\theta)}{\tan(\theta) - \mu}$ pour $\tan\theta \neq \mu$ (sinon, pas de solution). Cette valeur pour r_y n'est solution que si elle est strictement positive, autrement dit le problème possède une solution si et seulement si $\tan\theta > \mu$. ∎

Ceci est conforme à l'intuition : lorsque $\tan(\theta) > \mu$, le couple exercé par la force de frottement agit sur la barre dans le sens anti-horaire, et permet de compenser la force de gravité qui tend à faire pénétrer la barre dans le sol. Dans le cas limite où $\tan(\theta) = \mu$, la force de contact exerce un couple nul sur la barre et ne joue aucun rôle. Enfin, lorsque $\tan(\theta) < \mu$, la force de contact exerce un couple dans le sens horaire, qui ne compense pas l'effet de la gravité mais au contraire l'aggrave. Il n'est donc pas possible de trouver une force de contact qui empêche la barre de pénétrer dans le sol, et le problème n'a pas de solution.

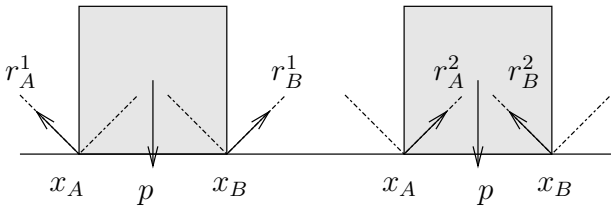


FIG. 1.13 – Un problème d'équilibre statique avec un continuum de solutions

Comme pour la non-unicité, cette situation de non-existence remet en cause notre modèle et ses hypothèses (corps rigides, loi de Coulomb, etc). En pratique, il arrive que les algorithmes dédiés à la résolution du problème incrémental (1.33) échouent, sans qu'il soit possible de dire si le problème n'admet pas de solution (dans ce cas, il faut remettre en cause notre modèle) ou s'il en existe (au moins) une mais que l'algorithme n'a pas été capable de la trouver (dans ce cas, il faut améliorer notre algorithme). Dans les deux cas, la question de savoir *a priori* si le problème (1.33) admet une solution est importante. Dans le chapitre 3, on donnera un critère théorique pour répondre à cette question.

1.4.5 Non-unicité avec continuum de solutions

Un deuxième problème très simple (deux contacts, trois degrés de liberté, en dimension 2) illustre la possibilité de solutions multiples au problème (1.33) ; mais cette fois, il existe un segment de solutions, qui ne sont donc pas isolées. Il suffit de considérer l'équilibre statique d'un carré rigide soumis à son poids p selon $-Oy$ et posé sur un plan en dimension 2 comme sur la figure 1.13. On suppose que le plan exerce deux forces r_A et r_B en x_A et x_B avec $\mu > 0$ en ces points. En discrétisant l'équation du mouvement comme dans la sous-section (1.4.4), on trouve que la solution en force n'est pas unique. Toutes les combinaisons convexes des couples de forces (r_A^1, r_B^1) et (r_A^2, r_B^2) de la figure 1.13 sont solutions. Plus généralement, dès qu'un corps rigide possède deux points de contact frottant ($\mu > 0$) avec un objet extérieur (comme le sol) ou avec un autre corps rigide, ce phénomène de coincement devient possible et il faut s'attendre à un continuum de solutions.

Du point de vue numérique, cette situation est tout aussi inquiétante que la non-unicité de la sous-section 1.4.4 où les deux solutions étaient isolées. Par exemple, si l'on calcule les forces de contact en cherchant un zéro d'une certaine fonction, la matrice jacobienne de cette fonction ne sera certainement pas inversible sur le continuum de solutions et on ne pourra pas utiliser la méthode de Newton.

1.4.6 Caractère NP-complet

Cette sous-section est adaptée de [Bar91]. On considère le système de la figure 1.14. Le socle rectangulaire de masse m_s coulisse sans frottement selon l'axe Ox et subit une force extérieure $2F_0$ dirigée selon $-Ox$. Les $k+2$ barres rigides identiques de longueur l portent une masse ν à leur extrémité x_{B_i}, $i \in 1, \dots k$ (ou x_{B_g}, x_{B_d} pour les barres

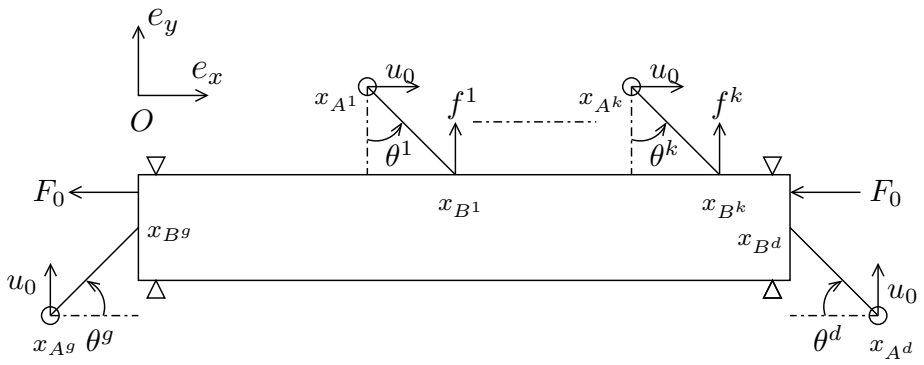


FIG. 1.14 – Un problème de contact difficile à résoudre !

de gauche et de droite). Elles sont articulées par une liaison pivot parfaite située à leur extrémité supérieure A_i (ou A_g, ou A_d) à un point de mouvement imposé à vitesse u_0 ; ce point se déplace selon Ox pour les barres d'indices $i = 1, \dots k$ et selon Oy pour les deux autres. Les barres d'indice $i = 1, \dots k$ sont soumises à des forces extérieures f^i dirigées vers le haut exercées en x_{B_i}. Les deux dernières barres ne sont soumises à aucune force extérieure. L'équation différentielle du mouvement pour la barre d'exposant $i = 1, \dots k$ est

$$\nu l^2 \ddot{\theta}^i = \cos(\theta)\, r_x^i + \sin(\theta)\, (r_y^i + f^i) \tag{1.46}$$

avec des relations similaires pour les barres de gauche et de droite, et l'équation du mouvement du socle est

$$m_s\, \ddot{x}_s = -\sum_{i=1}^{l} r_x^i - r_x^g - r_x^d - 2F_0. \tag{1.47}$$

On discrétise ces équations comme à la sous-section 1.4.4 en introduisant un pas de temps h et en approximant l'accélération angulaire $\ddot{\theta}^i$ de la barre d'indice i par $(v^i - v_0^i)/h$ où v^i approxime $\dot{\theta}^i$ au pas de temps courant et v_0^i l'approxime au pas de temps précédent ; le même traitement est appliqué à v^g et v^d. La vitesse du socle selon Ox étant notée v^s, le problème incrémental (1.33) met en jeu

- la vitesse généralisée $v := (v^1, \dots, v^k, v^g, v^d, v^s)$,
- les forces de contact $r := (r^1, \dots r^k, r^g, r^g)$ exercées par le socle sur les barres ; les barres exercent en retour des forces $(-r^1, \dots - r^k, -r^g, -r^d)$ sur le bloc,
- et les vitesses relatives $u := (u^1, \dots u^k, u^g, u^g)$ des barres par rapport au bloc aux points de contact.

On prend les valeurs numériques suivantes : $\nu = 1/2$, $m_s = 1$, $l = \sqrt{2}$, $\theta = \pi/4$, $1 = 1$, $h = 1$ et $\mu = 2$ à tous les contacts avec des conditions initiales nulles $v_0^i = v_0^g = v_0^d = v_0^s = 0$. Les valeurs des forces $(f^1, \dots, f^k, F_0) \in \mathbb{N} \setminus \{0\}$ ne sont pas prescrites.

L'équation dynamique discrétisée est

$$\left\{\begin{array}{ll} v^i = r_x^i + r_y^i + f^i & i = 1, \dots k \\ v^g = -r_x^g + r_y^g & \\ v^d = r_x^d + r_y^d & \\ v_s = -\sum_{i=1}^k r_x^i - r_x^g - r_x^d - 2F_0. & \end{array}\right. \tag{1.48}$$

L'équation cinématique discrétisée est

$$\left\{\begin{array}{l} u_x^i = v^i - v^s + 1 \\ u_y^i = v^i \end{array}\right. (i = 1, \dots k), \quad \left\{\begin{array}{l} u_x^g = v^g - v^s \\ u_y^g = v^g + 1 \end{array}\right. , \text{ et } \quad \left\{\begin{array}{l} u_x^d = v^g - v^s \\ u_y^d = v^g + 1 \end{array}\right. . \tag{1.49}$$

On ajoute la loi de Coulomb à chaque contact

$$(u^i, r^i)_{i=1,\dots k} \in \mathcal{C}(e_y, 2), \quad (u^g, r^g) \in \mathcal{C}(-e_x, 2), \quad (u^d, r^d) \in \mathcal{C}(e_x, 2). \tag{1.50}$$

Lemme 1.4. *Soit $f_1, \dots f_k, F_0 \in \mathbb{N} \setminus \{0\}$. Le problème incrémental défini par l'équation dynamique (1.48), l'équation cinématique (1.49) et la loi de frottement (1.50) a une solution si et seulement si*

$$\exists\ (\delta_1, \dots \delta_k) \in \{0,1\}^k \ : \ \sum_{i=1}^k \delta_i\, f_i = F_0. \tag{1.51}$$

Démonstration. Supposons qu'une solution (v, u, r) au problème incrémental (1.48)-(1.49)-(1.50) existe. On s'intéresse au contact en B^g. (1) Décollage : $r^g = 0$ donc $v^g = 0$ et $u_g = (-v^s, 1)$ qui est acceptable seulement si $v^s \geq 0$. (2) Adhérence : $u^g = 0$ donc $v^g = -1$ et $v^s = 1 \geq 0$. (3) Glissement : $u_x^g = 0$ donc $v^g + v^s = 0$. Si $v^s \leq 0$, on a $u_y^g = 1 - v^s > 0$, r_y^g est strictement négatif et au bord du cône, donc r^g vérifie $-r_x^g + r_y^g = -v^s$ et $2r_x^g - r_y^g = 0$, dont la solution est $r_x^g = -v_s > 0$: la force de contact n'est pas répulsive, c'est absurde, donc $v_s \geq 0$. Dans les trois cas (1-2-3), on a nécessairement $v^s \geq 0$. Le même raisonnement sur le contact symétrique en A^g montre que $v^s \leq 0$, donc $v^s = 0$. Revenons au contact en x_{B^g} : la solution (1) en décollage est possible avec $v^g = 0$, $u^g = (0, 1)$ et $r^g = 0$; la solution (2) en adhérence n'est pas possible ($v^s \neq 0$) ; et la solution (3) en glissement n'est pas possible (on a montré que $v^s = 0$, or $v^s \leq 0$ mène à une force de contact non-répulsive). On a donc nécessairement $v^s = 0$, $v^g = 0$ et de la même manière $v^d = 0$.

On s'intéresse maintenant au contact $i \in 1, \dots n$. (1) Décollage : $r^i = 0$ donc $v^i = f^i$ et $u^i = (1 + f^i, f^i)$, ceci est acceptable. (2) Adhérence : $u^i = 0$ est impossible car $v^s = 0$. (3) Glissement : $u_y^i = 0$ donc $v^i = 0$ et $u_x^i = 1$. La force de contact r^i doit vérifier $r_x^i + 2r_y^i = 0$ et $r_x^i + r_y^i = -f^i$ dont la solution est $r^i = (-2f^i, f^i)$ qui est acceptable ($r_y^i > 0$ car $f^i > 0$). Il y a donc deux solutions possibles, vérifiant $r_x^i = -2\,\delta^i\, f^i$ avec $\delta^i \in \{0, 1\}$.

Enfin, intéressons-nous à l'équation dynamique du socle : $0 = v^s = -\sum_{i=1}^k r_x^i - 2F_0 = 2\sum_{i=1}^k \delta^i\, f^i - F_0$ donc $\sum_{i=1}^k \delta^i\, f^i = F_0$. Ainsi, si (v, u, r) est solution du problème incrémental (1.48)-(1.49)-(1.50), alors le k-uplet $(\delta^i)_{i=1\dots k}$ défini par $\delta^i := -\frac{r_x^i}{2f^i}$ est solution du

problème (1.51). Réciproquement, si $(\delta^i)_{i=1\dots k}$ est solution de (1.51), la solution (v, u, r) construite ci-dessus est solution du problème incrémental. ■

Le problème de décision (1.51) s'appelle le problème de « somme de sous-ensembles » (*subset sum* en anglais) et on peut montrer qu'il est NP-complet. Si on savait résoudre en temps polynomial le problème (1.48)-(1.49)-(1.50), on saurait aussi résoudre (1.51) en temps polynomial. Il est donc peu probable que ce soit possible : le problème incrémental, en toute généralité, est donc extrêmement difficile.

Remarque 1.5. Le problème de la figure 1.14 est très proche de l'infaisabilité : une perturbation infime de la barre de gauche (par exemple si le socle s'en approchait, même avec un $v_s < 0$ minuscule, ou si on exerçait sur elle un tout petit couple négatif en laissant le socle immobile) produit une situation de non-existence similaire à celle de la sous-section 1.4.4. La barre de droite est dans la même situation. Ce problème est donc mal posé, au sens où une perturbation infime des données suffit à le rendre infaisable, et il n'est pas surprenant qu'il soit particulièrement difficile.

1.4.7 Difficultés prévisibles

Récapitulons : le problème (1.33) peut être un programme quadratique ordinaire, et il contient comme cas particulier le problème « subset sum » qui est NP-complet. Il peut admettre zéro, une, plusieurs ou une infinité de solutions, dont certaines peuvent être isolées et d'autres former un continuum. Toutes ces propriétés sont décourageantes, et c'est probablement pour cela que de nombreuses modifications et simplifications du modèle de Coulomb ont été proposées (certaines sont exposées dans la section suivante 1.5). Cependant, la riche structure du problème incrémental (1.33) permet d'envisager de nombreuses méthodes de résolution, et l'expérience prouve que ces méthodes fonctionnent bien en pratique et sont souvent capables de fournir une solution, même pour des problèmes de grande taille. En effet, les problèmes réels ne sont pas, en général, aussi pathologiques que l'exemple de la figure 1.14 ci-dessus dont on a montré que sa résolution revenait à résoudre un problème NP-complet. Malgré toutes les difficultés théoriques qui viennent d'être citées, nous affronterons donc directement le problème (1.33).

1.5 Problèmes et travaux connexes

Plusieurs problèmes de mécanique du contact sont proches de celui qui nous intéresse. Par exemple, d'autres modèles de frottement que la formulation (1.22) de la loi de Coulomb peuvent être utilisés. Ou alors, on peut s'intéresser seulement à l'équilibre statique du système, ou son évolution quasi-statique, en négligeant les effets inertiels. Cette section passe donc en revue quelques problématiques similaires à la nôtre.

1.5.1 Autres modèles dits « de Coulomb »

En raison des nombreuses difficultés théoriques illustrées dans la section précédente, et parce qu'il n'existe pas d'algorithme capable de trouver à coup sûr une solution au problème incrémental (1.33) lorsqu'il en existe, ce dernier est considéré comme un problème difficile. Dans les applications, en particulier en informatique graphique où la vitesse et la robustesse des algorithmes sont des critères déterminants, de nombreux modèles simplifiés plus ou moins similaires à la formulation impulsion-vitesse (1.22) ont été proposés afin de rendre le problème incrémental plus facile.

Lödstedt [Löt84] propose un modèle basé sur la programmation quadratique. En remplaçant le cône de Coulomb par un cône polyédral (en dimension 3), Anitescu et Potra [AP97] proposent un modèle de frottement qui se formule comme un LCP bien structuré qui peut être résolu efficacement. Baraff [Bar91] propose un algorithme qui « imite » d'une certaine manière le comportement attendu de la loi de Coulomb. La formulation de Kaufmann [KEP05] aboutit à un problème incrémental qui est un QP convexe, comme dans le cas sans frottement. De façon générale, les approximations de la loi de Coulomb appartiennent souvent à l'une des deux catégories suivantes : celles qui la transforment pour faire apparaître de la convexité (de manière à ce que la force de frottement dérive d'un potentiel, on parle alors de frottement *associé*) et celles qui la linéarisent en remplaçant le cône du second ordre par un cône polyédral.

1.5.2 Statique et évolution quasi-statique

Si on ne s'intéresse pas à la dynamique mais à la statique ou l'évolution quasi-statique du système, la formulation dite « gap-force » de la loi de Coulomb est la plus similaire au modèle continu. Elle est formellement identique à la formulation (1.22), à ceci près que les vitesses sont remplacées par les déplacements et que la complémentarité est écrite non plus entre la force et la vitesse, mais entre la force et le *gap* défini comme la distance entre les deux objets en contact potentiel. Bien que les variables soient différentes, on aboutit au même problème incrémental et les mêmes algorithmes peuvent être utilisés.

1.5.3 Détection de collisions

Avant de s'attaquer au problème incrémental (1.33), il est nécessaire d'assembler les données H et w et de connaître les normales e^i aux points de contact. Une étape préalable à ces calculs est donc de détecter les contacts présents dans le système. Selon la représentation géométrique utilisée et la complexité du système mécanique, cette étape peut être extrêmement coûteuse. De nombreux algorithmes et implémentations sont déjà disponibles pour ce problème et nous ne nous y intéresserons pas.

1.5.4 Détection de sous-systèmes

S'il est possible de diviser le système en deux sous-systèmes indépendants (qui n'interagissent par aucune de force de contact ou d'un autre type), il est possible de résoudre le problème incrémental en découplant les différents sous-problèmes. Lorsque seules les

forces de contact sont couplantes, ceci revient à rechercher une division en composantes connexes du graphe des objets en contact.

1.5.5 Problème continu

Dans toute cette thèse, on s'intéresse au problème en temps discret, à l'existence de solutions au problème incrémental, et à la manière de les calculer le cas échéant. Le problème en temps continu est également intéressant, et les mêmes problèmes d'existence et d'unicité de trajectoire se posent avec les même conséquences sur la validité ou l'invalidité du modèle. Moreau [Mor99] expose une étude complète d'un problème académique mais très instructif, le « processus de rafle » (*sweeping process*, en anglais) qui illustre bien les difficultés théoriques que pose l'étude de l'équation différentielle du mouvement assortie d'une contrainte unilatérale dans l'espace des coordonnées généralisées et d'une loi de frottement non-régulière. En raison des impacts qui interviennent lorsqu'une contrainte unilatérale devient active, les trajectoires ne sont pas régulières. Plus précisément, on les suppose continues mais pas C^2, de sorte que les dérivées en temps doivent être prises au sens des distributions et les trajectoires cherchées dans des espaces adaptés.

Chapitre 2

Dynamique non-régulière

Ce chapitre est consacré aux équations de la mécanique non-régulière en temps continu, non pas dans le but d'en faire une étude théorique, mais pour expliquer les méthodes de construction pratique des données du problème incrémental grâce à la discrétisation en temps des équations continues.

2.1 Équations du mouvement

Plusieurs méthodes sont disponibles pour la mise en équations d'un système mécanique. La mécanique lagrangienne, très générale, est celle que nous exposerons le plus en détail, mais la méthode des « coordonnées maximales » et la méthode d'Euler seront aussi évoquées. D'autres techniques de mise en équation et d'autres approches, comme l'approche modale, sont aussi utilisables : il suffit que le système étudié possède un nombre fini de degrés de liberté, et que l'on soit capable d'écrire les équations dynamique et cinématique discrétisées. On pourra consulter [LB08] et [BL08] pour plus de détails.

2.1.1 Approche lagrangienne

Systèmes mécaniques lagrangiens réguliers

On considère un système mécanique S en dimension d avec un nombre fini m de degrés de liberté paramétrés par un vecteur $\mathbf{q}(t) \in \mathbb{R}^m$. Pour l'instant, seules les contraintes bilatérales parfaites (*i.e.*, sans frottement) sont autorisées dans le système. Les contraintes unilatérales, éventuellement avec frottement, seront introduites plus tard. Le système est soumis à des forces extérieures F_j quelconques (ponctuelles, volumiques, constantes ou pas, etc). On définit les *vitesses généralisées* par

$$\mathbf{v}(t) := \frac{d\mathbf{q}}{dt}(t) \tag{2.1}$$

et les équations du mouvement sont données par le formalisme de la mécanique lagrangienne, plus précisément par le principe des puissances virtuelles ; en notant $E_c(q, \dot{q})$

l'énergie cinétique du système et $F_j(s)$ les forces qui s'appliquent au point $M(s)$ l'équation de Lagrange s'écrit

$$\frac{\partial}{\partial t}\Big(\frac{\partial E_c}{\partial \dot{q}}\Big) - \frac{\partial E_c}{\partial q} = \sum_j \int_{M\in S} \Big(\frac{\partial M}{\partial q}\Big)^\top dF_j(M) \tag{2.2}$$

et sa forme générale, tous calculs faits (c'est-à-dire après avoir explicité les dérivées et les intégrales dans (2.2)) est

$$\mathbf{M}(\mathbf{q}(t))\frac{d\mathbf{v}}{dt}(t) + \mathbf{N}(\mathbf{q}(t),\mathbf{v}(t)) + \mathbf{F}_{\text{int}}(t,\mathbf{q}(t),\mathbf{v}(t)) = \mathbf{F}_{\text{ext}}(t). \tag{2.3}$$

La matrice $\mathbf{M}(\mathbf{q})$, appelée *matrice de masse*, contient les masses et les moments d'inertie ; dans la plupart des applications, on a $\mathbf{M}(\mathbf{q}) \in \mathbf{S}_m^{++}$ (autrement dit, $\mathbf{M}(\mathbf{q})$ est symétrique définie positive). Le vecteur $\mathbf{F}_{\text{int}}$ représente les forces intérieures et le vecteur $\mathbf{F}_{\text{ext}} : \mathbb{R} \to \mathbb{R}^n$ représente les forces extérieures. Enfin le vecteur $\mathbf{N}(\mathbf{q},\mathbf{v})$ contient des termes non-linéaires appelés accélérations gyroscopiques. Dans la suite, l'équation du mouvement (2.3) sera notée sous la forme plus condensée

$$\mathbf{M}(\mathbf{q}(t))\frac{d\mathbf{v}}{dt}(t) + \mathbf{F}(t,\mathbf{q}(t),\mathbf{v}(t)) = 0, \tag{2.4}$$

où le vecteur $\mathbf{F}$ rassemble les termes $\mathbf{N}$, $\mathbf{F}_{\text{int}}$ et $-\mathbf{F}_{\text{ext}}$.

Systèmes mécaniques lagrangiens non-réguliers

Les contraintes unilatérales peuvent être introduites dans le système et modélisées soit en utilisant des multiplicateurs de Lagrange, soit les forces de contact en chaque point où une contrainte unilatérale est active ; nous adopterons la seconde approche.

En chaque point de contact $i \in 1,\dots n$, on étiquette les corps impliqués par A^i et B^i et on note $M_A^i = M_B^i$ le point de contact, avec M_A^i appartenant au corps A et M_B^i au corps B. Supposons que les corps A et B appartiennent tous les deux au système S (par opposition à la situation où l'un des corps est un objet extérieur à la simulation, comme un mur fixe jouant le rôle de conditions aux limites). Les paramètres $\mathbf{q}$ peuvent alors être utilisés pour déterminer la position de M_A^i et M_B^i

$$\exists f_A^i,\, f_B^i \,:\, \mathbb{R}^m \longrightarrow \mathbb{R}^d \,:\, M_A^i(t) = f_A^i(\mathbf{q}(t)),\;\; M_B^i(t) = f_B^i(\mathbf{q}(t)). \tag{2.5}$$

En dérivant par rapport au temps, on peut déterminer leur vitesse

$$\frac{dM_A^i}{dt} = \text{Jac}[f_A^i](\mathbf{q}(t))\;\mathbf{v}(t), \quad \frac{dM_B^i}{dt} = \text{Jac}[f_B^i](\mathbf{q}(t))\;\mathbf{v}(t) \tag{2.6}$$

(où Jac[·] est l'opérateur Jacobien) et finalement la vitesse relative $\mathbf{u}^i(t)$ de B^i par rapport à A^i au point de contact

$$\mathbf{u}^i(t) = \frac{dM_B^i}{dt} - \frac{dM_A^i}{dt} = (\text{Jac}[f_B^i] - \text{Jac}[f_A^i])(\mathbf{q}(t))\;\mathbf{v}(t). \tag{2.7}$$

Inversement, l'un des objets (disons A^i) peut être extérieur à la simulation avec un mouvement imposé. Par exemple, il peut s'agir d'un mur fixe, ou d'un tapis roulant de vitesse constante, etc. Dans ce cas, la vitesse de M_A^i ne dépend pas de $\mathbf{q}$, elle est imposée par une fonction extérieure

$$M_A^i(t) = f_A^i(t) \tag{2.8}$$

et la vitesse relative $\mathbf{u}^i(t)$ de B^i par rapport à A^i au point de contact est

$$\mathbf{u}^i(t) = \frac{dM_B^i}{dt} - \frac{dM_A^i}{dt} = \mathrm{Jac}[f_B^i](\mathbf{q}(t))\ \mathbf{v}(t) - \frac{d}{dt} f_A^i(t). \tag{2.9}$$

Que les corps appartiennent tous les deux au système, ou que l'un des deux soit un élément extérieur, la relation entre les vitesses relatives et les vitesses généralisées est linéaire ou affine. En rassemblant toutes les vitesses relatives locales $\mathbf{u}^i(t)(i = 1, \dots n)$ en un seul vecteur $\mathbf{u}(t) := (\mathbf{u}^1(t), \dots \mathbf{u}^n(t))$, cette relation peut être récrite sous forme matricielle

$$\mathbf{u}(t) = \mathbf{H}(q(t))\,\mathbf{v}(t) + \mathbf{w}(t) \tag{2.10}$$

où le second membre $\mathbf{w}(t)$ est nul lorsque les éventuels objets extérieurs à la simulation sont immobiles.

On introduit la force[1] de contact locale $\mathbf{r}^i(t)$ exercée par B^i sur A^i et on rassemble toutes les forces de contact locales dans un vecteur $\mathbf{r}(t) := (\mathbf{r}^1(t), \dots \mathbf{r}^n(t))$. Le principe des puissances virtuelles fournit la nouvelle équation du mouvement

$$\mathbf{M}(\mathbf{q}(t))\frac{d\mathbf{v}}{dt}(t) + \mathbf{F}(t, \mathbf{q}(t), \mathbf{v}(t)) = \mathbf{H}(t)^\top \mathbf{r}(t). \tag{2.11}$$

Les équations (2.10)-(2.11) (cinématique et dynamique), associées à la loi de Coulomb continue (sous-section 1.2.4 du chapitre 1), forment le problème en temps continu.

2.1.2 Coordonnées maximales et multiplicateurs de Lagrange

Dans l'approche lagrangienne présentée dans la sous-section précédente, on utilise autant de paramètres que le système possède de degrés de libertés. Pour cette raison, on appelle ces paramètres les *coordonnées réduites*, ou coordonnées généralisées. La méthode des *coordonnées maximales* consiste au contraire à utiliser plus de coordonnées que le nombre de degrés de liberté, et à imposer des forces de liaisons (appelées multiplicateurs de Lagrange) pour maintenir les contraintes satisfaites. Par exemple, considérons en dimension 2 un simple pendule rigide maintenu par une liaison pivot parfaite à son extrémité.

- On peut prendre comme paramètre l'angle θ que fait le pendule avec une direction donnée. C'est une approche lagrangienne, avec un seul paramètre (une coordonnée réduite). La liaison pivot est automatiquement satisfaite et la force de liaison est éliminée.

[1] Comme le mouvement est non-régulier, ce devrait être une impulsion, modélisée par une distribution et non une fonction, pour autoriser les impacts ; mais cette discussion ne fait pas partie de nos objectifs.

– On peut aussi prendre trois paramètres pour orienter le pendule (par exemple, deux pour la position du centre de la barre et un pour l'angle qu'elle fait avec une direction donnée). Pour maintenir la liaison pivot, on doit imposer des efforts de liaison, inconnus *a priori*, entre la barre et son point d'ancrage. Par rapport à l'approche lagrangienne, on a donc deux paramètres en plus (trois au lieu d'un) et une contrainte bilatérale de dimension deux à satisfaire, ce qui impose d'ajouter deux multiplicateurs de Lagrange.

Les avantages et inconvénients respectifs des deux approches sont bien discutés dans [LB08, BL08, Bar96] ; ce dernier article contient notamment un éclairage intéressant sur l'influence bénéfique de l'approche par multiplicateurs sur la simplicité et la modularité du code informatique qui doit être produit pour implémenter la simulation. Les deux paragraphes suivants résument ce comparatif.

L'approche lagrangienne, pour commencer, nécessite *a priori* des calculs de taille plus réduite : en effet, les forces de liaisons bilatérales parfaites sont éliminées automatiquement des équations, et le nombre de paramètres est réduit au nombre de degrés de liberté, il y a donc moins d'inconnues que dans l'approche par multiplicateurs. Lorsqu'un système possède un grand nombre de coordonnées maximales et peu de degrés de libertés (système très contraint), l'approche lagrangienne est donc *a priori* préférable. De plus, elle impose les contraintes de manière intrinsèque, au sens où – par construction – aucun point dans l'espace des configurations ne viole les contraintes bilatérales parfaites. Au contraire, dans l'approche par multiplicateurs, seul un sous-ensemble de l'espace des configurations est admissible, et c'est le rôle des multiplicateurs de Lagrange de forcer le système à rester dans ce sous-ensemble. Au fur et à mesure que les erreurs s'accumulent (erreurs dues à l'intégration en temps essentiellement, et dans une moindre mesure, erreurs d'arrondi), la configuration courante s'éloigne de l'ensemble admissible. Ce phénomène de dérive doit alors être compensé par des techniques de stabilisation qui compliquent nettement l'intégration en temps. De plus, pour limiter la dérive, on peut être amené à choisir des pas de temps plus petits que ceux autorisés par l'approche lagrangienne.

L'approche par coordonnées maximales et multiplicateurs de Lagrange, de son côté, est plus facile à mettre en oeuvre car elle évite le travail qui consiste à trouver un bon jeu de coordonnées réduites. Ce travail nécessite une certaine expertise, ou l'utilisation d'outils de calcul symbolique, et dans certaines situations (comme une contrainte de type « deux surfaces lisses déformables doivent glisser sans frottement l'une sur l'autre ») il est presque impossible de trouver un tel jeu de coordonnées réduites. Selon [Bar96], il est aussi plus facile d'exploiter le caractère creux de la matrice de masse en utilisant des coordonnées maximales. Enfin, l'approche par multiplicateurs permet de gérer des contraintes non-holonomes, contrairement à l'approche lagrangienne.

Si l'on choisit d'utiliser les coordonnées maximales au lieu des coordonnées réduites, ou un mélange des deux, cela revient simplement à ajouter des degrés de liberté au système et des multiplicateurs de Lagrange pour maintenir les contraintes bilatérales. On peut remplacer le vecteur des vitesses généralisées v par (v, λ) où le vecteur λ représente

les multiplicateurs introduits, et la matrice de masse discrétisée devient de la forme

$$M = \begin{bmatrix} \bar{M} & J^\top \\ J & 0 \end{bmatrix} \tag{2.12}$$

où $\bar{M} \in \mathbf{S}_m^{++}$ et J est une matrice qui exprime les contraintes bilatérales linéarisées en v. Les lignes de J sont supposées linéairement indépendantes, ce qui signifie que les contraintes linéarisées ne sont ni redondantes, ni contradictoires. Cette modification (2.12) de la matrice de masse maintient son caractère symétrique et inversible, mais *pas* son caractère défini positif.

2.1.3 Approche eulérienne

Cas d'application

La méthode d'Euler diffère essentiellement de la méthode de Lagrange par le paramétrage du champ de vitesses du système : au lieu de définir les vitesses généralisées comme les dérivées par rapport au temps des paramètres choisis pour décrire le système (les coordonnées généralisées), on les définit par la donnée de la vitesse d'un point particulier du solide, et de son vecteur rotation. Cette méthode ne s'applique donc qu'aux solides rigides.

Lorsque le système considéré comporte des solides rigides non contraints par des liaisons bilatérales (ou peu contraints, ou « simplement » contraints, de sorte qu'on soit capable d'exprimer les forces de liaison), la méthode d'Euler peut être préférable à la méthode de Lagrange pour écrire les équations du mouvement. D'une part, elle permet d'obtenir une matrice de masse particulièrement simple (diagonale et constante au cours du temps), et d'autre part la mécanique eulérienne se prête bien au paramétrage de l'orientation du solide par les quaternions au lieu des traditionnels angles d'Euler. Ceci est avantageux car

- les angles d'Euler permettent de représenter toutes les *orientations* d'un solide (avec redondance, une orientation donnée correspondant à plusieurs valeurs des angles) ;
- mais la donnée des angles d'Euler et de leur dérivée temporelle ne permet pas de représenter toutes les *vitesses instantanées* d'un solide ; pour certaines valeurs des angles d'Euler, il existe des vecteurs-rotation qui ne peuvent être représentés par aucune valeur de la dérivée temporelle des angles d'Euler ; autrement dit, les angles d'Euler sont satisfaisants pour représenter une simple orientation (statique) mais pas pour représenter une orientation et un vecteur rotation.

Ceci se traduit en pratique par une singularité artificielle (au sens où elle est introduite par le paramétrage, et n'appartient pas intrinsèquement au problème mécanique) qui rend la matrice de masse semi-définie positive mais pas définie positive (ce problème est parfois appelé *gimbal lock* en anglais, et « perte d'un degré de liberté » en français). Lorsque l'on ne peut pas s'assurer que les solides resteront éloignés de ces configurations singulières, comme c'est le cas pour l'orientation d'un satellite ou celle d'un grain dans un tas de sable, il n'est donc pas recommandé d'utiliser les angles d'Euler.

Méthode d'Euler

Pour étudier le mouvement d'un solide rigide S de masse m dans l'espace tridimensionnel, on considère un référentiel R défini par son origine O et une base orthonormée $B := (u_1, u_2, u_3)$ et un second référentiel R' défini par O' et $B' := (u'_1, u'_2, u'_3)$. Le référentiel R est supposé inertiel (galiléen), tandis que R' est un référentiel que l'on va choisir lié au solide et qui ne sera donc généralement pas inertiel. Le solide est considéré comme libre, au sens où il dispose de ses six degrés de liberté ; si le solide est contraint et que l'on veut tout de même utiliser l'approche eulérienne, il faut être capable d'exprimer les forces de contraintes et de les intégrer aux équations du mouvement en les considérant comme des forces extérieures.

Le solide possède six degrés de liberté, on va donc former une équation différentielle du mouvement de dimension six, sous la forme de deux blocs de dimension trois qui correspondent intuitivement au mouvement de translation et de rotation respectivement.

Les équations du mouvement de translation sont très simples ; on note a_G l'accélération du centre de gravité G du solide, et on considère l'équation de Newton (somme des force égale masse fois accélération), que l'on écrit symboliquement dans le référentiel inertiel R

$$\mathrm{d}f = (a_M)_{/R}\,\mathrm{d}m \tag{2.13}$$

où $\mathrm{d}m$ est la masse d'une particule, $(a_M)_{/R}$ son accélération dans le référentiel inertiel R et $\mathrm{d}f$ l'ensemble des forces (intérieures et extérieures) qui lui sont appliquées. On intègre (2.13) sur tout le solide S pour obtenir

$$\sum F_{\text{ext}} = \int_S \mathrm{d}f = \int_S (a_M)_{/R}\,\mathrm{d}m = m\,(a_G)_{/R} \tag{2.14}$$

où les forces de liaison internes au solide n'apparaissent pas car elles s'annulent deux à deux. L'équation (2.14) affirme que le mouvement du centre de gravité du solide S est le même que celui d'un point matériel qui serait positionné en G, affecté de toute la masse m du solide S et soumis aux forces extérieures F_{ext}. Elle constitue notre première équation différentielle de dimension trois, l'équation du mouvement en translation.

Considérons maintenant le mouvement de rotation ; on note ω le vecteur rotation de la base B' par rapport à la base B. La formule de Varignon, qui permet de relier la dérivée par rapport au temps d'un vecteur W quelconque dans les deux référentiels, s'écrit

$$\frac{\mathrm{d}}{\mathrm{d}t}(W)_{/B} = \frac{\mathrm{d}}{\mathrm{d}t}(W)_{/B'} + \omega \wedge W. \tag{2.15}$$

On note sous la forme d'un indice $/B$ ou $/B'$ la base dans laquelle le vecteur est dérivé par rapport au temps ; de même, on note $/R$ ou $/R'$ le référentiel dans lequel la position d'un point est dérivée par rapport au temps. Soit M un point mobile quelconque, dont la position dans le référentiel R est notée $r := OM$ et la position dans le référentiel R' est notée $r' := O'M$. Notons $(v_M)_{/R}$ la vitesse de M dans R, et $(v_M)_{/R'}$ sa vitesse dans R'. En appliquant cette formule à $W = r'$, on obtient la formule de composition des vitesses

$$(v_M)_{/R} = (v_{O'})_{/R} + \omega \wedge r' + (v_M)_{/R'} \tag{2.16}$$

et en l'appliquant à

$$W = (v_M)_{/R} - (v_{O'})_{/R} = \omega \wedge r' + (v_M)_{/R'}, \tag{2.17}$$

on obtient la formule de composition des accélérations

$$(a_M)_{/R} = (a_{O'})_{/R} + (a_M)_{/R'} + \frac{\mathrm{d}}{\mathrm{d}t}(\omega)_{/R'} \wedge r' + 2\,\omega \wedge (v_M)_{/R'} + \omega \wedge (\omega \wedge r'). \tag{2.18}$$

On choisit maintenant de lier le référentiel R' au solide S, et on prend pour M un point matériel de S. Ainsi, la position de M dans R' est constante, sa vitesse et son accélération dans R' sont donc nulles. La formule de composition des accélérations (2.18) devient

$$(a_M)_{/R} = (a_{O'})_{/R} + \frac{\mathrm{d}}{\mathrm{d}t}(\omega)_{/R'} \wedge r' + \omega \wedge (\omega \wedge r'). \tag{2.19}$$

Puis on utilise de nouveau la formule de Newton (2.13) et on intègre sur tout le solide S le moment des forces par rapport au point O. Autrement dit, on écrit :

$$\int_S r \wedge \mathrm{d}f = \int_S r \wedge (a_M)_{/R}\ \mathrm{d}m \tag{2.20}$$

et on remplace $(a_M)_{/R}$ par son expression simplifiée (2.19) pour obtenir, après quelques calculs et en notant I le tenseur d'inertie du solide S dans sa configuration actuelle, la formule

$$\int_S r \wedge \mathrm{d}f = m\,(O'G) \wedge (a_{O'})_{/R} + I\,\frac{\mathrm{d}}{\mathrm{d}t}(\omega)_{/R'} + \omega \wedge (I\,\omega). \tag{2.21}$$

Cette équation constitue notre deuxième équation différentielle de dimension trois, l'équation du mouvement en rotation. Si l'on choisit O' fixe dans R, ou si on prend $O' = G$, (2.21) se simplifie en

$$\int_S \mathrm{d}f \wedge r = I\,\frac{\mathrm{d}}{\mathrm{d}t}(\omega)_{/R'} + \omega \wedge (I\,\omega). \tag{2.22}$$

Notons symboliquement $\mathcal{M}_{\text{ext}}$ le moment des forces extérieures, et notons abusivement (*i.e.* en ne précisant pas le référentiel) $\dot{v}_G := \frac{\mathrm{d}}{\mathrm{d}t}(v_G)_{/R}$ et $\dot{\omega} := \frac{\mathrm{d}}{\mathrm{d}t}(\omega)_{/R'}$. En rassemblant (2.14) et (2.22), on obtient les équations d'Euler

$$\begin{cases} m\,\dot{v}_G & = \sum F_{\text{ext}} \\ I\,\dot{\omega} + \omega \wedge (I\,\omega) & = \sum \mathcal{M}_{\text{ext}} \end{cases} \tag{2.23}$$

qui déterminent le mouvement (il faut bien sûr leur ajouter une équation différentielle qui relie les vitesses (v_G, ω) aux paramètres choisis pour décrire le système ; ces paramètres sont typiquement la position du centre de gravité du solide, et des angles d'Euler ou un quaternion unité pour paramétrer son orientation). Comme l'orientation du solide S change au cours du mouvement, son tenseur d'inertie I change aussi. Cependant, la technique usuelle est de représenter les vecteurs qui interviennent dans les équations (2.21) ou (2.22) en les exprimant dans la base B' liée au solide ; de cette manière, la base (mobile) suit le solide dans son mouvement et, bien que le tenseur d'inertie change

au cours du mouvement, sa matrice relativement à la base B' ne change pas. On peut donc calculer la matrice du tenseur I une seule fois au début de la simulation, dans sa configuration de référence, et ne plus y revenir. Si de plus on choisit d'orienter la base B' selon les directions principales de S, alors la matrice d'inertie est diagonale. Par rapport à la méthode de Lagrange, la méthode d'Euler permet donc d'obtenir une matrice de masse constante et diagonale.

Calcul de l'orientation du solide

Les équations d'Euler gouvernent l'évolution du vecteur rotation $\omega_{R'/R}$ (noté simplement ω), à partir de laquelle on voudrait calculer l'évolution de l'orientation du solide. Pour représenter cette orientation, on peut utiliser

- les angles d'Euler (ψ, θ, ϕ),
- la matrice de passage Q de la base fixe B à la base liée au solide B',
- ou un quaternion unité q.

L'annexe B rappelle la définition des angles d'Euler, et l'annexe C contient les éléments essentiels de la théorie des quaternions, utiles pour paramétrer l'orientation d'un solide. On trouvera dans ces annexes la démonstration des trois formules suivantes, qui permettent de relier le vecteur rotation à la dérivée en temps de (ψ, θ, ϕ), Q et q respectivement. En notant $\underline{\Omega}$ le vecteur des coordonnées de ω dans la base mobile B', l'évolution de la matrice de passage Q de B à B' est gouvernée par

$$\dot{Q} = QS(\underline{\Omega}) \quad \text{avec} \quad S(\underline{\Omega}) := \begin{bmatrix} 0 & -\underline{\Omega}_3 & \underline{\Omega}_2 \\ \underline{\Omega}_3 & 0 & -\underline{\Omega}_1 \\ -\underline{\Omega}_2 & \underline{\Omega}_1 & 0 \end{bmatrix}. \tag{2.24}$$

La formule suivante (cf annexe B) relie les coordonnées $\underline{\omega}$ dans la base B du vecteur rotation $\omega_{B'/B}$ à la dérivée temporelle $\dot{a}$ des angles d'Euler $a := (\psi, \theta, \phi)$:

$$\underline{\omega} = M\dot{a} \quad \text{avec} \quad M_a = \begin{bmatrix} 0 & \cos\psi & \sin\theta\cos\psi \\ 0 & \sin\psi & -\sin\theta\cos\psi \\ 1 & 0 & \cos\theta \end{bmatrix}. \tag{2.25}$$

On constate que la matrice M est singulière lorsque $\sin\theta = 0$ (« gimbal lock »), ce qui fait que l'équation d'évolution des angles d'Euler (ψ, θ, ϕ) suivante peut ne pas avoir de sens

$$\dot{a} = M_a^{-1}\,\underline{\omega}. \tag{2.26}$$

Enfin, on trouve en annexe C la formule d'évolution du quaternion unité q qui correspond à la matrice Q :

$$\dot{q} = M_{\underline{\Omega}} q \quad \text{avec} \quad M_{\underline{\Omega}} = \frac{1}{2}\begin{bmatrix} 0 & -\underline{\Omega}_1 & -\underline{\Omega}_2 & -\underline{\Omega}_3 \\ \underline{\Omega}_1 & 0 & \underline{\Omega}_3 & -\underline{\Omega}_2 \\ \underline{\Omega}_2 & -\underline{\Omega}_3 & 0 & \underline{\Omega}_1 \\ \underline{\Omega}_3 & \underline{\Omega}_2 & -\underline{\Omega}_1 & 0 \end{bmatrix}. \tag{2.27}$$

Quel que soit le mode de représentation choisi, on peut ainsi retrouver l'orientation du solide à partir de l'évolution du vecteur rotation par intégration d'une équation différentielle ordinaire. Les avantages et inconvénients de chaque méthode sont discutés dans l'annexe C ; une première différence essentielle concerne le nombre de paramètres utilisés pour paramétrer l'orientation du solide : 3 avec les angles d'Euler (ce qui est minimal), 4 avec les quaternions et 9 avec les matrices de rotation (la représentation est donc redondante dans ces deux derniers cas). Les angles d'Euler présentent l'inconvénient d'introduire une singularité artificielle (le « gimbal lock »), tandis que les quaternions et les matrices de rotations souffrent du problème de dérive au cours de l'intégration ; en intégrant par exemple l'équation (2.24) de manière approximative avec un schéma numérique quelconque, il est probable que la matrice Q obtenue après un certain nombre de pas de temps ne sera plus orthogonale.

2.2 Discrétisation en temps

On s'intéresse maintenant à l'étape de discrétisation temporelle, qui permet de passer des équations dynamique et cinématique continues aux équations discrétisées qui forment le problème incrémental.

2.2.1 Méthode de Moreau

On considère un intervalle de temps $[t_i, t_f]$ de longueur δt. On suppose connues des approximations (q_i, v_i) des coordonnées et vitesses généralisées à t_i, et on voudrait calculer leurs approximations (q_f, v) au temps t_f. Les notation sont celles de la sous-section 2.1.1 ; on va traiter différemment les termes $\mathbf{M}$, $\mathbf{F}$, $\mathbf{H}$ et $\mathbf{w}$ d'une part (dont on suppose qu'ils varient de manière régulière) et les termes $\mathbf{v}$, $\mathbf{u}$ et $\mathbf{r}$ d'autre part, pour lesquels des sauts sont possibles.

Pour discrétiser les termes réguliers, on introduit le temps moyen $t_m := \frac{t_i+t_f}{2}$ et la position moyenne $q_m := q_i + \frac{\delta t}{2} v_i$ (cette formule suppose que les vitesses généralisées sont les dérivées temporelles des coordonnées généralisées, comme en mécanique lagrangienne ; lorsque ce n'est pas le cas, comme en mécanique eulérienne, on l'adapte). On approxime ensuite les termes réguliers $\mathbf{M}$, $\mathbf{F}$, $\mathbf{H}$ et $\mathbf{w}$ sur le pas de temps courant par leur valeur en (t_m, q_m) en posant

$$M := \mathbf{M}(q_m), \quad F := \mathbf{F}(t_m, q_m), \quad H = \mathbf{H}(t_m, q_m), \quad w = \mathbf{w}(t_m, q_m). \tag{2.28}$$

Puis on remplace l'accélération généralisée $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{v}(t)$ au cours du pas de temps courant par une différence finie

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{v}(t) \approx \frac{v_f - v_i}{\delta t} \tag{2.29}$$

et les variables non-régulières $\mathbf{v}$, $\mathbf{u}$ et $\mathbf{r}$ par leur valeur approchée (inconnue) v, u, r au temps t_f. En posant $f := F - \frac{M v_i}{\delta t}$, les équations (2.10)-(2.11) deviennent

$$Mv + f = H^\top r \tag{2.30}$$

$$u = Hv + w \tag{2.31}$$

qui sont exactement de la forme attendue dans (1.33).

Remarque 2.1. On remarque que $\mathbf{F}$ (plus précisément, le terme $\mathbf{F}_{\text{int}}$ qui représente les forces intérieures) peut dépendre de $\mathbf{v}$; comme on s'attend à des sauts dans les vitesses, donc dans les vitesses généralisées $\mathbf{v}$, il est possible que traiter $\mathbf{F}$ comme un terme régulier soit déraisonnable. L'introduction de cette difficulté supplémentaire rendrait le problème incrémental encore plus délicat, on préfère donc supposer (en le vérifiant pour chaque cas d'application) que le terme $\mathbf{F}$ peut raisonnablement être classé dans les termes réguliers, et on le note $\mathbf{F}(t_m, q_m)$ au lieu de $\mathbf{F}(t_m, q_m, v_m)$.

2.2.2 Précision des schémas de discrétisation

En raison de la nature non-régulière des fonctions et des lois qui interviennent dans l'équation du mouvement, il est délicat de définir une notion de solution continue, de préférence unique, par rapport à laquelle on pourrait mesurer l'erreur d'un schéma numérique et étudier son imprécision et sa dérive par rapport à la solution continue. En dehors de la technique empirique qui consiste à réaliser une simulation avec un très petit pas de temps et à la considérer comme « la » solution exacte pour étudier l'erreur commise lors des simulations avec un pas de temps plus grand, il semble que l'on dispose de peu de méthodes et de résultats pour étudier la précision des schémas de discrétisation.

2.3 Exemples

Quelques systèmes mécaniques avec contact et frottement sont décrits dans cette section afin de donner une idée des situations qu'il est possible de traiter, et de fournir des problèmes-tests pour les méthodes numériques qui feront l'objet des deux chapitres suivants.

2.3.1 Pendule double

La figure 2.1 décrit un premier exemple très simple de système mécanique avec contact unilatéral et frottement ; il s'agit d'un pendule double susceptible de frotter contre le sol immobile. L'annexe A contient la mise en équation et la discrétisation en temps de ce système ; elle illustre donc la technique générale de construction des données du problème incrémental.

2.3.2 Systèmes multicorps

Les systèmes composés de nombreux corps rigides, comme les milieux granulaires et les bâtiments en maçonnerie non-cohésive (figure 2.2), sont une application historique et importante de la mécanique du contact [Mor03]. Il existe aussi de nombreuses applications en informatique graphique. Lorsqu'il n'y a pas de force de liaison inconnue entre les corps rigides, les équations d'Euler décrites à la sous-section 2.1.3 suffisent à déterminer le mouvement. Il suffit de savoir calculer la matrice d'inertie de chacun des objets.

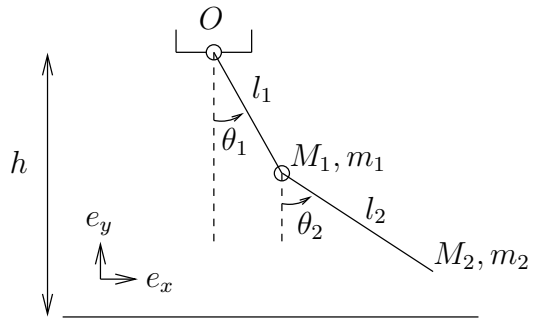


FIG. 2.1 – Un pendule double avec frottement

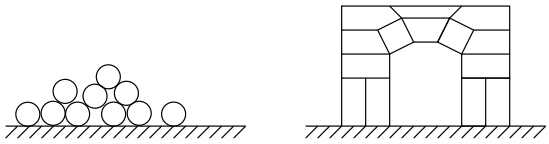


FIG. 2.2 – Milieu granulaire et maçonnerie non-cohésive faite de blocs rigides

2.3.3 Élasticité linéaire

Si l'on dispose d'un logiciel de calcul par la méthode des éléments finis, et que ce logiciel permet d'exporter les données qu'il produit (géométrie, matrice de masse, matrice de rigidité ...), on peut traiter facilement des problèmes de corps déformables en élasticité linéaire. Par exemple, on peut considérer un carré élastique qui glisse avec frottement sur un plan sous l'effet de forces extérieures (figure 2.3). Comme la matrice de masse et la matrice de rigidité sont alors constantes, il suffit de les faire assembler une seule fois et hors-ligne par le logiciel d'éléments finis. Ensuite, le plus simple est d'imposer des forces de frottement au niveau des noeuds du bord du maillage pour lesquels le contact unilatéral est actif. La détection de ces contacts actifs, et le calcul des normales aux points de contact peuvent être problématiques (on n'aborde pas ce problème) ; mais la discrétisation par le schéma de Moreau est ensuite directe et on aboutit de nouveau au problème incrémental (1.33).

2.3.4 Super hélices

Pour les expériences numériques, on a aussi utilisé le modèle de tige élastique dit des *super-hélices*, proposé par F. Bertails [Ber06] en informatique graphique pour représenter

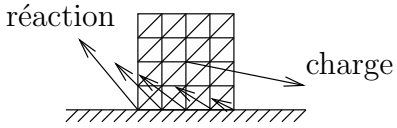


FIG. 2.3 – Un carré déformable tracté sur un plan

les structure fines comme les cheveux. Grâce au code déjà existant, les données du problème incrémental étaient déjà quasi disponibles ; cela nous a permis d'expérimenter les algorithmes proposés sur des exemples variés, tout en évitant une bonne partie du coût de développement.

2.3.5 Solution explicite

Pour certains systèmes simples, comme un point matériel glissant avec frottement de Coulomb sur un plan incliné, on peut facilement résoudre de manière exacte les équations du mouvement continues. Cependant, dès qu'on considère des systèmes mécaniques un peu plus complexes, il est rare qu'une solution analytique soit disponible. Dans [Spe75], une telle solution est présentée pour un problème de poinçon rigide cylindrique qui indente un demi-espace élastique linéaire. Cet exemple peut servir de référence pour comparer la solution obtenue par un code de calcul numérique à une solution exacte. Sinon, on peut comparer qualitativement ses résultats numériques à ceux déjà publiés, ou se contenter d'évaluer l'erreur *a posteriori* après la résolution du problème incrémental.

Chapitre 3

Résultat d'existence

Ce chapitre présente une nouvelle formulation du problème incrémental (1.33) comme un problème de point fixe non régulier. La nouveauté de l'approche est de traiter directement le modèle de frottement comme un problème d'optimisation convexe paramétrique sous contraintes de cône du second ordre, couplé avec une équation de point fixe. Cette formulation nouvelle nous permet de démontrer un résultat d'existence sous des hypothèses relativement faibles. Nous montrerons que ces hypothèses sont en fait nécessaires (et pas seulement suffisantes) dans deux situations pour lesquelles on peut étudier « à la main » l'existence de solutions. Il s'agit du cas sans frottement ($\mu = 0$ à tous les points de contact) et d'un exemple très simple inspiré du problème classique de Painlevé. On s'intéressera aussi aux techniques disponibles pour vérifier le critère en pratique.

3.1 Résultat d'existence de Klarbring et Pang

Dans [KP98], Klarbring et Pang donnent les résultats d'existence les plus avancés (à notre connaissance) pour le problème incrémental 1.33. La matrice de masse n'est supposée que *semi*-définie positive, et le cône de frottement n'est pas nécessairement du second-ordre ; lorsque ce cône est polyédral, un résultat spécifique (théorème 4.4 de [KP98]) est même disponible. Avec nos notations, et lorsque la matrice de masse M est définie positive (comme on le suppose ici, hypothèse 1.1), le théorème 4.1 de [KP98] se réduit à ce qui suit.

Théorème 3.1. *Si $w \in (\ker(H^\top) \cap L)^*$, alors il existe une solution au problème incrémental.*

Comme on a $(\ker(H^\top) \cap L)^* = \mathrm{cl}(\mathrm{Im}(H) + L^*)$, l'hypothèse du théorème 3.1 est automatiquement satisfaite si

$$\exists v \in \mathbb{R}^m \colon Hv + w \in L^*. \tag{3.1}$$

Cette hypothèse (3.1) va jouer un rôle central dans tout ce chapitre. Nous allons démontrer (théorème 3.19) un résultat d'existence un peu moins fort que celui de Klarbring et

Pang (car nous supposerons M définie positive, et parce que nous ferons une hypothèse de qualification supplémentaire), mais en utilisant des techniques différentes et plus simples (en particulier, nous n'utiliserons pas la notion d'inéquation quasi-variationnelle). De plus, l'étude théorique qui va suivre nous permettra de traduire la démonstration d'existence sous la forme d'un nouvel algorithme de résolution du problème incrémental.

3.2 Formulation par complémentarité conique

Dans cette section, on reformule le problème incrémental (1.33) du chapitre 1 comme un problème d'optimisation conique paramétrique couplé avec une équation de point fixe. On utilisera ensuite cette reformulation dans la section Section 3.3 pour établir le résultat d'existence.

On rappelle l'hypothèse 1.1 selon laquelle la matrice de masse M du problème incrémental est définie positive.

3.2.1 Contrainte de complémentarité conique

En s'inspirant du *bipotentiel* de De Saxcé [DSF98] évoqué en 1.3.3, on effectue le changement de variables $u^i \to \widetilde{u}^i$ défini pour tout $i = 1, \dots n$ par

$$s^i := \|u_{\mathrm{T}}^i\| \quad \text{et} \quad \widetilde{u}^i := u^i + \mu^i s^i e^i. \tag{3.2}$$

Remarque 3.2. Lorsque $\mu^i = 0$, ce changement de variables est sans effet. La « bonne » manière de faire consiste à n'introduire s^i que si $\mu^i > 0$, comme cela est fait dans l'article en préparation qui reprend les résultats de ce chapitre [ACLM09]. Cependant, cela introduit des difficultés techniques supplémentaires que nous éviterons ici en faisant simplement l'hypothèse suivante :

$$\forall i \in 1, \dots n \colon \mu^i > 0.$$

Tous les résultats à venir dans ce chapitre peuvent être adaptés au cas où il existe des contacts sans frottement.

En posant $\widetilde{u} := (\widetilde{u}^1, \dots, \widetilde{u}^n) \in \mathbb{R}^{nd}$, $s := (s^1, \dots, s^n) \in \mathbb{R}^n$ et $E := \mathrm{Diag}(\mu^i e^i)$, on récrit le changement de variables (3.2) sous la forme plus compacte

$$\tilde{u} = u + Es.$$

On va maintenant reformuler (1.33) en utilisant $\tilde{u}$ et s. D'une part, l'équation cinématique (1.31) se récrit

$$\widetilde{u} = H\, v + w + E\, s. \tag{3.3}$$

D'autre part, la loi de Coulomb (1.22) est équivalente à la contrainte de complémentarité conique (1.29) rappelée ici

$$\forall i,\ (K^i)^* \ni \tilde{u}^i \perp r^i \in K^i \iff L^* \ni \tilde{u} \perp r \in L$$

où le cône L est défini par $L := K_1 \times \cdots \times K_{\mathrm{N}} \subset \mathbb{R}^{nd}$ (les K_i étant les cônes de Coulomb aux n points de contact) et son cône dual L^* est $L^* = K_1^* \times \cdots \times K_{\mathrm{N}}^*$. Le problème incrémental est donc équivalent à : trouver $(v, \widetilde{u}, r, s)$ dans $\mathbb{R}^{m+(nd)+(nd)+n}$ tel que

$$\left\{ \begin{array}{l} M\,v + f = H^\top r \\ \widetilde{u} = H\,v + w + E\,s \\ L^* \ni \widetilde{u} \perp r \in L \\ s^i = \|\widetilde{u}_{\mathrm{T}}^i\|, \quad \text{pour tout } i \in 1, \ldots n. \end{array} \right. \tag{3.4}$$

L'idée est maintenant d'extraire de (3.4) le problème de complémentarité conique suivant, où $s \in \mathbb{R}^n$ est considéré comme un paramètre : trouver $(v, \widetilde{u}, r)$ dans $\mathbb{R}^{m+(nd)+(nd)}$ tel que

$$\left\{ \begin{array}{l} M\,v + f = H^\top r \\ \widetilde{u} = H\,v + w + E\,s \\ L^* \ni \widetilde{u} \perp r \in L. \end{array} \right. \tag{3.5}$$

Cette double manipulation (le changement de variables $u \to \widetilde{u}$ et le fait de fixer la valeur de $\|\widetilde{u}_{\mathrm{T}}^i\|$ sous la forme d'un paramètre s) est motivée par la remarque suivante : (3.5) constitue exactement les conditions d'optimalité d'un problème d'optimisation quadratique convexe posé sur un produit de cônes du second ordre (voir le théorème 3.5 ci-dessous).

3.2.2 Optimisation quadratique paramétrique du second ordre

On introduit la fonction quadratique

$$J(v) := \frac{1}{2} v^\top M v + f^\top v. \tag{3.6}$$

Sachant que M est définie positive, J est fortement convexe. Il se trouve que cette fonction joue un rôle intrinsèque dans notre problème ; comme le prouve le théorème 3.5, elle apparaît dans le problème d'optimisation associé à (3.5).

Définissons maintenant l'ensemble convexe ouvert

$$C(s) := \{v \in \mathbb{R}^m : Hv + w + Es \in \operatorname{int} L^*\} \tag{3.7}$$

et l'ensemble convexe fermé

$$\bar{C}(s) := \{v \in \mathbb{R}^m : Hv + w + Es \in L^*\}. \tag{3.8}$$

L'ensemble $\bar{C}(s)$ est exactement l'ensemble réalisable du problème (3.11) ci-dessous, et l'ensemble $C(s)$ est son intérieur. Ils seront tous les deux utilisés fréquemment dans la suite, en particulier à travers l'hypothèse

$$C(0) \neq \emptyset \tag{H}$$

et l'hypothèse plus faible

$$\bar{C}(0) \neq \emptyset. \tag{$\bar{H}$}$$

Une propriété cruciale est la monotonie de la multi-application $\bar{C}(\cdot)$.

Lemme 3.3. *Les multifonctions $C\colon \mathbb{R}^n \to \mathbb{R}^{nd}$ et $\bar{C}\colon \mathbb{R}^n \to \mathbb{R}^{nd}$ sont croissantes : pour $s, t \in \mathbb{R}^n$ tels que $s^i \leq t^i$ pour tout i, on a $C(s) \subset C(t)$.*

Démonstration. Soient $s, t \in \mathbb{R}^n$ tels que $s^i \leq t^i$ et $v \in C(s)$. On a donc par définition

$$Hv + w + Es \in \operatorname{int} L^*. \tag{3.9}$$

De plus, $(t^i - s^i)e^i \in K^i$ car $(t^i - s^i) \geq 0$ et $\mu^i e^i \in K^i$ par construction. On a donc aussi

$$E(t - s) \in L^*. \tag{3.10}$$

Combinant (3.9) et (3.10), on obtient (lemme III-2.1.6 de [HUL93], vol 1)

$$\frac{1}{2}(Hv + w + Es) + \frac{1}{2}E(t - s) = \frac{1}{2}(Hv + w + Et) \in \operatorname{int} L^*.$$

Par dilatation d'un facteur 2, ceci montre que $v \in C(t)$. Le même raisonnement fonctionne pour $\bar{C}(\cdot)$. ∎

Un corollaire immédiat du lemme 3.3 est le suivant.

Corollaire 3.4.

$$(H) \;\Rightarrow\; \forall s \in \mathbb{R}^n_+,\ C(s) \neq \emptyset$$
$$(\bar{H}) \;\Rightarrow\; \forall s \in \mathbb{R}^n_+,\ \bar{C}(s) \neq \emptyset.$$

Définissons maintenant le problème d'optimisation suivant

$$q(s) := \begin{cases} \min \quad J(v) & \text{(quadratique, fortement convexe)} \\ Hv + w + Es \in L^* & \text{(contraintes coniques)} \end{cases} \tag{3.11}$$

dont la valeur optimale définit la fonction

$$q\colon \begin{cases} \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\} \\ s \longmapsto \text{val}(3.11) \end{cases} \tag{3.12}$$

et le problème

$$p(s) := \begin{cases} \min \quad \frac{1}{2} r^\top (HM^{-1}H^\top) r + (w + Es - HM^{-1}f)^\top r & \text{(quadratique, convexe)} \\ r \in L & \text{(contraintes coniques)} \end{cases} \tag{3.13}$$

dont la valeur optimale définit la fonction

$$p\colon \begin{cases} \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{-\infty\} \\ s \longmapsto \text{val}(3.13). \end{cases} \tag{3.14}$$

On rappelle – voir l'équation (1.34) – qu'on a posé $W := HM^{-1}H^\top$ et $q := w - HM^{-1}f$, de sorte que la fonction objectif de (3.13) se récrit

$$\frac{1}{2}r^\top Wr + (q + Es)^\top r.$$

Sous l'hypothèse ($\bar{H}$), on définit aussi les fonctions

$$v\colon \begin{cases} \mathbb{R}^n_+ & \longrightarrow \mathbb{R}^m \\ s & \longmapsto v(s) := \operatorname{argmin}_{v\in\bar{C}(s)} J(v) \end{cases} \tag{3.15}$$

et

$$F\colon \begin{cases} \mathbb{R}^n_+ & \longrightarrow \mathbb{R}^n \\ s & \longmapsto (\|\widetilde{u}^1_{\mathrm{T}}(s)\|, \ldots, \|\widetilde{u}^n_{\mathrm{T}}(s)\|). \end{cases} \tag{3.16}$$

qui seront utiles dans le suite (à partir du lemme 3.8).

En fait, (3.11) et (3.13) sont duaux l'un de l'autre, et équivalents à (3.5).

Théorème 3.5 (Dualité et problèmes quadratiques)**.** *On suppose $M \in \mathbf{S}^{++}_m$ et on considère s tel que $C(s) \neq \emptyset$. Alors il existe une solution $(\widetilde{\bar{u}}, \bar{v}, \bar{r})$ de (3.5) et le couple $(\widetilde{\bar{u}}, \bar{v})$ est unique. Plus précisément, $\bar{v}$ est la solution unique de (3.11), $\widetilde{\bar{u}}$ est déterminé par (3.3) et $\bar{r}$ est une solution de (3.13).*

De plus, la valeur optimale finie $q(s)$ de (3.11) et la valeur optimale finie $p(s)$ de (3.13) satisfont

$$q(s) + p(s) + \frac{1}{2}f^\top M^{-1}f = 0. \tag{3.17}$$

Démonstration. L'existence d'une solution unique à (3.11) découle directement des hypothèses ; appliquons maintenant le théorème de dualité de Fenchel ([HUL93], 2.3.2 p.63) à la somme $J + i_{\bar{C}(s)}$ (où $i_{\bar{C}(s)}(\cdot)$ désigne la fonction indicatrice de $\bar{C}(s)$). Noter que $\operatorname{dom}(J) = \mathbb{R}^m$, donc l'hypothèse du théorème de Fenchel, à savoir

$$\operatorname{ri}\operatorname{dom}(J) \cap \operatorname{ri}\operatorname{dom}\big(i_{\bar{C}(s)}\big) \neq \emptyset$$

est satisfaite automatiquement. Le théorème affirme que

$$\min_{v\in\bar{C}(s)} J(v) = -\min_{z\in\mathbb{R}^m} J^*(z) + \big(i_{\bar{C}(s)}\big)^*(-z). \tag{3.18}$$

Des calculs directs donnent

$$J^*(z) = \frac{1}{2}(z-f)^\top M^{-1}(z-f).$$

D'autre part,

$$i_{\bar{C}(s)}(v) = i_{L^*}(Hv + w + Es).$$

On utilise le théorème relatif à la conjuguée de la pré-composition par une application affine ([HUL93] vol. 1, théorèmes 2.2.1 p.56 et 2.2.3 p.58) : sous l'hypothèse $C(s) \neq \emptyset$, on a

$$\begin{aligned}(i_{\bar{C}(s)})^*(-z) &= \min_{H^\top r=-z} (i_{L^*})^*(r) - (w+Es)\cdot r \\ &= \min_{H^\top r=-z} i_{-L}(r) - (w+Es)\cdot r \\ &= \min_{-r\in L, H^\top r=-z} -(w+Es)\cdot r \\ &= \min_{r\in L, H^\top r=z} (w+Es)\cdot r.\end{aligned} \tag{3.19}$$

On a donc, d'après (3.18)

$$\begin{aligned}\min_{v\in\bar{C}(s)} J(v) &= -\min_{z\in\mathbb{R}^m}[J^*(z) + \min_{r\in L, H^\top r=z} (w+Es)\cdot r] \\ &= -\min_{r\in L, H^\top r=z} J^*(z) + (w+Es)\cdot r \\ &= -\min_{r\in L} J^*(H^\top r) + (w+Es)\cdot r \\ &= -\min_{r\in L}[\frac{1}{2} r^\top H M^{-1} H^\top r + (w - HM^{-1}f + Es)^\top r] \;-\; \frac{1}{2} f^\top M^{-1} f.\end{aligned} \tag{3.20}$$

Ceci prouve que $p(s)$ est fini et démontre l'égalité (3.17). D'autre part, l'équation (3.21) ci-dessous est une condition nécessaire et suffisante d'optimalité pour (3.11) : v est optimal si et seulement si

$$-\nabla J(v) \in N_{\bar{C}(s)}(v). \tag{3.21}$$

Explicitons le membre de droite :

$$\begin{aligned}N_{\bar{C}(s)}(v) &= \partial\Big(i_{L^*}\circ(H\cdot + w + Es)\Big)v \\ &= H^\top \partial i_{L^*}(Hv + w + Es) \\ &= H^\top N_{L^*}(\widetilde{u})\end{aligned} \tag{3.22}$$

et en utilisant le fait que

$$N_{L^*}(\widetilde{u}) = -L \cap \widetilde{u}^\perp$$

on voit que (3.21) est équivalente à

$$Mv + f \in H^\top(L\cap\widetilde{u}^\perp) \quad \text{avec} \quad \widetilde{u} = Hv + w + Es. \tag{3.23}$$

Autrement dit, v est (l'unique) solution de (3.11) si et seulement s'il existe $(r, \widetilde{u})$ tels que (3.5) soit vérifié. Ceci prouve l'unicité de v, donc celle de $\widetilde{u}$, dans (3.5). Considérons enfin une solution $\bar{r}$ de (3.13), et posons

$$\bar{v} := M^{-1}(H^\top \bar{r} - f) \quad \text{et} \quad \bar{\widetilde{u}} := W\bar{r} + q + Es = H\bar{v} + w + Es.$$

L'optimalité de $\bar{r}$ dans (3.13) implique $\bar{\tilde{u}} \perp \bar{r}$ donc

$$\bar{r} \in L \cap \bar{\tilde{u}}^{\perp},$$

et comme $M\bar{v} + f = H^{\top}\bar{r}$ on voit que $\bar{v}$ satisfait (3.23) : c'est donc la solution unique de (3.11). ■

Le théorème ci-dessous, qui est maintenant évident au vu du théorème 3.5, affirme que résoudre (3.4), donc le problème incrémental, revient à résoudre l'équation de point fixe (3.24) ci-dessous.

Théorème 3.6. *Soit $s \in \mathbb{R}^n_+$; on suppose que $C(s) \neq \emptyset$, donc $F(s)$ est bien définie par (3.16). Alors s est un point fixe de F, c'est-à-dire*

$$F(s) = s \tag{3.24}$$

si et seulement si le quadruplet $(v, r, \tilde{u}, s)$ (qui n'est pas unique) associé à s par le théorème 3.5 est solution de (3.4).

Remarque 3.7 (Manipulation pratique des contraintes du second ordre)**.** En utilisant des astuces classiques (voir [BTN01] et le chapitre 4), les deux problèmes (3.11) et (3.13) peuvent être reformulés comme des problèmes SOCP (« second order-cone programs ») standards, c'est-à-dire des problèmes d'optimisation de la forme

$$\begin{cases} \min c^{\top}x & \text{(fonction-objectif linéaire)} \\ Ax = b & \text{(contraintes affine)} \\ x \in \prod_j K_j & \text{(contraintes coniques)} \end{cases}$$

où les K_j sont soit des cônes du second ordre, soit des cônes polyédriques. Le problème SOCP a été beaucoup étudié [AG03] et on dispose de codes efficaces pour le résoudre [Stu99]. On peut aussi envisager d'utiliser des algorithmes spécifiques qui exploitent la structure de (3.11) et (3.13) [Kuc07]. On voit aussi que dans le cas bidimensionnel ($d = 2$), (3.11) et (3.13) deviennent de simples programmes quadratiques (QP) pour lesquels on peut utiliser tous les algorithmes usuels et leurs nombreuses implémentations.

3.3 Existence d'une solution au problème incrémental

Dans cette section, on prouve le résultat d'existence principal. La contribution essentielle de ce chapitre est que, sous l'hypothèse ($\bar{H}$), le problème de point fixe (3.24) admet une solution. Cette hypothèse admet une interprétation mécanique simple : elle signifie qu'il doit être cinématiquement possible qu'en chaque point de contact i, la vitesse relative u^i appartienne à $(K^i)^*$ (voir la sous-section 3.4.1).

On prouve tout d'abord l'existence sous l'hypothèse (H) via une série de lemmes techniques (sous-section 3.3.1), visant à appliquer le théorème de Brouwer. Pour passer de (H) à ($\bar{H}$), on utilise ensuite un argument de perturbation.

Quelques exemples d'applications sont donnés dans la sous-section 3.4.

3.3.1 Existence d'un point fixe

Commençons par quelques lemmes, qui aboutissent à la démonstration de la continuité de F définie à l'équation (3.16) ci-dessous.

Lemme 3.8 (Caractère borné)**.** *Sous l'hypothèse ($\bar{H}$) ou l'hypothèse plus forte (H), les fonctions v et F définies par (3.15) et (3.16) sont bien définies et bornées.*

Démonstration. L'hypothèse ($\bar{H}$) assure l'existence d'un $\bar{v} \in \bar{C}(0)$. D'après le lemme (3.3), $\bar{C}(s)$ est non vide pour tout $s \geq 0$; (3.15) a donc une solution unique $v(s)$, et $J(v(s)) \leq J(\bar{v}) < +\infty$. Par coercivité de J, ceci entraîne que $v(\cdot)$ est bornée. Le caractère borné de F est ensuite immédiat car $\widetilde{u}_{\mathrm{T}}(s) = [Hv(s) + w]_{\mathrm{T}}$. ∎

Lemme 3.9 (Convexité de la valeur optimale)**.** *Sous l'hypothèse ($\bar{H}$), la fonction $q\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ définie par (3.12) est convexe sur $\mathbb{R}^n$, et continue sur $(\mathbb{R}^*_+)^n$.*

Démonstration. On considère le sous-ensemble de $\mathbb{R}^n \times \mathbb{R}$ défini par

$$\Gamma := \big\{(s,t) \in \mathbb{R}^n \times \mathbb{R} : \quad \exists v \in C(s)\colon J(v) \leq t\big\}.$$

On va montrer que Γ est convexe et fermé. Pour démontrer la convexité, prenons $\alpha \in [0,1]$, (s_1,t_1) et (s_2,t_2) dans Γ et considérons v_1 et v_2 qui leur sont associés. En utilisant la convexité de L^* et J, on voit que $v_\alpha := \alpha v_1 + (1-\alpha)v_2$ satisfait $v_\alpha \in C(\alpha s_1 + (1-\alpha)s_2)$ et $J(v_\alpha) \leq \alpha t_1 + (1-\alpha)t_2$; ainsi, $\alpha(s_1,t_1) + (1-\alpha)(s_2,t_2)$ appartient à Γ.

On prouve maintenant que Γ est fermé. Considérons une suite $(s_k,t_k) \in \Gamma$ qui converge vers $(\bar{s},\bar{t}) \in \mathbb{R}^n \times \mathbb{R}$, et une suite associée $(v_k)_k$. Pour k assez grand, on a $t_k \leq 2\bar{t}$, donc $J(v_k) \leq 2\bar{t}$; ainsi la suite $(v_k)_k$ est incluse dans un ensemble de sous-niveau de J. Par coercivité de J, on peut extraire une sous-suite $(v_{k'})_{k'}$ qui tend vers $\bar{v} \in \mathbb{R}^m$. En passant à la limite $k' \to +\infty$ dans

$$v_{k'} \in C(s_{k'}) \quad \text{et} \quad J(v_{k'}) \leq t_{k'},$$

et en utilisant la fermeture de L^* et la continuité de J, on voit que $(\bar{s},\bar{t})$ appartient à Γ (et est associé à $\bar{v}$). Ainsi, Γ est un ensemble convexe fermé.

On observe maintenant que $q(s)$ peut être récrit de la manière suivante

$$q(s) = \min\{t : (s,t) \in \Gamma\}.$$

Ainsi q est la fonction *lower-bound* de Γ ; le théorème IV-1.3.1 de [HUL93] montre que q est convexe et semi-continue inférieurement sur $\mathbb{R}^n$.

Sous l'hypothèse ($\bar{H}$), q est finie sur $\mathbb{R}^n_+$ car son ensemble réalisable n'est pas vide. Ainsi $\mathbb{R}^n_+$ est inclus dans le domaine de q, et cette fonction est donc continue sur $(\mathbb{R}^n_+)^*$ d'après le théorème IV.2.1.2 de [HUL93] (qui affirme qu'une fonction convexe est continue sur l'intérieur relatif de son domaine). ∎

Remarque 3.10. Une autre manière de voir la convexité de q est la suivante : la fonction p définie par (3.14) est clairement concave (c'est un min, sur une famille indexée par r, de fonctions affines – donc concaves – de s ; or on sait qu'un min de fonctions concaves est concave). D'après l'équation (3.17), q est donc convexe ; mais cet argument ne tient que si $C(s) \neq \emptyset$ (cette hypothèse est faite dans le théorème 3.5) alors que la démonstration qui vient d'être donnée pour le lemme 3.9 ne nécessite pas cette hypothèse.

Lemme 3.11 (Continuité de la solution optimale)**.** *Sous l'hypothèse (H), la fonction v définie par (3.15) et la fonction F définie par (3.16) sont continues sur $\mathbb{R}^n_+$.*

Démonstration. Soit $\bar{s} \in \mathbb{R}^n_+$ et $\mathbb{R}^n_+ \ni s_k \to \bar{s}$; on veut montrer que $v(s_k) \to v(\bar{s})$. Soit $\bar{v}$ un point d'accumulation de la suite $(v(s_k))_k$ (qui est bornée, d'après le lemme 3.8) ; considérons une sous-suite $(v(s_{k'}))_{k'}$ telle que $v(s_{k'}) \to \bar{v}$.

En passant à la limite $k \to \infty$ dans l'inclusion $Hv(s_k) + w + Es_k \in L^*$ et en utilisant le fait que H est continue et L^* est fermé, on obtient : $H\bar{v} + w + E\bar{s} \in L^*$. Ceci prouve que $\bar{v}$ est réalisable pour (3.11) avec le paramètre $\bar{s}$.

D'autre part, la continuité de J implique

$$q(s'_k) = J(v(s_{k'})) \to J(\bar{v}).$$

Par continuité de q (3.9), on a aussi $q(s'_k) \to q(\bar{s})$, donc $J(\bar{v}) = q(\bar{s}) = J(v(\bar{s}))$.

On vient donc de montrer que $\bar{v}$ est la solution unique de (3.11) avec le paramètre $\bar{s}$: $\bar{v} = v(\bar{s})$. Ainsi $v(\bar{s})$ est l'unique point d'accumulation de la suite $(v(s_k))_k$ qui, de plus, est bornée par le lemme 3.8. Ceci prouve la continuité de v. Celle de F découle immédiatement de sa définition (3.16) et de la définition (3.2) de $\widetilde{u}$. ∎

Nous sommes maintenant en mesure de démontrer notre résultat d'existence sous l'hypothèse (H).

Théorème 3.12 (Existence d'une solution)**.** *On rappelle que M est supposée définie positive (hypothèse 1.1). Si (H) est vérifiée, alors la fonction F définie par (3.16) admet au moins un point fixe. Autrement dit, le problème de point fixe (3.4) et par conséquent (au vu du théorème 3.6) le problème incrémental (1.33) possèdent une solution.*

Démonstration. La fonction $F(\cdot)$ est positive et le lemme 3.8 montre qu'elle est bornée. Soit $R \geq 0$ tel que $\operatorname{Im} F \subset \mathbb{R}^n_+ \cap B(0,R)$. En particulier, $F(\mathbb{R}^n_+ \cap B(0,R)) \subset \mathbb{R}^n_+ \cap B(0,R)$. Le lemme 3.11 affirme que sous l'hypothèse (H), $F(\cdot)$ est aussi continue. On peut donc appliquer le théorème du point fixe de Brouwer à F sur $\mathbb{R}^n_+ \cap B(0,R)$ pour obtenir le résultat. ∎

On démontre ainsi de manière assez simple l'existence de solution au problème incrémental sous l'hypothèse (H). La sous-section suivante est consacrée au passage à l'hypothèse ($\bar{H}$). Elle éclaire les problèmes qui peuvent survenir lorsqu'on se passe de la qualification assurée par (H) : la dualité conique entre (3.11) et (3.13) n'a plus nécessairement lieu, et une solution du problème de point fixe (3.24) peut ne pas correspondre à une solution du problème incrémental.

3.3.2 Argument de perturbation

Afin de généraliser le théorème 3.12, nous nous intéressons aux propriétés de continuité de la multi-application $\bar{C}(\cdot)$. Commençons par rappeler les différentes notions de continuité pour les multi-applications [HUL93, RW98].

La distance $d(x,S)$ d'un point x à un ensemble convexe fermé S est définie par $d(x,S) := \min_{s\in S} \|x-s\|$. On définit aussi l'*excès* d'un ensemble S_1 par rapport à un ensemble S_2 par

$$e_H(S_1/S_2) := \sup\{d(x,S_2),\ x\in S_1\}$$

et la *distance de Hausdorff* entre S_1 et S_2

$$\Delta_H(S_1,S_2) := \max\big(e_H(S_1/S_2), e_H(S_2/S_1)\big).$$

Une multi-application $S\colon D\subset\mathbb{R}^n \to \mathbb{R}^m$ est dite *semi-continue extérieurement*[1] en $\bar{x}\in D$ lorsque pour tout $\varepsilon>0$, il existe un voisinage N de $\bar{x}$ dans D tel que pour tout $x\in N$

$$S(x)\subset S(\bar{x})+B(0,\varepsilon) \qquad \text{ou encore} \qquad e_H(S(x)/S(\bar{x}))\le\varepsilon.$$

De la même manière, S est dite *semi-continue intérieurement* en $\bar{x}\in D$ lorsque pour tout $\varepsilon>0$, il existe un voisinage N de $\bar{x}$ dans D tel que pour tout $x\in N$

$$S(\bar{x})\subset S(x)+B(0,\varepsilon) \qquad \text{autrement dit} \qquad e_H(S(\bar{x})/S(x))\le\varepsilon.$$

De plus, S est dite *continue* en $\bar{x}$ quand elle est à la fois semi-continue intérieurement et extérieurement en $\bar{x}$. Enfin S est dite *fermée* quand son graphe est fermé, c'est-à-dire

$$\forall (x_k)_k\in D \text{ avec } x_k\to\bar{x},\ \forall (v_k)_k\in S(x_k) \text{ avec } v_k\to\bar{v},\ \bar{x}\in D \text{ et } \bar{v}\in S(\bar{x}) \tag{3.25}$$

et *bornée* lorsque $S(D)$ est un ensemble borné.

Il est facile de montrer qu'une multi-application fermée et bornée est semi-continue extérieurement. Par ailleurs, une fonction (ordinaire) continue sur un ensemble compact est uniformément continue sur cet ensemble ; le résultat suivant généralise cette propriété aux multi-applications.

Lemme 3.13. *Soit S une multi-application fermée, bornée et semi-continue intérieurement définie sur un ensemble compact C. Alors S est uniformément continue sur C :*

$$\forall\,\varepsilon>0, \exists\,\delta>0,\ \forall x,y\in C,\ \|x-y\|\le\delta \ \Rightarrow\ \Delta_H(S(x),S(y))\le\varepsilon.$$

Démonstration. On va montrer ce résultat par l'absurde. Supposons qu'il existe $\varepsilon>0$ tel que pour tout $k=1,2\ldots$ avec $\delta_k\to 0$ quand $k\to\infty$, il existe x_k et y_k avec $\|x_k-y_k\|\le\delta_k$ et $\Delta_H(S(x_k),S(y_k))>\varepsilon$. Ceci signifie que, pour tout k : soit il existe $u_k\in S(x_k)$ tel que $d(u_k,S(y_k))>\varepsilon$; soit il existe $v_k\in S(y_k)$ tel que $d(v_k,S(x_k))>\varepsilon$.

[1] La terminologie « semi-continue supérieurement » est plus traditionnelle.

Au moins l'une des deux suites u_k et v_k (disons u_k) est infinie, et quitte à changer les indices on peut supposer qu'elle est définie pour tout $k \in \mathbb{N}$. Ainsi on a

$$d(u_k, S(y_k)) > \varepsilon \quad \text{pour tout } k \in \mathbb{N}. \tag{3.26}$$

Comme C est compact et S est bornée (donc u_k est bornée), on peut supposer (quitte à extraire une sous-suite) que $x_k \to \ell$ et $u_k \to \bar{u}$. La multi-application S étant fermée par hypothèse, on a $\bar{u} \in S(\ell)$. Comme $\|x_k - y_k\| \to 0$, on a aussi $y_k \to \ell$. Pour k suffisamment grand, on a $d(\bar{u}, S(y_k)) \leq \varepsilon/3$ (grâce à la semi-continuité intérieure) et on a aussi $\|u_k - \bar{u}\| \leq \varepsilon/3$ (car $u_k \to \bar{u}$). Ainsi

$$d(u_k, S(y_k)) \leq d(u_k, \bar{u}) + d(\bar{u}, S(y_k)) \leq \frac{2\varepsilon}{3} < \varepsilon$$

ce qui contredit (3.26) et achève la démonstration. ∎

Appliquons ce résultat à notre multi-application $\bar{C}$. Pour la borner, on introduit la multi-application suivante

$$G\colon \begin{cases} \mathbb{R}^n_+ & \longrightarrow \Lambda \\ s & \longmapsto \bar{C}(s) \cap \Lambda, \end{cases} \tag{3.27}$$

où Λ est l'ensemble convexe et compact (voir le lemme 3.8)

$$\Lambda := \{v \in \mathbb{R}^m \ : \ J(v) \leq J(v(0))\}.$$

Commençons par une propriété élémentaire de G.

Lemme 3.14. *Le graphe de la multi-application G définie par (3.27) est convexe et fermé.*

Démonstration. Soient $(s_1, v_1), (s_2, v_2) \in \text{graph}(G)$ et $\alpha \in [0, 1]$. On a donc $u_1 := Hv_1 + w + Es_1 \in L^*$ et $u_2 := Hv_2 + w + Es_2 \in L^*$. La convexité de L^* implique $\alpha u_1 + (1-\alpha)u_2 \in L^*$. Autrement dit

$$H(\alpha v_1 + (1-\alpha)\, v_2) + w + E(\alpha s_1 + (1-\alpha)\, s_2) \in L^*,$$

donc $(\alpha v_1 + (1-\alpha)\, v_2) \in \bar{C}(\alpha s_1 + (1-\alpha)\, s_2)$. Par ailleurs, la convexité de Λ implique que $(\alpha v_1 + (1-\alpha)\, v_2) \in \Lambda$, donc $\alpha(s_1, v_1) + (1-\alpha)(s_2, v_2) \in \text{graph}(G)$. Ceci prouve que le graphe de G est convexe.

Soit maintenant $s_k \in \mathbb{R}^n_+$ avec $s_k \to \bar{s}$ et $v_k \in G(s_k)$ avec $v_k \to \bar{v}$. Par définition on a $Hv_k + w + Es_k \in L^*$; en vertu de la continuité de H et de la fermeture de L^*, on a $H\bar{v} + w + E\bar{s} \in L^*$. De plus Λ est fermé, donc $\bar{v} \in \Lambda$. Finalement, $\bar{v} \in G(\bar{s})$, donc le graphe de G est fermé. ∎

Passons à la semi-continuité intérieure de G qui est la partie la plus difficile, et utilise de façon cruciale le caractère monotone de la multi-application $\bar{C}$.

Lemme 3.15. *Sous l'hypothèse ($\bar{H}$), la multi-application G définie par (3.27) est semi-continue intérieurement sur $\mathbb{R}^n_+$.*

Démonstration. Soit $\bar{s} \in \mathbb{R}^n_+$ et $\varepsilon > 0$. Il suffit de montrer que

$$\exists \delta > 0 \ : \ \forall s \in \mathbb{R}^n_+, \quad \|s - \bar{s}\|_\infty \leq \delta \ \Rightarrow \ G(\bar{s}) \subset G(s) + B(0, \varepsilon).$$

Si $\bar{s} = 0$, ceci est évident car G est croissante. Sinon, soit $\chi := \min_i\{\bar{s}^i : \bar{s}^i > 0\} > 0$. Soit aussi $\bar{v} \in G(\bar{s}) \neq \emptyset$, et fixons $v_0 \in G(0) \neq \emptyset$; on peut supposer que $v_0 \neq \bar{v}$, sinon $\bar{v} \in G(0) \subset G(s)$ pour tout $s \geq 0$ et il n'y a rien à montrer. On va démontrer que la valeur

$$\delta := \min(\chi, \frac{\chi\varepsilon}{\|v_0 - \bar{v}\|}) > 0$$

convient ; noter que $\bar{s}_i - \delta \geq 0$ pour tout i tel que $\bar{s}_i > 0$. Formons maintenant une combinaison convexe

$$s_\alpha := (1 - \alpha)\, 0 + \alpha\, \bar{s} = \alpha\, \bar{s}$$

de $0 \in \mathbb{R}^m$ et de $\bar{s}$, où $\alpha \in [0, 1]$ est choisi de manière à avoir $s_\alpha \leq s$ (donc $G(s_\alpha) \subset G(s)$) pour tout $s \geq 0$ tel que $\|s - \bar{s}\|_\infty \leq \delta$ (fig. 3.1). On pose donc $\alpha := 1 - \delta/\chi$, ce qui assure

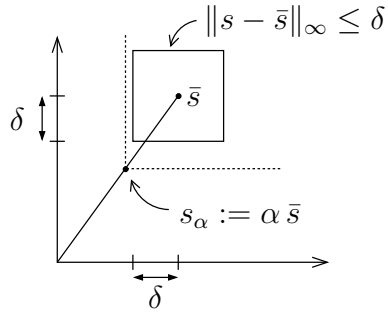


FIG. 3.1 – Choix de la valeur de α

bien $\alpha\bar{s}_i \leq \bar{s}_i - \delta$ pour tout i tel que $\bar{s}_i > 0$ (pour les i tels que $\bar{s}_i = 0$, on a évidemment $\alpha\bar{s}_i \leq s_i$ pour tout $s \geq 0$ quel que soit $\alpha \in [0, 1]$).

Posons maintenant $v_\alpha := (1 - \alpha)\, v_0 + \alpha\, \bar{v}$ (cette notation étant bien compatible avec celle de v_0). En vertu de la convexité du graphe de G (lemme 3.14), $v_\alpha \in G(s_\alpha)$. Ainsi, pour tout $s \geq 0$ tel que $\|s - \bar{s}\|_\infty \leq \delta$ on a $v_\alpha \in G(s)$.

Récapitulons ; on a fixé δ, puis pour tout $\bar{v} \in G(\bar{s})$, on a construit v_α qui appartient à $G(s)$ pour tout $s \geq 0$ tel que $\|s - \bar{s}\|_\infty \leq \delta$. De plus $\bar{v} = v_\alpha + (\bar{v} - v_\alpha)$ avec

$$\|\bar{v} - v_\alpha\| = (1 - \alpha)\|v_0 - \bar{v}\| = \frac{\delta}{\chi}\|v_0 - \bar{v}\| \leq \epsilon$$

ce qui achève la démonstration. ∎

Nous sommes maintenant en mesure de démontrer la continuité de G, qui sera utilisée dans la démonstration du lemme 3.17 ci-dessous.

Lemme 3.16. *La multi-application G définie par (3.27) est uniformément continue sur tout compact.*

Démonstration. D'après les lemmes 3.14 et 3.15, G est fermée et semi-continue intérieurement, et elle est manifestement bornée puisque $G(s) \subset \Lambda$ pour tout $s \in \mathbb{R}^n_+$. Le lemme 3.13 permet de conclure. ∎

Revenons à l'étude du problème incrémental. On suppose que les données H, w, E, μ satisfont l'hypothèse $(\bar{H})$ mais peut-être pas l'hypothèse (H). L'argmin correspondant $v(s)$ et la fonction F associée sont bien définis par (3.15) et (3.16), mais pour l'instant nous n'avons pas démontré que F est continue. Nous allons le faire grâce à un argument de perturbation qui nous permettra de revenir au cas précédemment étudié où l'hypothèse (H) est vérifiée. Une fois montrée la continuité de F, le théorème de Brouwer s'applique de la même manière.

Pour $\delta > 0$, soit $\Delta := (\delta, \dots, \delta)$ et $w_\delta := w + E\Delta$. Les données H, w_δ, E, μ satisfont l'hypothèse (H) (on rappelle qu'on a supposé $\mu^i > 0$ pour tout i, voir la remarque 3.2) de sorte que l'argmin correspondant $v_\delta(s)$ est bien défini et la théorie précédente s'applique : la fonction correspondante F_δ est continue. On va montrer que $F_\delta(s) \to F(s)$ uniformément par rapport à s sur un ensemble compact suffisamment grand, prouvant ainsi que F est continue sur ce compact. Pour commencer, on montre la convergence uniforme sur tout compact de $v_\delta(\cdot)$ vers $v(\cdot)$.

Lemme 3.17. *Sous l'hypothèse $(\bar{H})$, soient v et v_δ comme ci-dessus. Soit $D \subset \mathbb{R}^n_+$ un ensemble compact.*

$$\forall \epsilon > 0,\ \exists \bar{\delta} > 0,\ \forall \delta \in [0, \bar{\delta}],\ \forall s \in D :\ \|v_\delta(s) - v(s)\| \leq \epsilon.$$

Démonstration. Sachant que J est fortement convexe (hypothèse 1.1), il existe $\alpha > 0$ tel que pour tout $s \in \mathbb{R}^m_+$, pour tout $\delta > 0$ et pour tout $v \in \mathbb{R}^m$

$$J(v) \geq J(v_\delta(s)) + \nabla J(v_\delta(s)) \cdot (v - v_\delta(s)) + \alpha \|v - v_\delta(s)\|^2.$$

Notons alors $s_\delta := s + (\delta, \dots, \delta)$; si $v \in G(s) \subset G(s_\delta)$, alors l'optimalité de $v_\delta(s)$ implique $\nabla J(v_\delta(s)) \cdot (v - v_\delta(s)) \geq 0$. On a donc la condition de croissance

$$\eta := \alpha \|v_\delta(s) - v(s)\|^2 \leq J(v(s)) - J(v_\delta(s)). \tag{3.28}$$

De plus, J est lipschitzienne sur Λ avec une certaine constante de Lipschitz κ. Soit $\epsilon > 0$; on pose

$$\gamma := \frac{\alpha \epsilon^2}{\kappa}.$$

L'uniforme continuité de G sur le compact D (lemme 3.16) implique l'existence d'un $\bar{\delta} > 0$ tel que, pour tout $\delta \in [0, \bar{\delta}]$ et pour tout $s \in D$ tel que $s_\delta \in D$, on a

$$G(s_\delta) \subset G(s) + B(0, \gamma).$$

Comme $v_\delta(s) \in G(s_\delta)$, il existe donc un $\omega_\delta \in G(s)$ tel que $\|\omega_\delta - v_\delta(s)\| \leq \gamma$. Introduisons ω_δ dans (3.28) ; on obtient

$$\eta \leq [J(v(s)) - J(\omega_\delta)] + [J(\omega_\delta) - J(v_\delta(s))] \leq J(\omega_\delta) - J(v_\delta(s))$$

car $J(v(s)) \leq J(\omega_\delta)$ par définition de $v(s)$, et finalement

$$\eta \leq \kappa \|\omega_\delta - v_\delta(s)\| \leq \kappa\gamma$$

en utilisant la propriété de Lipschitz. Vu la définition de η, ceci démontre que

$$\|v_\delta(s) - v(s)\| \leq \sqrt{\frac{\kappa\gamma}{\alpha}} = \epsilon$$

et achève la démonstration. ∎

Ce résultat de convergence uniforme permet alors d'étendre le lemme 3.11 de continuité de F.

Lemme 3.18. *Soit $R \geq 0$ tel que $\operatorname{Im} F \subset B(0,R)$. Alors*

$$F_\delta(s) \underset{\delta \to 0}{\longrightarrow} F(s)$$

uniformément par rapport à $s \in \mathbb{R}^n_+ \cap B(0,R)$. En particulier, F est continue sur $\mathbb{R}^n_+$.

Démonstration. Pour tout $i \in \{1, \dots n\}$ on a

$$\begin{aligned}
|F^i_\delta(s) - F^i(s)| &= |\|\widetilde{u}^i_{\delta,\mathrm{T}}\| - \|\widetilde{u}^i_{\mathrm{T}}\|| \\
&\leq \|\widetilde{u}^i_{\delta,\mathrm{T}} - \widetilde{u}^i_{\mathrm{T}}\| \\
&= \|(\widetilde{u}^i_\delta - \widetilde{u}^i)_{\mathrm{T}}\| \\
&\leq \|\widetilde{u}^i_\delta - \widetilde{u}^i\| \\
&\leq \|\widetilde{u}_\delta - \widetilde{u}\| \\
&= \|Hv_\delta(s) + w + E(s+\Delta) - (Hv(s) + w + Es)\| \\
&= \|H(v_\delta(s) - v(s)) + E\Delta\| \\
&\leq \|H\| \, \|v_\delta(s) - v(s)\| + \|E\| \, \|\Delta\|.
\end{aligned} \tag{3.29}$$

Comme $v_\delta(s)$ converge vers $v(s)$ uniformément par rapport à s (avec $\|s\| \leq R$) quand δ tend vers zéro, ceci prouve la convergence uniforme de F_δ vers F sur $\mathbb{R}^n_+ \cap B(0,R)$, donc la continuité de F sur cet ensemble. Comme R peut être choisi arbitrairement grand, F est continue sur $\mathbb{R}^n_+$. ∎

On a donc la généralisation suivante du théorème 3.12, qui affirme l'existence d'une solution au problème incrémental sous l'hypothèse faible ($\bar{H}$).

Théorème 3.19 (Existence d'un point fixe)**.** *Sous l'hypothèse 1.1 et si ($\bar{H}$) est vérifiée, la fonction F définie par (3.16) admet au moins un point fixe.*

Démonstration. F est continue d'après le lemme 3.18 et bornée d'après le lemme 3.8. La démonstration est identique à celle du théorème 3.12. ∎

La question est maintenant de savoir si la solution s du problème de point fixe correspond à une solution du problème incrémental (autrement dit, si on peut appliquer le théorème 3.6 bien que l'hypothèse $C(s) \neq \emptyset$ ne soit pas satisfaite).

Théorème 3.20. *On suppose que ($\bar{H}$) est satisfaite ; soit v tel que $u := Hv + w \in L^*$. D'après le théorème 3.19, la fonction F définie par (3.16) admet (au moins) un point fixe s. Si $s > 0$, ou plus généralement si $u^i \in \operatorname{int} \Lambda^*_{\mu^i}$ pour tous les indices i tels que $s^i = 0$, alors il existe r tel que $(v(s), Hv(s) + w, r)$ soit solution du problème incrémental.*

Démonstration. La démonstration est identique à celle du théorème (3.6) : si $u^i \in \operatorname{int} \Lambda^*_{\mu^i}$ pour tous les indices i tels que $s^i = 0$, alors l'optimum $v(s)$ dans (3.11) est qualifié et il existe une solution duale r. ∎

Dans le cas où l'un des s^i est nul et la qualification n'a pas lieu pour le cône correspondant, alors on ne peut pas conclure : il se peut qu'aucun r (solution duale) correspondant au couple (v, u) (solution primale) n'existe, comme le prouve l'exemple suivant.

Exemple 3.21. Soit $m = 2$, $n = 1$, $d = 3$; la composante normale est la troisième ($e = (0, 0, 1)$) et le coefficient de frottement est $\mu = 1$. On considère les données suivantes :

$$M = I_2,\ f = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},\ w = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

L'hypothèse (H) est satisfaite : $\exists v = (0, 0)$: $Hv + w = 0 \in L^*$; les fonctions v et F sont bien définies par (3.15) et (3.16). En $s = 0$, on voit facilement que

$$v := \left\{ \begin{array}{l} \operatorname{argmin}(\frac{1}{2}\|v\|^2 + v_2) \\ v_1 \geq 0, v_2 = 0 \end{array} \right. = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

de sorte que $v(0) = 0$, et $F(0) = 0$: la fonction F admet un point fixe en 0. Pourtant, ce point fixe ne *correspond pas* à une solution du problème incrémental : en effet,

$$Mv(s) + f = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \notin H^\top L$$

car on voit facilement que $H^\top L = (0, 0) \cup \mathbb{R}^+_* \times \mathbb{R}$.

Ce exemple montre que lorsque seule l'hypothèse faible est satisfaite, l'équation de point fixe peut posséder des solutions qui ne correspondent pas à des solutions du problème incrémental. Dans la mise en oeuvre pratique de cette technique, si on obtient numériquement un point fixe de F qui ne satisfait pas les hypothèses du théorème 3.19, il faut donc vérifier *a posteriori* l'existence de la solution duale r. En pratique, les problèmes arriveront alors dès l'évaluation de la fonction F au point litigieux, car les algorithmes usuels échoueront à résoudre le problème (3.11) (pourtant bien posé, mais non qualifié). Le chapitre suivant discute le problème de la reconstruction de la solution duale r à partir de la résolution du problème primal (3.11) (voir la remarque 4.20 et la sous-section 4.7.3).

3.4 Applications

Dans cette section, on applique le critère d'existence[2] à quelques situations simples afin d'illustrer son « mode d'emploi » et de montrer son large champ d'application ; en particulier, on verra dans les sous-sections 3.4.2 et 3.4.3 deux exemples dans lesquels le critère d'existence est non seulement suffisant, mais aussi nécessaire. Ce n'est malheureusement pas un cas général, comme le prouve la sous-section 3.4.4 : il arrive que le problème incrémental possède une solution sans que $(\bar{H})$ soit vérifié.

3.4.1 Interprétation mécanique

L'hypothèse $(\bar{H})$ exige qu'il soit cinématiquement possible qu'en chaque point de contact i, la vitesse relative appartienne à $(K_{\mu^i,e^i})^*$ (le cône dual du cône de frottement). Ceci est inattendu ; une hypothèse plus naturelle est la suivante

$$\exists v \in \mathbb{R}^m \ : \ \forall i \in 1,\dots n \ : \ [Hv + w]^i_{\mathrm{N}} \geq 0, \tag{3.30}$$

qui signifie simplement qu'il doit être possible d'empêcher la pénétration. Ceci n'est pas le cas, par exemple, pour le problème représenté sur la figure 3.2 où un disque rigide est écrasé entre le sol immobile et un plan mobile de vitesse $u_0 > 0$. Quand tous les

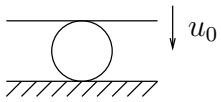


FIG. 3.2 – Il est impossible d'éviter la pénétration

coefficients de frottement sont nuls, l'hypothèse $(\bar{H})$ est équivalente à l'hypothèse (3.30) (sous-section 3.4.2), mais dans les autres cas elle est strictement plus exigeante : en effet, le fait qu'il soit cinématiquement possible d'éviter la pénétration ne garantit pas l'existence d'une solution, comme le prouve l'exemple de Painlevé vu au chapitre 1.

[2]Disons, celui de Klarbring-Pang, théorème 3.1, dont l'énoncé est plus simple et plus puissant que notre théorème 3.19.

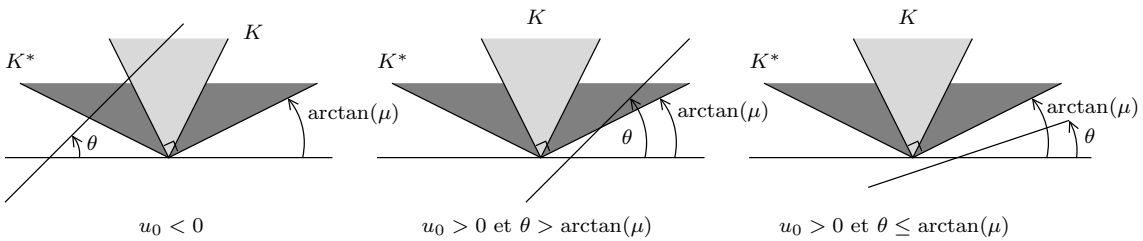


FIG. 3.3 – Application de notre critère à l'exemple de la barre de Painlevé

3.4.2 Cas sans frottement

Lorsque le coefficient de frottement μ^i est nul pour tous les contacts, la matrice E est la matrice nulle et les variables s et $\widetilde{u}$ disparaissent du problème (3.4). Si (v, u, r) est solution du problème (3.4), en particulier $v \in \bar{C}(0)$ donc $(\bar{H})$ est vérifiée. Ceci prouve que la condition $(\bar{H})$ est *nécessaire et suffisante* dans le cas sans frottement.

On voit que dans ce cas, le critère est trivial : il suffit de vérifier qu'il est cinématiquement possible d'éviter l'interpénétration (en effet, L^* est alors le demi-espace supérieur et $u = Hv+w \in L^*$ signifie simplement que toutes les vitesses normales u_{N}^i sont positives ou nulles). Il est évident que ce critère est nécessaire ; le théorème 3.19 affirme qu'il est aussi suffisant (ce que l'on peut voir directement en considérant le fait que dans le cas sans frottement, le problème incrémental est un QP, voir (1.36)).

3.4.3 Caractère nécessaire sur l'exemple de Painlevé

Pour l'exemple de la barre de Painlevé, déjà étudié à la sous-section 1.4.4, on sait calculer à la main les solutions du problème incrémental et on peut donc les comparer aux conséquences du théorème 3.19.

On a vu que le problème incrémental (1.43) possède une solution si et seulement si

$$u_0 \leq 0 \quad \text{ou} \quad [u_0 > 0 \quad \text{et} \quad \tan\theta > \mu]. \tag{3.31}$$

De son côté, le théorème 3.19 affirme qu'il existe une solution si

$$\exists\, v \in \mathbb{R}\ ;\ (u_0, 0) + (\cos\theta, \sin\theta)\, v \in K^*. \tag{3.32}$$

La figure 3.3 représente les trois situations possibles. On voit que la condition suffisante (3.32) est en fait équivalente à la condition nécessaire et suffisante (3.31) : sur cet exemple, le théorème 3.19 donne donc une condition nécessaire et suffisante.

3.4.4 Contre-exemple au caractère nécessaire

Les deux sous-sections précédentes suggèrent que l'hypothèse $(\bar{H})$ est peut-être une condition nécessaire et suffisante dans le cas général, mais il n'en est rien : l'exemple qui suit, qui est encore une variante de la barre de Painlevé, montre une situation où une solution au problème incrémental existe bien que cette hypothèse ne soit pas vérifiée.

Il suffit de reprendre le problème incrémental (1.40), dans lequel la gravité est dirigée vers le haut ($g = 1$). Par rapport à (1.43), seul le signe du terme constant f est modifié ($f = \pm\sin(\theta)$). Or, f n'intervient pas dans l'hypothèse ($\bar{H}$). La sous-section précédente montre que si $\tan(\theta) < \mu$, cette hypothèse n'est pas vérifiée ; pourtant, le lemme 1.2 affirme que deux solutions existent dans ce cas.

Remarque 3.22. L'hypothèse ($\bar{H}$) est purement cinématique et ne dépend ni de la matrice de masse M ni du terme f qui contient les forces et les conditions initiales. Le critère du théorème 3.19 n'utilise donc pas toute l'information disponible, et il n'est pas surprenant qu'il ne constitue qu'une condition suffisante en général.

3.4.5 Caractère intrinsèque

Indépendamment de la valeur de H et w, l'hypothèse ($\bar{H}$) est vérifiée lorsqu'il est cinématiquement possible d'imposer à chaque contact une vitesse relative qui appartienne à K_μ^*. En particulier, si on peut imposer une vitesse nulle ou verticale à chaque point de contact, le critère est vérifié et une solution existe. Ce critère ne concerne donc que la cinématique du système, et pas son paramétrage : il est *intrinsèque* au système. Par exemple, le cas où w est nul est rendu trivial par le théorème 3.19 : il suffit de prendre $v = 0$ dans ($\bar{H}$) (voir la sous-section suivante). En général, avec les paramétrages usuels, les systèmes mécaniques pour lesquels tous les objets extérieurs à la simulation sont immobiles vérifient $w = 0$. Cependant, si l'on prend un paramétrage qui dépend du temps, on peut avoir $w \neq 0$ même lorsque les objets extérieurs sont immobiles. Dans ce cas, prendre $v = 0$ ne suffit pas pour vérifier ($\bar{H}$). Cependant, grâce au caractère intrinsèque du critère, on voit qu'une solution existe (et qu'il suffirait de changer le paramétrage ou le référentiel pour avoir $w = 0$).

3.4.6 Objets extérieurs en mouvement de solide rigide

On vient de voir que le problème incrémental admet toujours une solution lorsque les objets extérieurs à la simulation sont tous immobiles ; plus généralement, supposons que les objets extérieurs sont animés d'un mouvement de solide rigide : il suffit d'imposer ce même mouvement aux objets internes à la simulation pour que toutes les vitesses relatives soient nulles, et que ($\bar{H}$) soit satisfaite. Ainsi, dans les trois situations classiques représentées sur la figure 3.4, *il existe une solution au problème incrémental à chaque pas de temps*. Sur cette figure, le premier schéma représente l'exemple très classique d'une grande quantité de solides (en général rigides) qui tombent d'un silo et forment un tas sur le sol. Le silo et le sol étant immobiles, le problème incrémental admet toujours une solution. Le deuxième schéma représente une expérience de plaque vibrante sur laquelle on a disposé des solides. La plaque étant animée d'un mouvement de solide rigide, une solution existe à chaque pas de temps. Le dernier schéma représente une expérience de « machine à laver » : des solides sont disposés dans un tambour qui tourne avec une vitesse imposée. Là encore, le seul objet extérieur (le tambour) est animé d'un champ de vitesse de solide rigide, donc une solution existe à chaque pas de temps.

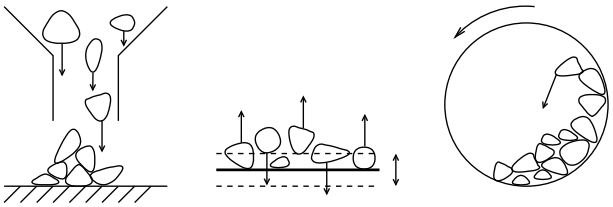


FIG. 3.4 – Trois situations classiques où le critère s'applique

3.4.7 Solides déformables

Supposons que le système est constitué d'un solide déformable discrétisé en espace, et que ses degrés de liberté correspondent à la position des noeuds d'un certain maillage. Supposons de plus que la détection de collisions soit faite au niveau des noeuds et qu'on ne déclare pas plus d'une collision par noeud. Alors, en chaque point de contact, il suffit de donner à chaque noeud une vitesse verticale (selon e^i) pour que (H) soit vérifiée. Ceci prouve qu'il existe alors une solution au problème incrémental *pour toute valeur de* μ.

Plus généralement, quand un système dispose de suffisamment de degrés de liberté pour qu'on puisse régler les vitesses relatives à tous les contacts indépendamment les unes des autres en jouant sur les degrés de liberté, on peut imposer une vitesse relative dirigée selon e^i et on obtient l'existence d'une solution à tous les pas de temps. Ce résultat d'existence inconditionnelle pour le problème discrétisé en temps et en espace lorsque le système est déformable est comparable à celui dû à Haslinger [Has83] dans le cas statique.

3.4.8 Conditions sur H

On peut donner des conditions suffisantes de vérifications de (H) et ($\bar{H}$) en considérant seulement la matrice H, ou plutôt $\mathrm{Ker}(H^\top)$. Si les lignes de H sont linéairement indépendantes, on a

$$\mathrm{Ker}(H^\top) = \{0\} = (\mathrm{Im}\, H)^\perp \tag{3.33}$$

donc $\mathrm{Im}\, H = \mathbb{R}^{nd}$. Par conséquent, l'intersection $\mathrm{Im}\, H + w \cap \mathrm{int}\, L^*$ n'est pas vide : (H) est satisfaite. Ceci prouve qu'une solution au problème incrémental existe, indépendamment de la valeur de w et des coefficients de frottement μ^i.

Une hypothèse plus faible que (3.33) est la suivante

$$\mathrm{Ker}(H^\top) \cap L = \{0\}. \tag{3.34}$$

Le corollaire 16.4.2 de [Roc70] affirme que pour tous cônes convexes fermés $K_1, \ldots K_p$ on a

$$(K_1 \cap \cdots \cap K_p)^\circ = \mathrm{cl}(K_1^\circ + \cdots + K_p^\circ).$$

En prenant $p = 2$, $K_1 := \mathrm{Ker}(H^\top)$ et $K_2 := L$, on trouve

$$(\mathrm{Ker}(H^\top) \cap L)^\circ = \mathrm{cl}(\mathrm{Im}\, H + L^\circ)$$

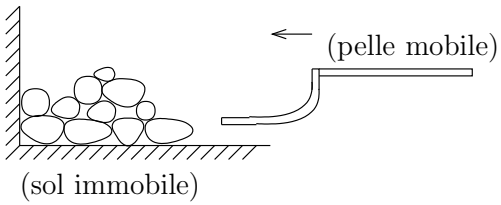


FIG. 3.5 – Deux objets extérieurs de vitesses différentes

Ainsi, si (3.34) est satisfaite, alors

$$\mathrm{cl}(\mathrm{Im}\, H + L^\circ) = \{0\}^\circ = \mathbb{R}^{nd}$$

donc $\mathrm{Im}\, H + L^\circ = \mathbb{R}^{nd}$ (un ensemble convexe dense est égal à tout l'espace, comme on peut le démontrer à partir du théorème de Carathéodory), et (H) est satisfaite.

3.4.9 Nombre de contacts et nombre de degrés de liberté

On a vu dans la sous-section précédente 3.4.8 que le problème incrémental admet toujours une solution lorsque H est surjective. Comme H est de taille $nd \times m$, ceci n'est certainement pas vérifié lorsque $nd > m$, mais peut arriver si $nd \leq m$.

De plus quand le sous-espace vectoriel $\mathrm{Im}(H)$ n'est pas de dimension pleine, il y a intuitivement peu de chances qu'il intersecte $L^* - w$; en effet, L^* est inclus dans le produit de n demi-espaces en dimension d et occupe donc au maximum une fraction $1/2^n$ du « volume disponible ».

Le caractère non-surjectif de H implique aussi la non-injectivité de W. En particulier quand $nd > m$, W est toujours singulière ; son conditionnement devient infini, ce qui pose des difficultés supplémentaires aux algorithmes de résolution que nous verrons au chapitre 4. On soupçonne donc une certaine corrélation entre le rapport $(nd)/m$ d'une part, et la satisfaction de (H) ou la « difficulté » du problème d'autre part. Ceci sera confirmé par les expériences numériques du chapitre 5.

3.4.10 Quand le critère ne s'applique pas

Lorsque le système mécanique est plus compliqué, par exemple s'il existe plusieurs objets extérieurs animés de vitesses différentes comme sur la figure 3.5, il n'est pas possible de vérifier le critère de façon directe. Dans ce cas, on peut essayer de le vérifier par le calcul, sur un ordinateur. La section suivante montre comment effectuer cette vérification en utilisant des algorithmes d'optimisation.

3.5 Vérification du critère par optimisation

Dans cette section, on va voir comment il est possible d'appliquer le résultat d'existence 3.19 en utilisant des calculs numériques pour vérifier les hypothèses ($\bar{H}$) et (H).

On suggère deux idées, qui consistent à résoudre : l'une, un problème de faisabilité, et l'autre, un problème d'optimisation.

3.5.1 Faisabilité ou optimisation

La vérification de l'hypothèse ($\bar{H}$) consiste à déterminer si l'intersection d'un espace affine et d'un produit de cônes du second ordre est vide ou non. Il s'agit donc d'un problème de faisabilité ; étant donnés H, w, les normales e^i et les coefficients μ^i, on cherche $v \in \mathbb{R}^m$ tel que

$$Hv + w \in L^*. \tag{3.35}$$

Définissons également $e := [(e^1)^\top, \dots, (e^n)^\top]^\top$, puis considérons le problème d'optimisation

$$\begin{cases} \max s \\ Hv - es + w \in L^* \end{cases} \tag{3.36}$$

Il existe $s \geq 0$ tel que $Hv + es + w \in L^*$ si et seulement si ($\bar{H}$) est vérifiée. De même, il existe $s > 0$ tel que $Hv + es + w \in L^*$ si et seulement si (H) est vérifiée. Autrement dit, résoudre le problème (3.36) permet de vérifier les hypothèses du critère d'existence :

- si la valeur optimale de (3.36) est strictement négative, alors ni (H) ni ($\bar{H}$) n'ont lieu
- si cette valeur optimale est nulle, alors ($\bar{H}$) a lieu mais pas (H)
- enfin si cette valeur est strictement positive, (H) et ($\bar{H}$) ont lieu toutes les deux.

3.5.2 Notion de conditionnement

Lors de la résolution pratique de (3.36), on peut bien sûr arrêter l'algorithme dès qu'une valeur strictement positive est atteinte (il n'est pas nécessaire de terminer l'optimisation). D'autre part, le cas où la valeur optimale est exactement zéro est indétectable en pratique, puisque tous les calculs faits sur ordinateur sont approximatifs et que la valeur s^* issue de la résolution numérique de (3.36) n'est pas exactement égale à sa valeur optimale. La situation est donc plutôt la suivante (où $\delta > 0$ est une une tolérance fixée).

- Si $s^* \leq -\delta$, on ne peut rien dire. Il est possible qu'une solution existe, mais on ne sait pas le démontrer ; il n'est même pas sûr que le critère d'existence soit violé : l'algorithme n'a peut-être simplement pas atteint une précision suffisante.
- Si $-\delta < s^* < \delta$, l'hypothèse ($\bar{H}$) est peut-être vérifiée mais une petite perturbation des données suffirait à ce qu'elle ne le soit plus.
- Enfin si $s^* \geq \delta$, une solution existe.

Dans le troisième cas, on a envie de déclarer le problème « bien conditionné » car il continuerait à posséder une solution malgré une petite perturbation des données. Cette notion de conditionnement serait cependant imparfaite, car l'hypothèse ($\bar{H}$) n'est qu'une condition *suffisante* d'existence. Il est possible qu'une solution existe sans que ce critère ne soit vérifié. Autrement dit, il arrive qu'on déclare un problème « mal conditionné » de façon exagérément pessimiste.

D'autre part, même quand ($\bar{H}$) est une condition nécessaire et suffisante comme pour le problème de Painlevé de la sous-section 3.4.3, cette notion de conditionnement est critiquable. Certes, lorsque seule l'hypothèse ($\bar{H}$) est vérifiée mais pas (H) (c'est-à-dire lorsque $u_0 = 0$ et $\tan\theta = \mu$, voir la figure 3.3), une minuscule modification des données comme remplacer $u_0 = 0$ par $u_0 = \varepsilon > 0$ suffit à ce que le problème n'admette plus de solution. Pourtant, on voit aussi sur la figure 3.3 que pour $u_0 > 0$ très petit et $\tan(\theta)$ légèrement supérieur à μ, la valeur optimale de (3.36) vaut $+\infty$ alors que le problème semble presque aussi mal conditionné que dans le cas où $u_0 = 0$ et $\tan(\theta) = \mu$. Ceci est dû au fait que v n'est pas contraint dans (3.36) et peut prendre des valeurs extrêmes pour permettre à s d'atteindre $+\infty$ même lorsque le problème est, intuitivement, mal conditionné. On propose donc de borner v en imposant par exemple $\|v\|_\infty \leq R$ dans (3.36), où R est une constante positive qui limite v à une valeur raisonnable (pour le problème mécanique considéré). On propose donc de considérer qu'un problème incrémental est d'autant mieux conditionné que la valeur optimale du problème suivant est grande

$$\begin{cases} \max s \\ Hv + Es + w \in L^* \\ \|v\|_\infty \leq R. \end{cases} \tag{3.37}$$

Ainsi, pour vérifier l'hypothèse (H) ou ($\bar{H}$) du théorème 3.19 par le calcul, on peut : soit résoudre le problème de faisabilité (3.35), ce qui est moins coûteux mais ne fournit pas d'information sur le conditionnement du problème ; soit résoudre le problème d'optimisation (3.37), ce qui est plus coûteux mais permet de s'assurer que le problème possède une certaine robustesse. Les deux sous-sections suivantes expliquent comment résoudre en pratique le problème de faisabilité (3.35) et le problème d'optimisation (3.37).

3.5.3 Dimension 2

Le cône du second ordre est ici polyédral et la contrainte

$$x \in L^*$$

se récrit trivialement sous forme de $2n$ contraintes d'inégalité linéaires. Ainsi, vérifier (3.35) consiste seulement à s'assurer qu'un certain polyèdre n'est pas vide ; ce problème de faisabilité est classique et on peut le résoudre en effectuant par exemple la phase 1 de l'algorithme du simplexe (qui permet de trouver une base réalisable).

3.5.4 Dimension 3

Une première idée que nous avons utilisée est d'attaquer le problème de faisabilité par un algorithme de projection. On peut par exemple résoudre

$$\begin{cases} \min \frac{1}{2}\|v\|^2 \\ Hv + w \in L^* \end{cases} \tag{3.38}$$

qui est un problème connu (SOCLS, pour « second-order cone least squares » [Mal09]). Il peut être résolu par exemple par un algorithme dual : on se ramène ainsi à minimiser

sans contraintes une fonction convexe et C^1. L'inconvénient de cette méthode est qu'elle ne fournit en général qu'un point du bord, et ne permet donc de vérifier que ($\bar{H}$) mais pas (H).

Une autre idée très simple consiste à remplacer L^* dans (3.35) par un cône polyédral inclus dans L^*. Si on parvient à résoudre le problème de faisabilité linéaire (ce qui est fait en pratique avec un solveur LP), alors on résout du même coup le problème de faisabilité non linéaire.

Enfin, si l'on préfère s'attaquer à (3.37) en dimension 3, il est nécessaire de disposer d'un solveur SOCP comme SeDuMi [Stu99] ; le coût de la vérification devient donc nettement supérieur.

Remarque 3.23. En dimension $d = 2$, le problème de faisabilité est linéaire : s'il n'admet pas de solution, alors le lemme de Farkas permet d'exhiber un certificat prouvant que le polyèdre correspondant est vide. De plus, un tel certificat est généralement fourni par les solveurs qui utilisent l'algorithme du simplexe. En revanche en dimension $d = 3$, il n'est pas évident de produire un certificat de non-satisfaction lorsque le problème de faisabilité (3.35) n'a pas de solution. Supposons par exemple que l'on résout (3.38) par un algorithme dual ; on arrête cet algorithme lorsqu'on observe que la valeur de la fonction duale devient très grande, et on suppose alors que le problème dual est non borné, que le primal est infaisable et que (3.35) n'a pas de solution. Mais on ne peut pas en avoir la certitude : en attendant plus longtemps, on aurait peut-être trouvé une solution au problème dual.

Conclusion

En reformulant le problème incrémental sous la forme (3.4) d'un problème d'optimisation convexe « facile » couplé avec une équation de point fixe, on obtient une démonstration d'existence sous l'hypothèse ($\bar{H}$). Cette hypothèse permet de traiter de nombreux exemples (sous-section 3.4), et elle est nécessaire dans le cas sans frottement et pour un certain nombre d'autres cas particuliers. De plus, lorsqu'on ne parvient pas à vérifier l'hypothèse ($\bar{H}$) par le raisonnement, on peut recourir à une vérification par le calcul en utilisant un ordinateur. Enfin, puisqu'on dispose d'une condition suffisante d'existence, on peut définir une notion (imparfaite) de conditionnement : on déclare un problème mal conditionné lorsque la condition suffisante d'existence est « presque insatisfaite ».

La reformulation (3.4) semble être une bonne manière de considérer le problème incrémental : sous cette forme, la démonstration du résultat d'existence est raisonnablement simple et intuitive. En généralisant l'argument de perturbation, on pourrait sans doute généraliser le résultat d'existence au cas où la matrice de masse est seulement semi-définie ($M \in \mathbf{S}_m^+$ au lieu de $M \in \mathbf{S}_m^{++}$), mais ce travail reste à réaliser.

La reformulation proposée consiste, au prix de l'introduction d'une variable supplémentaire $s \in \mathbb{R}^n$, à diviser le problème incrémental en deux : une partie facile, convexe, pour laquelle des algorithmes rapides et surtout robustes sont disponibles ; et une partie

difficile (le problème de point fixe), de petite dimension, qui concentre les problèmes de non-régularité et de non-convexité. Ainsi, si l'on utilise notre reformulation pour effectuer des calculs pratiques, on peut espérer

- d'une part, un gain de robustesse (puisque la partie qui peut échouer est maintenant de taille plus réduite) ;
- d'autre part, un gain de vitesse (puisqu'une grosse partie du problème est convexe et peut être résolue efficacement).

Le chapitre suivant est donc consacré à l'étude numérique du problème incrémental. Après avoir dressé un panorama des algorithmes existants, on proposera un algorithme nouveau basé sur la formulation (3.4).

Chapitre 4

Résolution pratique du problème incrémental

Ce chapitre décrit l'aspect algorithmique de la résolution du problème incrémental (1.33). Il contient d'une part un état de l'art des techniques numériques existantes à ce jour et utilisées en pratique par les mécaniciens du contact (on pourra consulter aussi [AB08] pour une liste de ces méthodes), et d'autre part une approche nouvelle basée sur la formulation (3.4). On verra que de nombreuses méthodes numériques ont été proposées dans la littérature pour résoudre le problème incrémental ; ce chapitre est consacré à leur classification en différentes familles. Cette classification pourra sembler parfois un peu arbitraire ; elle a été choisie de manière à correspondre à celle établie au chapitre 1 pour les différentes formulations équivalentes de la contrainte $(u, r) \in \mathcal{C}(e, \mu)$.

Organisation du chapitre

La section 4.1 traite du cas particulier de la dimension 2 ($d = 2$), dans lequel on dispose d'une reformulation du problème incrémental sous forme d'un problème de complémentarité linéaire (LCP).

La section 4.2 décrit les méthodes de résolution fonctionnelles, c'est-à-dire celles qui consistent à transformer le problème incrémental en un système d'équations non linéaires et non régulières dont on recherche ensuite un zéro. La méthode d'Alart et Curnier, qui attaque la reformulation fonctionnelle décrite à la sous-section 1.3.1 par l'algorithme de Newton, en fait partie. Cette méthode, dont d'autres versions similaires on été proposées (voir par exemple [CKPS98]), est une référence pour le problème tridimensionnel. On présentera aussi les méthodes basées sur la reformulation de De Saxcé (sous-section 1.3.3) qui transforme le problème incrémental en un problème de complémentarité du second ordre ; en introduisant une fonction de complémentarité, on peut alors de nouveau se ramener à un système d'équations non linéaires.

La section 4.3 est consacrée aux techniques de résolution par optimisation, qui reviennent à reformuler le problème incrémental comme un problème de minimisation avec ou sans contraintes. A ce jour, ces techniques ne semblent pas compétitives avec les ap-

proches fonctionnelles, comme on le verra dans le chapitre suivant ; cependant, elles ont été encore relativement peu étudiées et ce verdict n'est peut-être pas définitif.

La section 4.6 décrit la méthode de Haslinger, qui revient à résoudre un problème d'optimisation convexe paramétrique couplé avec un problème de point fixe. Elle présente donc une certaine ressemblance avec celle que nous proposons dans la section 4.7 et qui consiste à résoudre l'équation de point fixe (3.24) du chapitre précédent.

Enfin, la section 4.8 explique la technique dite « de Gauss-Seidel » (en raison d'une certaine ressemblance avec la méthode du même nom consacrée à la résolution des systèmes linéaires), qui permet de résoudre le problème incrémental en résolvant itérativement une suite de problèmes de taille réduite, comme si le système ne comportait qu'un seul contact. Ce système réduit peut être résolu en utilisant n'importe laquelle des approches générales décrites dans les sections 4.2 à 4.7.

Ainsi, de nombreuses idées ont été proposées pour résoudre le problème incrémental ; cependant, on s'attaque à un problème difficile (on a vu par exemple son caractère NP-complet au chapitre 1) et dont les propriétés théoriques sont assez inconfortables, notamment à cause de l'absence de régularité et de convexité : on ne dispose donc pas pour l'instant de méthode satisfaisante pour le résoudre, même lorsqu'on sait (par exemple grâce au théorème 3.19) qu'une solution existe, et les algorithmes existants (à l'exception de la méthode de Haslinger, sous certaines hypothèses) ne sont malheureusement pas soutenus par de solides garanties théoriques.

On rappelle qu'en éliminant v dans l'équation dynamique $Mv + f = H^\top r$ comme à l'équation (1.34) et en l'injectant dans l'équation cinématique $u = Hv + w$, on obtient

$$u = Wr + q, \quad \text{avec} \quad W = HM^{-1}H^\top \quad \text{et} \quad q = w - HM^{-1}f. \tag{4.1}$$

4.1 Méthodes spécifiques à la dimension 2

En dimension 2, on a vu au chapitre 1 qu'il est possible d'écrire la loi de Coulomb (1.22) sous forme d'une contrainte de complémentarité linéaire, en introduisant des variables supplémentaires qui correspondent intuitivement aux parties positives et négatives de la force tangentielle r_{T} et de la vitesse tangentielle u_{T}. On va maintenant montrer que le problème incrémental complet peut être reformulé comme un problème de complémentarité linéaire (LCP), et discuter brièvement les méthodes disponibles pour le résoudre.

4.1.1 Formulation LCP

Quitte à effectuer un changement de base orthonormale, on peut supposer qu'en tout point de contact i, on a $u^i = (u^i_{\mathrm{T}}, u^i_{\mathrm{N}})$. Autrement dit, on suppose que la première coordonnée de u^i correspond à sa composante tangentielle, et sa seconde coordonnée à sa composante normale. On suppose de même que $r^i = (r^i_{\mathrm{T}}, r^i_{\mathrm{N}})$. On note i_{T} les indices des coordonnées tangentielles dans u et r (c'est-à-dire les indices impairs) et i_{N} les indices des coordonnées normales (c'est-à-dire les indices pairs). On note ensuite $u_{\mathrm{T}} := \{u_j,\ j \in i_{\mathrm{T}}\}$ et $u_{\mathrm{N}} := \{u_j,\ j \in i_{\mathrm{N}}\}$ les coordonnées tangentielles et normales de u, et r_{T}, r_{N}, q_{T} et q_{N}

celles de r et q. De même, on extrait de W les sous-matrices W_{TT}, W_{TN}, W_{NT} et W_{NN}. Enfin on pose $D := \mathrm{diag}(\mu^i)$. En utilisant (1.30), on obtient facilement la reformulation suivante.

Lemme 4.1. *En posant*

$$A := \begin{bmatrix} (W_{TN} - W_{TT}D) & 2W_{TT} & \mathrm{I}_{n,n} \\ (W_{NN} - W_{NT}D) & 2W_{NT} & 0_{n,n} \\ D & -\mathrm{I}_{n,n} & 0_{n,n} \end{bmatrix}, \quad b := \begin{bmatrix} q_T \\ q_N \\ 0_{n,1} \end{bmatrix} \tag{4.2}$$

le problème incrémental (1.33) est équivalent au LCP suivant : trouver (x, y) dans $\mathbb{R}^{6n}$ tel que

$$\begin{cases} y = Ax + b \\ 0 \le x \perp y \ge 0 \\ x = (r_N, \lambda^+, \eta^-), \ \ y = (\eta^+, u_N, \lambda^-) \end{cases} \tag{4.3}$$

au sens où (v, u, r) est solution de (1.33) si et seulement s'il existe $(\eta^-, \eta^+, \lambda^-, \lambda^+)$ tels que (4.3) soit vérifiée.

Démonstration. On voit que résoudre le problème incrémental dans lequel on a éliminé v est équivalent à trouver u_{T}, u_{N}, r_{T}, r_{N}, η^-, η^+, λ^-, λ^+ dans $\mathbb{R}^n$ qui satisfont

$$\begin{cases} u_{\mathrm{T}} & = & W_{\mathrm{TT}}\, r_{\mathrm{T}} + W_{\mathrm{TN}}\, r_{\mathrm{N}} + q_{\mathrm{T}} \\ u_{\mathrm{N}} & = & W_{\mathrm{NT}}\, r_{\mathrm{T}} + W_{\mathrm{NN}}\, r_{\mathrm{N}} + q_{\mathrm{N}} \\ r_{\mathrm{T}} & = & \lambda^+ - \lambda^- \\ Dr_{\mathrm{N}} & = & \lambda^+ + \lambda^- \\ u_{\mathrm{T}} & = & \eta^+ - \eta^- \\ 0 \le \lambda^- & \perp & \eta^- \ge 0 \\ 0 \le \lambda^+ & \perp & \eta^+ \ge 0 \\ 0 \le u_{\mathrm{N}} & \perp & r_{\mathrm{N}} \ge 0 \end{cases} \tag{4.4}$$

On reformule (4.4) comme un LCP en éliminant les variables r_{T} et u_{T} grâce à

$$r_{\mathrm{T}} = \lambda^+ - \lambda^- = 2\lambda^+ - D\, r_{\mathrm{N}}, \quad \text{et} \quad u_{\mathrm{T}} = \eta^+ - \eta^-.$$

On obtient le système suivant.

$$\begin{cases} \eta^+ & = & \eta^- + W_{\mathrm{TT}}\, (2\lambda^+ - D\, r_{\mathrm{N}}) + W_{\mathrm{TN}}\, r_{\mathrm{N}} + q_{\mathrm{T}} \\ u_{\mathrm{N}} & = & W_{\mathrm{NT}}\, (2\lambda^+ - D\, r_{\mathrm{N}}) + W_{\mathrm{NN}}\, r_{\mathrm{N}} + q_{\mathrm{N}} \\ \lambda^- & = & Dr_{\mathrm{N}} - \lambda^+ \\ 0 \le \lambda^- & \perp & \eta^- \ge 0 \\ 0 \le \lambda^+ & \perp & \eta^+ \ge 0 \\ 0 \le u_{\mathrm{N}} & \perp & r_{\mathrm{N}} \ge 0 \end{cases} \tag{4.5}$$

ce qui achève la démonstration. ∎

La théorie des LCP est très riche, on pourra par exemple consulter [CPS92]. Une question naturelle, reliée à l'étude d'existence que nous proposons au chapitre 3, est

de savoir quand ce LCP possède une solution, et quand elle est unique. On dispose en particulier des définitions et des théorèmes (donnés sans démonstration) qui suivent [CPS92].

Définition 4.2. On dit qu'une matrice carrée réelle est une *P-matrice* lorsque tous ses mineurs principaux sont strictement positifs.

Théorème 4.3. *Soit $M \in \mathcal{M}_n(\mathbb{R})$ une P-matrice, et $b \in \mathbb{R}^n$. Le LCP*

$$y = Ax + b,\ 0 \leq x \perp y \geq 0$$

possède une unique solution.

L'utilisation pratique de ce théorème nécessite malheureusement de connaître tous les mineurs principaux, dont l'évaluation directe est trop coûteuse pour être effectuée en pratique. De plus, on ne s'attend pas à ce que le LCP (4.3) issu du problème incrémental admette une solution unique : on a vu au chapitre 1 que la multiplicité de solutions est une situation courante, notamment lorsqu'on manipule des solides rigides (la situation de non-existence de solution est également possible, mais elle est plus pathologique ; voir le chapitre 3).

Définition 4.4. On dit qu'une matrice carrée réelle $M \in \mathbb{R}^{n\times n}$ est *copositive* lorsque

$$\forall x \in \mathbb{R}^n_+,\ x^\top Mx \geq 0.$$

On dit que $M \in \mathbb{R}^{n\times n}$ est *strictement copositive* lorsque

$$\forall x \in \mathbb{R}^n_+\backslash\{0\},\ x^\top Mx > 0.$$

Théorème 4.5. *Soit $M \in \mathbb{R}^{n\times n}$ une matrice strictement copositive. Alors, pour tout $q \in \mathbb{R}^n$, le LCP*

$$0 \leq Mx + q \perp x \geq 0$$

a une solution.

Théorème 4.6. *Soit $M \in \mathbb{R}^{n\times n}$ une matrice copositive, et soit $q \in \mathbb{R}^n$. Si*

$$\forall v \in \mathbb{R}^n,\ [v \geq 0,\ Mv \geq 0,\ v^\top Mv = 0] \Rightarrow [v^\top q \geq 0]$$

alors le LCP

$$0 \leq Mx + q \perp x \geq 0$$

a une solution.

Ce dernier résultat, accompagné de calculs pénibles, permet de retrouver notre résultat d'existence 3.19 dans le cas bidimensionnel (ou en 3D lorsque le cône du second ordre est remplacé par un cône polyédral) ; voir [Ste98]. On remarque d'ailleurs que l'hypothèse faite sur q dans le théorème 4.6 ressemble à l'hypothèse $(\bar{H})$ du chapitre 3 : il s'agit d'une condition d'appartenance de q à un certain cône dual.

4.1.2 Résolution des LCP

Les premiers algorithmes proposés pour résoudre les LCP, en particulier ceux qui apparaissaient comme conditions d'optimalité de programmes quadratiques (QP), sont très similaires à l'algorithme du simplexe utilisé pour résoudre les programmes linéaires ([Fle87], chap. 10). Les techniques plus récentes, quant à elles, utilisent des méthodes de points intérieurs. Certains auteurs diffusent leurs implémentations, ce qui permet de les utiliser en boîte noire ; le solveur commercial Path [DF95] est une référence.

On peut citer la méthode de Dantzig-Wolfe, celle « des pivots principaux » (*principal pivoting method*) et l'algorithme de Lemke [LH64]. Dans tous les cas, des hypothèses sont nécessaires pour assurer la convergence de l'algorithme, la situation la plus agréable étant celle où la matrice du LCP est symétrique définie positive ; dans ce cas, le LCP est équivalent à un QP fortement convexe. Le LCP (4.3) ne vérifie malheureusement pas cette hypothèse en général. Cependant, sous l'hypothèse (assez restrictive) que w est nul, Stewart [Ste98, ST00] propose une reformulation LCP équivalente à (4.5) telle que l'algorithme de Lemke converge toujours vers une solution (il en existe une car l'hypothèse ($\bar{H}$) est automatiquement satisfaite quand $w = 0$). Ce résultat utilise le théorème 4.4.13 de [CPS92], qui affirme que sous une certaine hypothèse de non-dégénérescence, l'algorithme de Lemke fournit une démonstration constructive du théorème 4.6.

4.2 Reformulations fonctionnelles

Passons maintenant au cas général ($d = 2$ ou $d = 3$). On a vu au chapitre 1 qu'il est possible d'exprimer la contrainte $(u, r) \in \mathcal{C}(e, \mu)$ sous forme fonctionnelle $f(u, r) = 0$, où f est différentiable presque partout avec une expression explicite pour sa matrice jacobienne. Ceci va nous permettre d'écrire le problème (1.33) sous la forme d'un problème de recherche de zéro d'une fonction, et d'envisager la résolution du système par les méthodes usuelles de résolution d'équations non-linéaires ; on citera la méthode du point fixe (la formulation d'Alart et Curnier et celle de De Saxcé avec projection prennent en effet naturellement la forme d'un problème de point fixe) et celle de Newton, mais la théorie des équations non-linéaires est riche et de nombreuses autres possibilités sont envisageables. Le problème majeur, en grande partie non résolu à ce jour, consiste à adapter la théorie développée dans le cadre de l'analyse lisse à notre cas où les fonctions sont non-régulières. Un tel exemple d'adaptation est donné dans [CKPS98] qui définit un algorithme dit « de semi-smooth Newton ».

Les formulations fonctionnelles sont

- la formulation d'Alart et Curnier (sous-section 4.2.1),
- la formulation de De Saxcé exprimée sous forme de projection (sous-section 4.2.2), que l'on peut voir comme un cas particulier du point suivant,
- la formulation de De Saxcé exprimée par une fonction de complémentarité du second ordre, comme la fonction de Fischer-Burmeister (sous-section 4.2.3).

4.2.1 Formulation d'Alart et Curnier

On rappelle que la fonction d'Alart et Curnier est définie aux équations (1.23) et (1.24) par les formules

$$f_{AC,\mathrm{N}}\colon \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ (u,r) \longmapsto g_{\mathrm{N}}(u,r) - r_{\mathrm{N}} \end{cases} \quad \text{avec } g_{\mathrm{N}}(u,r) := \mathrm{P}_{\mathbb{R}^+}(r_{\mathrm{N}} - \rho_{\mathrm{N}}\, u_{\mathrm{N}})$$

et

$$f_{AC,\mathrm{T}}\colon \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^{d-1} \\ (u,r) \longmapsto g_{\mathrm{T}}(u,r) - r_{\mathrm{T}} \end{cases} \quad \text{avec } g_{\mathrm{T}}(u,r) := \mathrm{P}_{B(0,\mu\, r_{\mathrm{N}})}(r_{\mathrm{T}} - \rho_{\mathrm{T}}\, u_{\mathrm{T}})$$

où $\rho_{\mathrm{N}} > 0$ et $\rho_{\mathrm{T}} > 0$ sont deux constantes. On a vu au chapitre 1 que la contrainte $f_{AC}(u,r) := (f_{AC,\mathrm{N}}, f_{AC,\mathrm{T}})(u,r) = 0$ est équivalente à la loi de Coulomb (1.22). On a donc trivialement le résultat suivant.

Lemme 4.7. *Le problème incrémental (1.33) est équivalent au système d'équations suivant : trouver (v,u,r) dans $\mathbb{R}^{m+nd+nd}$*

$$\begin{cases} Mv - H^\top r + f &= 0 \\ Hv - u + w &= 0 \\ f_{AC}(u,r) &= 0. \end{cases} \tag{4.6}$$

La formulation fonctionnelle (4.6) est un problème de recherche de zéro d'un système d'équations non-linéaires et non régulières. L'annexe D détaille le calcul de la matrice jacobienne (quand elle existe) de la fonction d'Alart et Curnier f_{AC}, ce qui permet d'attaquer (4.6) à l'aide de la méthode de Newton. C'est cette technique qui constitue la méthode originale d'Alart et Curnier.

D'autre part, comme la fonction d'Alart et Curnier f_{AC} s'écrit naturellement sous la forme $f_{AC} = g_{AC} - Q^{-1}\,\mathrm{I}$ où Q est une matrice orthogonale et $g_{AC} := (g_{\mathrm{N}}, g_{\mathrm{T}})$, on dispose d'une deuxième formulation équivalente du problème incrémental sous forme d'une équation de point fixe. Cette nouvelle formulation (4.7) est très similaire à (4.6), mais elle suggère des méthodes de résolution différentes utilisant des algorithmes de point fixe.

Lemme 4.8. *Le problème incrémental (1.33) est équivalent au problème de point fixe suivant : trouver $r \in \mathbb{R}^d$ tel que*

$$r = Q\, g_{AC}(Wr + q, r) \tag{4.7}$$

où $Q := \mathrm{Diag}([(P_N^i)^\top, (P_T^i)^\top])$, et P_N^i et P_T^i sont les projections sur la direction normale et le plan tangent au contact i.

Démonstration. Le problème incrémental est équivalent à, pour tout i

$$P_{\mathrm{N}}^i\, r^i = g_{\mathrm{N}}([Wr+q]^i, r^i) \quad \text{et} \quad P_{\mathrm{T}}^i\, r^i = g_{\mathrm{T}}([Wr+q]^i, r^i).$$

En posant $Q^i := [(P^i_{\mathrm{N}})^\top, (P^i_{\mathrm{T}})^\top]$, ceci se récrit $(Q^i)^\top r^i = g_{AC}([Wr+q]^i, r^i)$ ou encore, en utilisant le fait que la matrice Q^i est orthogonale, $r^i = Q^i\, g_{AC}([Wr+q]^i, r^i)$. ∎

Notre expérience (voir le chapitre 5) indique qu'il est beaucoup plus robuste d'attaquer l'équation (4.6) avec la méthode de Newton que l'équation (4.7) avec des itérations de point fixe. D'autre part, comme on s'y attend, on n'observe pas de convergence globale : même lorsqu'une solution existe, il arrive qu'aucune des deux techniques ne la trouve (si elles sont « mal » initialisées). Il arrive aussi que la méthode de point fixe ne converge même pas localement : même en lui donnant un excellent premier itéré, elle peut échouer. Quant à la méthode de Newton, elle se comporte mieux en pratique mais manque tout autant de justifications théoriques (voir tout de même [Ala97]).

4.2.2 Formulation de De Saxcé avec projection

Cette approche n'est qu'un cas particulier (voir la remarque 4.11 ci-dessous) de la méthode des fonctions de complémentarité exposée dans la sous-section suivante, mais elle présente une certaine analogie avec la méthode d'Alart et Curnier et constitue un exemple introductif aux fonctions de complémentarité générales. On la présente donc séparément.

On a vu au chapitre 1 que la loi de frottement pouvait être exprimée sous la forme $L^* \ni \widetilde{u} \perp r \in L$ avec $\widetilde{u}$ défini par (1.28). On a donc trivialement le résultat suivant.

Lemme 4.9. *Soit $\rho_{DS} > 0$. Le problème incrémental (1.33) est équivalent au problème de point fixe : trouver $r \in \mathbb{R}^{nd}$ tel que*

$$\mathrm{P}_L(r - \rho_{DS}\, \widetilde{u}(r)) = r. \tag{4.8}$$

Remarque 4.10. Comme $\widetilde{u}$ dépend de u par (1.28) et u est une fonction de r par (4.1), on se permet de noter $\widetilde{u}$ sous la forme $\widetilde{u}(r)$ d'une fonction de r.

On définit donc la fonction f_{DS} par

$$f_{DS} : \begin{cases} \mathbb{R}^{nd} & \longrightarrow \mathbb{R}^{nd} \\ r & \longmapsto \mathrm{P}_L(r - \rho_{DS}\, \widetilde{u}(r)) \end{cases} \tag{4.9}$$

et on essaie de trouver un point fixe f_{DS} ou d'annuler $f_{DS} - \mathrm{I}$. Comme pour la fonction d'Alart et Curnier, on peut tenter d'effectuer directement des itérations de point fixe : c'est la méthode proposée par De Saxcé et Feng [DSF98], sous le nom de « méthode d'Uzawa ». Cependant, l'expérience semble prouver que cette méthode n'est pas très robuste, du moins sur les exemples que nous avons considérés (chapitre 5).

En revanche, la résolution de (4.8) par la méthode de Newton (qui nécessite de différentier f_{DS}, voir l'annexe D) fonctionne bien en pratique. Son efficacité est comparable à celle de la méthode de Newton sur la formulation d'Alart et Curnier, et ses défauts sont les mêmes (existence et inversibilité de la matrice jacobienne, cyclage). Bizarrement, cette idée ne semble pourtant pas avoir été exploitée jusqu'à présent dans la littérature, sauf très récemment [JF08] pour résoudre le problème à un seul contact.

4.2.3 Fonctions de complémentarité générales

Pour utiliser des notations traditionnelles, on suppose dans cette sous-section que $n = 1$ et $\mu = 1$, de sorte que le produit de cônes L est réduit à un seul cône du second ordre $K = K^*$ (le cône de Lorentz usuel). Tout ce qui est dit ici est adaptable au cas général $n \geq 1$ et $\mu \in [0, \infty[$. On dit que $\phi : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ est une fonction de complémentarité pour K lorsque, pour tout x, y dans $\mathbb{R}^n$

$$x \cdot y = 0, \quad x \in K, \quad y \in K \quad \Longleftrightarrow \quad \phi(x, y) = 0. \tag{4.10}$$

On trouvera dans [FLT02] une technique générale pour construire de telles fonctions de complémentarité ϕ pour K en utilisant le formalisme des algèbres de Jordan. Outre cette famille de fonctions, on peut utiliser la fonction de Fischer-Burmeister

$$\phi_{\mathrm{FB}}(x, y) = x + y - \sqrt{x \circ x + y \circ y} \tag{4.11}$$

où l'opérateur $\circ$ représente le produit de Jordan associé à K et la racine carrée est associée à $x \mapsto x \circ x$ (voir [FLT02] pour ces notions). On peut donc remplacer la contrainte $K \ni \widetilde{u} \perp r \in K$ par la contrainte fonctionnelle $\phi(\widetilde{u}, r) = 0$ et utiliser la méthode de Newton comme précédemment, avec les mêmes problèmes théoriques liés à la non-régularité des fonctions employées.

Une autre approche, par régularisation, consiste à utiliser une approximation régulière (généralement, de classe C^1) ϕ_α de ϕ telle que, pour tout (x, y), la limite ponctuelle

$$\lim_{\alpha \to 0} \phi_\alpha(x, y)$$

existe et soit une fonction de complémentarité. L'approche par régularisation consiste alors à résoudre de manière approchée le problème incrémental dans lequel ϕ est remplacée par ϕ_α, par exemple en effectuant quelques itérations de la méthode de Newton, puis à faire décroître α. La difficulté de la méthode réside dans le choix des valeurs successives de α, et dans la résolution approchée du problème interne (à α fixé).

Les méthodes de complémentarité semblent relativement méconnues, ou en tout cas inutilisées, dans la communauté de la mécanique du contact pour s'attaquer au problème du frottement en dimension 3, avec un cône de frottement du second ordre. L'article récent de Kanno, Martins et Pinto Da Costa [KMPdC06], qui utilisent le solveur (Matlab) de problème de complémentarité conique d'Hayashi [HYF05], en est apparemment le seul exemple. Pourtant, de nombreux auteurs utilisent ces méthodes (souvent en boîte noire, par exemple grâce au solveur Path [FM97]) lorsque le problème de complémentarité est linéaire, soit parce qu'on se place en dimension $d = 2$, soit parce que le cône de Coulomb tridimensionnel est remplacé dans le modèle par un cône polyédral. Une extension naturelle de ces travaux serait donc d'utiliser les fonctions de complémentarité et leurs régularisées également dans le cas non-linéaire.

Remarque 4.11. La méthode de De Saxcé avec projection est en fait un cas particulier de cette méthode générale, dans laquelle la fonction de complémentarité choisie est la fonction-résidu naturelle (« natural residual function » [FLT02])

$$\phi(x, y) = \mathrm{P}_K(x - y) - x.$$

4.3 Approches par minimisation

Dans la section précédente, on a vu plusieurs techniques pour reformuler le problème incrémental comme un problème de recherche de zéro d'une fonction non-régulière. Dans cette section, on s'intéresse aux reformulations qui conduisent à un problème d'optimisation. On aura ainsi l'avantage de pouvoir stabiliser les algorithmes en assurant la décroissance d'une fonction de mérite.

On propose deux manières de produire de telles fonctions de mérite : la première (sous-section 4.3.1) aboutit à un problème d'optimisation sous contrainte conique, tandis que la seconde (sous-section 4.3.2) aboutit à un problème sans contrainte.

Une fois le problème incrémental reformulé comme un problème de minimisation, tout l'arsenal de l'optimisation peut être utilisé pour le résoudre, en prenant garde au fait que les résultats théoriques classiques ne s'appliquent en général pas à cause du caractère non-régulier des fonctions objectif employées. D'autre part, on a observé au chapitre 1 que l'ensemble des solutions du problème incrémental n'était pas nécessairement convexe, donc le problème d'optimisation associé ne sera certainement pas convexe et on n'a pas d'espoir d'obtenir un résultat de convergence globale.

4.3.1 Minimisation du bipotentiel

On utilise simplement le fait que $r \cdot \widetilde{u}$ est positif sur $L \times L^*$, et nul si et seulement si $r \perp \widetilde{u}$. Il suffit donc de minimiser $r \cdot \widetilde{u}$ sur $L \times L^*$, ou encore de résoudre le problème d'optimisation

$$\left\{ \begin{array}{ll} \min & r \cdot \widetilde{u}(r) = u(r) \cdot r + \sum_{i=1}^{n} \mu^i r_{\mathrm{N}}^i \|u_{\mathrm{T}}^i(r)\| \quad \text{avec} \quad u(r) = Wr + q \\ & u_{\mathrm{N}}(r) \geq 0 \\ & r \in L \end{array} \right. \tag{4.12}$$

où la contrainte (linéaire en r) $u_{\mathrm{N}} \geq 0$ impose $\widetilde{u} \in L^*$. On a besoin d'un algorithme d'optimisation non-linéaire capable de gérer la contrainte de cône du second ordre $r \in L$ et la contrainte d'inégalité linéaire $u_{\mathrm{N}}(r) \geq 0$; la méthode SQP est un candidat naturel. Dans la même veine, on pourra consulter aussi [AFMS08] où l'on se ramène à des problèmes avec contraintes de boîte.

4.3.2 Fonctions de mérite

Comme dans la sous-section 4.2.3, on suppose ici que $n = 1$ et $\mu = 1$. On dit que ϕ est une fonction de mérite pour $K = K^*$ lorsque

$$\phi(x, y) \geq 0 \text{ et } [\ \phi(x, y) = 0 \iff K \ni x \perp y \in K\ ]. \tag{4.13}$$

On peut construire une fonction de mérite en prenant (une fonction de) la norme d'une fonction de complémentarité ; typiquement, on prend le carré de sa norme euclidienne. Une autre technique consiste à élever au carré la fonction d'Alart et Curnier. On obtient un problème de moindres carrés sans contrainte, non régulier et non convexe mais que

l'on peut tout de même essayer de résoudre numériquement avec les techniques usuelles de l'optimisation. C'est ce que nous ferons au chapitre 5 consacré aux expériences numériques, en prenant pour ϕ la fonction de Fischer-Burmeister.

A propos de l'utilisation de fonctions de mérite pour résoudre les problèmes de complémentarité, on pourra consulter aussi [CT05], [PC09] et l'article de revue [FJ00].

4.4 Approches fonctionnelles : aspects numériques

On a vu dans la section précédente plusieurs manières de reformuler le problème incrémental sous la forme d'un problème de recherche de zéro ($\Phi(x) = 0$, avec Φ définie par exemple par (4.6)) ou un problème de point fixe ($p(x) = x$, avec p définie par (4.7) ou (4.8)). Cette section traite des algorithmes disponibles pour attaquer ces formulations fonctionnelles.

4.4.1 Méthode de Newton

La méthode originale d'Alart et Curnier consiste à chercher une solution du système (4.6) à l'aide de la méthode de Newton. On peut envisager plusieurs implémentations assez différentes de l'algorithme ; en particulier, il est certainement souhaitable d'éliminer la variable u du problème (4.6) grâce à la deuxième équation.

Par ailleurs, il faut garder à l'esprit qu'aucune des techniques basées sur la méthode de Newton (appliquée à telle ou telle reformulation fonctionnelle non-régulière) n'est confortée par des fondements théoriques solides. La sous-section D.1.4 montre par exemple que la situation où la matrice jacobienne de la fonction d'Alart et Curnier est bien définie, mais pas inversible, n'est pas un cas pathologique mais plutôt le cas général.

4.4.2 Recherche linéaire

On constate en pratique que même sur des exemples simples, et même lorsque les itérés de Newton sont tous bien définis (la matrice jacobienne de Φ existe et est inversible), la méthode de Newton appliquée à (4.6) peut ne pas converger ; on observe typiquement des trajectoires cycliques. Pour lutter contre ce phénomène, on peut modifier l'algorithme en lui ajoutant une recherche linéaire sur la fonction objectif des moindres carrés, définie par

$$g(x) := \frac{1}{2}\|\Phi(x)\|^2.$$

Cette idée est justifiée par l'espoir que Φ soit différentiable en x ; dans ce cas, on dispose du lemme suivant qui assure la décroissance locale de g dans la direction de Newton.

Lemme 4.12. *Soit $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^m$, U ouvert, et $x \in U$. On suppose que Φ est différentiable en x et que $J[\Phi](x)$ est inversible. Alors la direction de Newton, définie par*

$$dx := -\mathrm{Jac}[\Phi](x)^{-1}\Phi(x)$$

est une direction de descente.

Démonstration. On a

$$\nabla g(x) \cdot dx = (\mathrm{Jac}[\Phi](x)^\top \Phi(x)) \cdot (-\mathrm{Jac}[\Phi](x)^{-1}\Phi(x)) = -\|\Phi(x)\|^2$$

et dx est une direction de descente, à moins que x ne soit déjà une solution. ■

Dans notre cas, Φ n'est pas différentiable, mais on connaît souvent bien sa structure et elle est assez simple (penser à la fonction d'Alart et Curnier, par exemple). En faisant une étude spécifique à la fonction Φ utilisée et en s'inspirant des techniques et idées de l'optimisation non-régulière, on pourrait peut-être trouver une direction de descente dans le cas où l'itéré courant est un point de non différentiabilité. En effet, même si la fonction n'est pas différentiable au point courant, elle l'est au voisinage de ce point, et on connaît alors explicitement les matrices jacobiennes. Autrement dit, sous réserve que la fonction soit lipschitzienne, on connaît tout son sous-différentiel au sens de Clarke.

4.4.3 Convergence

On verra au chapitre 5 qu'en pratique, l'algorithme d'Alart et Curnier avec recherche linéaire s'avère souvent performant pour résoudre le problème incrémental. En revanche, on ne dispose pas de résultat théorique de convergence, même lorsqu'on fait l'hypothèse que les itérés de Newton sont tous bien définis. La seule garantie que l'on possède est la suivante : lorsqu'on utilise une recherche linéaire, on évite le problème du cyclage. En contrepartie, la recherche linéaire peut se bloquer en retournant plusieurs fois de suite un pas microscopique, indiquant que la direction choisie n'est pas une direction de descente.

4.4.4 Itérations de point fixe

On peut aussi envisager d'effectuer des itérations de point fixe pour résoudre (4.7) ou (4.8), ce qui évite les problèmes de définition et d'inversibilité de la matrice jacobienne de Φ.

La méthode de point fixe la plus simple pour résoudre l'équation de point fixe $p(r) = r$ consiste à effectuer des itérations de la forme $r \leftarrow p(r)$, mais des variantes plus élaborées sont possibles. On constatera au chapitre 5 que cette méthode converge parfois (sur de très petits problèmes et quand les coefficients de frottement ne sont pas trop gros, en général) mais que ceci est loin d'être toujours le cas ; en général, elle diverge brutalement. Cette technique n'est donc pas très séduisante ; cependant, utilisée pour résoudre les sous-problèmes à un seul contact produits par la méthode dite « de Gauss-Seidel » (section 4.8 ci-dessous), elle donne parfois de bons résultats.

4.5 Approche par minimisation : aspects numériques

4.5.1 Fonctions de mérite par moindres carrés

Revenons à la sous-section 4.3.2 ; numériquement, l'idée de minimiser $g : x \mapsto \frac{1}{2}\|\Phi(x)\|^2$ au lieu d'annuler Φ n'est pas très séduisante. En effet, elle revient essen-

tiellement à tenter d'annuler $\nabla g(x) = \mathrm{Jac}[\Phi](x)\,\Phi(x)$ au lieu d'annuler directement Φ. L'évaluation de la fonction g et de ses dérivées est coûteuse : évaluer ∇g (un vecteur) nécessite *a priori* d'évaluer $\mathrm{Jac}[\Phi]$ (une matrice), évaluer $\mathrm{Jac}[\nabla g]$ nécessite de calculer les dérivées secondes de Φ, etc. En pratique, il est rare que l'on puisse se permettre d'évaluer les dérivées secondes de Φ. On ne peut donc calculer que le gradient de g et pas sa hessienne. Autrement dit, pour minimiser g en pratique

- soit on utilisera une méthode d'optimisation générale d'ordre 1 (*i.e.*, qui ne nécessite que le calcul de ∇g), comme la méthode de quasi-Newton ;
- soit on utilise une méthode spécifique aux problèmes de moindres carrés, on pense alors à la méthode de Gauss-Newton.

Remarque 4.13. Il y a tout de même un avantage à l'idée de minimiser g au lieu d'annuler Φ : la fonction g peut être plus régulière que Φ. D'autre part, il arrive que l'on sache calculer $\nabla g(x) = \mathrm{Jac}[\Phi](x)\,\Phi(x)$ sans calculer explicitement $\mathrm{Jac}[\Phi](x)$, c'est-à-dire en ne faisant que des opérations sur des vecteurs et pas sur des matrices. C'est le cas pour la fonction de Fischer-Burmeister [CT05] que nous utiliserons dans les expériences numériques du chapitre 5.

4.5.2 Méthode de Gauss-Newton

On rappelle la définition d'un itéré de la méthode de Gauss-Newton pour minimiser la fonction de moindres carrés g.

Définition 4.14 (Méthode de Gauss-Newton)**.** Le pas de Gauss-Newton Newton dx associé à la fonction g et au point x est obtenu en négligeant les dérivées secondes de Φ dans la matrice hessienne $H[g](x)$ de g en x

$$H[g](x) = \mathrm{Jac}[\Phi](x)^\top \mathrm{Jac}[\Phi](x) + \sum_i \Phi_i(x) H[\Phi_i](x) \approx \mathrm{Jac}[\Phi](x)^\top \mathrm{Jac}[\Phi](x) =: \tilde{H}[g](x)$$

et en calculant ensuite un pas de Newton standard avec cette matrice hessienne approchée $\tilde{H}[g](x)$, c'est-à-dire

$$dx := -(\tilde{H}[g](x))^{-1}\,\nabla g(x) = -(\mathrm{Jac}[\Phi](x)^\top \mathrm{Jac}[\Phi](x))^{-1} \mathrm{Jac}[\Phi](x)^\top \Phi(x).$$

Cette méthode est raisonnable lorsque l'on veut minimiser une fonction de moindres carrés g où le nombre de carrés (la dimension de Φ) est plus grand que le nombre de paramètres (la dimension de x). Ici, le nombre de carrés est le même que le nombre de paramètres et on ne veut pas se contenter de minimiser g mais bien l'annuler ; il est donc probablement maladroit d'utiliser la méthode de Gauss-Newton. En effet, lorsque $\mathrm{Jac}[\Phi](x)$ est carrée et inversible (ainsi le pas de Newton pour annuler Φ, soit $-(\mathrm{Jac}[\Phi](x))^{-1}\Phi(x)$ est bien défini) la méthode de Gauss-Newton fournit exactement le même pas que la méthode de Newton pour annuler Φ :

$$\begin{aligned} dx &= -(\mathrm{Jac}[\Phi](x)^\top \mathrm{Jac}[\Phi](x))^{-1} \mathrm{Jac}[\Phi](x)^\top \Phi(x) \\ &= -\mathrm{Jac}[\Phi](x)^{-1} \mathrm{Jac}[\Phi](x)^{-\top} \mathrm{Jac}[\Phi](x)^\top \Phi(x) \\ &= -(\mathrm{Jac}[\Phi](x))^{-1} \Phi(x). \end{aligned}$$

Toutefois, ce calcul est fait de manière moins intelligente : le système linéaire à résoudre pour calculer le pas de Gauss-Newton

$$(\mathrm{Jac}[\Phi](x)^\top \mathrm{Jac}[\Phi](x))\, dx = -\nabla g(x) \tag{4.14}$$

est à la fois moins creux et moins bien conditionné, en général, que le système à résoudre pour calculer le pas de Newton

$$\mathrm{Jac}[\Phi](x)\, dx = -\Phi(x). \tag{4.15}$$

Les seuls avantages sont que

- le système (4.14) est symétrique, contrairement à (4.15), ce qui permet d'utiliser des algorithmes spécialisés pour le résoudre
- la méthode de Gauss-Newton permet d'envisager sa variante dite de Levenberg-Marquardt, qui consiste à remplacer $\tilde{H}[g](x)$ par $\tilde{H}[g](x) + \lambda \mathrm{I}$ où $\lambda > 0$ est une constante bien choisie qui décroît au cours des itérations.

Cette dernière idée rend nécessairement $\tilde{H}[g](x)$ inversible : elle permet donc de contourner le caractère singulier de la matrice $\mathrm{Jac}[\Phi](x)$ qui fait perdre son sens à la méthode de Newton sur Φ.

4.5.3 Méthode de quasi-Newton

Si l'on veut éviter de devoir évaluer la hessienne de g (ce qui est coûteux) ou de l'approcher comme dans la méthode de Gauss-Newton (ce qui dégrade le conditionnement et le caractère creux), il ne reste plus qu'à utiliser une méthode d'optimisation du premier ordre comme la méthode de quasi-Newton.

Asymptotiquement, on s'attend à ce que cette méthode converge super-linéairement sous réserve qu'au voisinage de l'optimum, la fonction objectif soit régulière et que sa matrice hessienne soit définie positive. Cependant, ces hypothèses n'ont pas de raison d'être réunies dans nos applications où la fonction objectif est non régulière. En pratique la décroissance de l'objectif est très lente et on observe rarement la convergence super-linéaire.

Cependant, la méthode de quasi-Newton a tout de même deux avantages : d'une part, elle ne nécessite pas de résoudre un système linéaire à chaque itération mais seulement d'effectuer un produit matrice-vecteur, elle est donc très économe en calculs. D'autre part elle possède des versions à mémoire limitée qui permettent de la rendre également économe en mémoire[1]. Malgré tout, nos expériences sur la fonction de Fischer-Burmeister au chapitre 5 sont très décevantes, et un travail supplémentaire est nécessaire pour savoir s'il est possible d'améliorer significativement les performances de cette méthode.

4.6 Méthode de Haslinger

Il existe plusieurs variantes [HT83, Has83, HKD04] de cette méthode, dont l'idée essentielle est de considérer la valeur de la force normale (qui intervient dans le seuil de

[1] L'idée de calculer des forces de contact par une méthode d'optimisation à mémoire limitée a déjà été proposée dans [FWA07], avec un modèle différent.

frottement) comme un paramètre. On présente ici celle qui consiste à effectuer des itérations de point fixe, en l'adaptant à « notre » problème incrémental (1.34). On généralise la définition (1.26) du demi-cylindre $T(s)$ au cas où $n > 1$ par

$$\forall\, s \in \mathbb{R}^n_+ \colon T(s) := \prod_{i=1}^{n} \{r^i \in \mathbb{R}^d \colon r^i_{\mathrm{N}} \geq 0,\ \|r^i_{\mathrm{T}}\| \leq s^i\}.$$

En reprenant la formulation (1.27), on voit que le problème incrémental (1.34) est équivalent à : trouver (u, r, s) dans $\mathbb{R}^{nd+nd+n}$ tel que

$$\begin{cases} r \in T(s) \\ u = Wr + q \in -N_{T(s)}(r) \\ s^i = \mu^i\, r^i_{\mathrm{N}} \quad \text{pour tout } i \in 1, \dots n. \end{cases} \tag{4.16}$$

Les deux premières équations constituent les conditions d'optimalité d'un programme quadratique sur $T(s)$, et le problème (4.16) est équivalent à

$$r \in \operatorname*{argmin}_{r \in T(s)} \frac{1}{2} r^\top W r + q^\top r \tag{4.17a}$$

$$s^i = \mu^i\, r^i_{\mathrm{N}} \quad \text{pour tout } i \in 1, \dots n. \tag{4.17b}$$

On peut ensuite tenter de résoudre le problème incrémental de la manière suivante : on résout (4.17a) à s fixé, ce qui fournit une valeur pour r ; puis on met à jour la valeur de s par une itération de point fixe sur (4.17b), *i.e.* on effectue $s^i \leftarrow \mu^i r^i_{\mathrm{N}}$, et on recommence. Contrairement à la méthode de point fixe sur la formulation d'Alart et Curnier (4.7) ou sur celle de De Saxcé (4.8), cette technique fonctionne extrêmement bien en pratique comme on le verra au chapitre 5.

Remarque 4.15 (Cas sans frottement)**.** Il n'est pas nécessaire d'introduire de variable supplémentaire s^i pour les contacts sans frottement : en effet, (4.17b) lui imposerait toujours une valeur nulle. L'équation de point fixe (4.17b) est donc limitée aux contacts frottants, et les autres sont automatiquement éliminés du problème. En particulier, si $\mu^i = 0$ pour tous les contacts, il suffit de résoudre une seule fois (4.17a) qui est un QP (et on retrouve le fait, vu à la sous-section 1.4.3, que dans le problème incrémental sans frottement est un QP). Ceci est un argument très fort en faveur de cette approche : dans le cas facile, sans frottement, elle attaque le problème de la bonne manière, c'est-à-dire en résolvant le programme quadratique par minimisation. La méthode que nous proposons dans la section suivante partage aussi cette propriété (mais ce n'est pas le cas des méthodes vues jusqu'à présent, comme celle d'Alart et Curnier).

Remarque 4.16 (Interprétation)**.** Chaque itération interne de cette méthode – le calcul de r dans (4.17a) à s fixé – peut-être vue comme la résolution du problème de frottement de Tresca avec un seuil s fixé. Cette technique consiste donc à résoudre une succession de problèmes de Tresca en ajustant la valeur du seuil par des itérations de point fixe qui visent à satisfaire également la loi de Coulomb.

Remarque 4.17 (Aspects numériques)**.** Cette méthode est attrayante du point de vue numérique pour les raisons suivantes.

- A la fin de chaque itération interne, on obtient $u \in -N_{T(s)}(r)$ donc $u_{\mathrm{N}} \geq 0$; on pourra donc se contenter d'une résolution tronquée du problème de point fixe sans pour autant introduire de pénétration dans le système. En revanche, il peut en résulter une violation de la contrainte $r \in L$.
- Le calcul explicite de W n'est pas obligatoire ; pour résoudre (4.17a), il suffit de savoir multiplier par W, ce qui ne nécessite pas de l'assembler.
- Cette méthode a été initialement proposée pour résoudre le problème de contact statique en élasticité, et dans ce contexte des démonstrations de convergence existent : on peut montrer que la fonction de point fixe est contractante. Malheureusement, les hypothèses nécessaires pour l'assurer dépendent (entre autres choses) du maillage [Has83]. Il serait extrêmement intéressant de voir si la démonstration du caractère contractant peut être adaptée à notre problème incrémental, et comment se traduisent ses hypothèses.

4.7 Méthode proposée

4.7.1 Motivation

Lorsqu'on introduit la variable $\widetilde{u}$, le problème incrémental se formule comme un problème de complémentarité. Ces problèmes sont bien connus en optimisation [FK98], ils apparaissent en particulier dans les conditions d'optimalité de Karush-Kuhn-Tucker. Les méthode usuelles pour les résoudre sont celles que l'on a présentées : on introduit des fonctions de complémentarité comme la fonction de Fischer-Burmeister, des fonctions de mérite, on les régularise, et on les utilise dans des algorithmes de type « points intérieurs », « central path », etc. L'idée que nous défendons dans la section suivante pour résoudre le problème incrémental est d'effectuer le cheminement inverse, et de voir la contrainte de complémentarité (assortie de l'équation dynamique et de l'équation cinématique) comme le problème d'optimisation convexe (3.11)-(3.13) couplé avec l'équation de point fixe (3.24). En ce sens, notre méthode ne rentre dans aucune des deux catégories ci-dessus (reformulation fonctionnelle et par optimisation), ou plutôt – comme la méthode de Haslinger décrite à la section 4.6 – elle mélange les deux idées : une partie du problème est traitée par minimisation (elle s'y prête bien, parce qu'elle présente de la convexité) et le reste par la méthode de Newton.

Remarque 4.18. Comme pour la méthode de Haslinger (voir la remarque 4.15), le cas particulier où tous les coefficients de frottement sont nuls illustre l'intérêt de notre approche. Dans cette situation, le problème incrémental est équivalent à un QP convexe ; voir l'équation (1.36). La méthode d'Alart et Curnier et ses variantes consistent alors à résoudre ce QP en appliquant la méthode de Newton aux équations KKT (ou plutôt à une reformulation fonctionnelle de ces équations). Ceci n'est certainement pas la bonne méthode pour s'attaquer à un programme quadratique : il est plus raisonnable d'utiliser des méthodes de minimisation comme la méthode « active set » ou celle des points

intérieurs par exemple, comme ceci est fait dans les solveurs spécifiques aux QP. Or, c'est exactement ce que fait notre méthode : dans le cas où tous les μ^i sont nuls, le problème d'optimisation (3.11) n'est plus couplé à un problème de point fixe et notre méthode consiste à résoudre une seule fois un QP à l'aide d'un solveur adapté. Autrement dit, comme celle de Haslinger, notre approche permet d'éliminer proprement les contacts sans frottements du problème.

Remarque 4.19 (Aspects numériques)**.** L'attrait numérique de cette méthode est comparable à celui de la méthode de Haslinger, exposé à la remarque 4.17.

- A la fin de chaque itération interne, on obtient cette fois $r \in L$; en revanche la contrainte $u_{\mathrm{N}} \geq 0$ peut être violée.
- Le calcul explicite de W peut être évité de la même manière.
- Par analogie avec la méthode de Haslinger, et au vu du succès des expériences numériques, on peut supposer que la fonction de point fixe F est « souvent » contractante. Un travail supplémentaire est nécessaire ici pour voir si ce fait peut être démontré, et sous quelles hypothèses.

On propose d'attaquer numériquement l'équation de point fixe (3.24) soit par des itérations de point fixe, soit par des itérations de l'algorithme de Newton. La mise en oeuvre de ce programme nécessite de savoir

- évaluer la fonction de point fixe F, c'est-à-dire résoudre le problème d'optimisation conique (3.11) ou son dual (3.13)
- différentier F (si l'on veut utiliser la méthode de Newton)
- retrouver la solution complète $(v, \widetilde{u}, r)$ à partir de la solution de (3.11) et (3.13) une fois que l'équation de point fixe $F(s) = s$ est satisfaite.

Dans toute cette section, les notations sont celles du chapitre 3.

4.7.2 Résolution du problème interne

L'évaluation de $F(s)$ pour une valeur donnée de s nécessite de calculer $\widetilde{u}(s)$, ce qui revient à résoudre l'un des deux problèmes (3.11) ou (3.13). Ceci peut être fait à l'aide d'algorithmes spécifiques, au prix d'un coût de développement supplémentaire, ou en utilisant une reformulation sous forme d'un problème d'optimisation classique (SOCP) pour lequel des codes existants sont disponibles. Selon nos expériences, les gains obtenus avec les algorithmes spécifiques sont très importants et justifient largement l'effort de développement, mais l'approche plus simple par reformulation SOCP est également décrite ci-dessous.

Algorithmes spécifiques

Le problème (3.13) est particulièrement simple, puisque la contrainte est de la forme $r \in L$ (où r est la variable) alors que le problème (3.11) est sous la forme $g(v) \in L^*$ (où v est la variable). En contrepartie, la fonction quadratique que l'on minimise est fortement convexe pour (3.11) alors qu'elle n'est que convexe au sens large pour (3.13).

Dans des expériences préliminaires, nous nous sommes contentés de résoudre (3.13) à l'aide d'un simple algorithme de descente de gradient avec projection et recherche linéaire d'Armijo, avec des résultats numériques bien meilleurs que ceux obtenus avec la formulation SOCP ci-dessous et le solveur SeDuMi. Il est probable que si l'on disposait de méthodes plus sophistiquées, comme celles décrites dans [Kuc07], l'amélioration serait encore plus significative. Dans les expériences numériques du chapitre 5, on se contentera de l'algorithme du gradient projeté pour résoudre (3.13).

Remarque 4.20. L'algorithme du gradient projeté ne fournit qu'une solution au problème dual (3.13), c'est-à-dire la valeur des forces ; en revanche, un solveur comme SeDuMi fournit un couple de solutions primal-dual. La sous-section 4.7.3 explique pourquoi cette propriété est intéressante : dans le cas où l'on résout numériquement le problème en v (3.11), elle permet de retrouver la valeur des forces r dans le système. Si l'on ne dispose pas d'une telle solution duale, il faut se satisfaire d'une solution en vitesses $(\widetilde{u}, v)$ et de l'assurance qu'un vecteur de forces r correspondant existe.

Reformulation SOCP

Les deux problèmes (3.11) et (3.13), qui sont des programmes quadratiques sur le cône du second ordre (SOQP), peuvent être reformulés comme des problèmes *linéaires* sur le cône du second ordre (SOCP).

Lemme 4.21 (Reformulation SOCP du problème dual)**.** *Le problème (3.13) est équivalent au programme linéaire sur le cône du second ordre suivant, dont les inconnues sont (t, z, x, y, r) dans $\mathbb{R}^{1+1+m+1+nd}$:*

$$\begin{cases} \min \frac{t}{2} + b^\top r \\ z = \frac{t+1}{2} \\ y = \frac{t-1}{2} \\ Cx = H^\top r \\ (z, (x, y)) \in K_{\mu=1} \\ r \in L \end{cases} \tag{4.18}$$

où $b := w + Es - HM^{-1}f$ et C est une matrice telle que $CC^\top = M$. Autrement dit, r est solution de (3.13) si et seulement s'il existe (t, x, y, z) tels que (t, x, y, z, r) soit solution de (4.18).

Démonstration. La fonction-objectif du problème (3.13) s'écrit

$$\frac{1}{2} r^\top H M^{-1} H^\top r + b^\top r.$$

Effectuons une décomposition de M sous la forme $M = CC^\top$ (par exemple la décomposition de Choleski). On a donc $M^{-1} = C^{-\top} C^{-1}$. La fonction-objectif devient

$$\frac{1}{2} \|x\|^2 + b^\top r$$

avec $x = C^{-1}H^\top r$. On utilise ensuite une astuce [BTN01] qui consiste à pousser le terme quadratique de la fonction-objectif dans les contraintes en introduisant une variable supplémentaire t soumise à

$$t \geq \|x\|^2 \iff \left(\frac{t+1}{2}\right)^2 \geq \left(\frac{t-1}{2}\right)^2 + \|x\|^2 \iff \frac{t+1}{2} \geq \|\left(x, \frac{t-1}{2}\right)\|$$

où on remarque que la dernière contrainte est une contrainte de cône du second ordre. Autrement dit, (3.13) devient (4.18) qui est un SOCP. ∎

Intéressons-nous maintenant au problème primal (3.11).

Lemme 4.22. *Le problème d'optimisation (3.11) est équivalent au programme linéaire sur le cône du second ordre suivant, dont les inconnues sont (t, y, v) dans $\mathbb{R}^{1+m+m}$:*

$$\begin{cases} \min t \\ (t, y) \in K_{\mu=1} \\ Hv + w + Es \in L^* \\ y = C^\top v + C^{-1} f \end{cases} \tag{4.19}$$

où C est une matrice telle que $CC^\top = M$.

Démonstration. La fonction-objectif $J(v)$ de (3.11) s'écrit

$$\begin{aligned} J(v) &= \frac{1}{2} v^\top M v + f^\top v \\ &= \frac{1}{2}[(C^\top v)^\top (C^\top v) + 2(C^\top v)^\top (C^{-1} f)] \\ &= \frac{1}{2}\|C^\top v + C^{-1} f\|^2 + \text{cte}. \end{aligned} \tag{4.20}$$

En posant $y = C^\top v + C^{-1} f$ et en introduisant la variable supplémentaire t soumise à $t \geq \|y\|$, on obtient la reformulation SOCP (4.19) de (3.11). ∎

Remarque 4.23. Bien entendu, il n'est pas nécessaire d'évaluer C^{-1} (qui apparaît dans les contraintes) mais seulement $C^{-1} f$, qui revient à résoudre un système triangulaire.

Cette technique de reformulation SOCP n'est pas très rapide, mais elle présente deux avantages : d'une part, il existe déjà des solveurs SOCP de bonne qualité (alors que nous ne connaissons pas de solveur dédié aux SOQP (3.11) et (3.13) qui soit disponible publiquement). D'autre part, les solveurs SOCP (comme SeDuMi) ne fournissent en général pas seulement une solution primale, mais un couple de solutions primal-dual. Grâce à cette propriété, il est possible de retrouver facilement la valeur de r lorsqu'on a résolu le problème en v (3.11), ou la valeur de $(v, \widetilde{u})$ lorsqu'on a résolu le problème en r (3.13). C'est à ce problème que nous nous attachons maintenant.

4.7.3 Reconstruction des inconnues et évaluation de F

Problème en r. La résolution du problème (3.13) fournit une valeur pour les forces de contact r. A l'aide de la première équation du problème incrémental (1.33), on peut alors retrouver v en résolvant le système linéaire symétrique, en la variable v

$$Mv = (H^\top r - f) = \text{cte}.$$

On retrouve ensuite u grâce à $u = Hv + w$, ce qui permet d'évaluer $F^i(s) := \|u^i_{\mathrm{T}}(s)\|$.

Cependant, admettons que le programme qui résout (3.13) fournisse non seulement une solution primale, mais aussi une solution duale ; on aimerait retrouver directement la valeur de $\widetilde{u}$ et v à partir de la solution duale sans avoir à résoudre un système linéaire.

Problème en v. D'autre part, si l'on résout (3.11), on obtient la valeur de v (solution primale) qui suffit à calculer celle de $\widetilde{u} = Hv + w + Es$ et donc à évaluer F. Cependant, si l'on parvient à résoudre l'équation de point fixe $F(s) = s$, on obtiendra seulement une solution partielle $(\widetilde{u}, v)$ du problème incrémental, et l'assurance qu'un r correspondant existe. Ce n'est pas satisfaisant du point de vue mécanique, car on peut légitimement se demander quelles sont les forces de contact dans notre système.

Cette situation est plus difficile que pour le problème en r (3.13), où l'on devait « seulement » résoudre un système linéaire $Mv = \text{cte}$. On veut maintenant trouver r tel que

$$H^\top r = Mv + f = \text{cte} \quad \text{et} \quad L \ni r \perp \widetilde{u} = \text{cte} \in L^*$$

où v et $\widetilde{u}$ sont connus. Il s'agit donc de résoudre un problème qui comporte deux contraintes linéaires ($H^\top r = \text{cte}$ et $r \perp \widetilde{u}$) et une contrainte conique ($r \in L$) ; sa résolution n'est pas immédiate.

On préférerait donc là aussi obtenir directement la valeur des forces à partir de la solution duale de (3.11) calculée par le solveur en même temps que la solution primale.

La sous-section suivante est consacrée à cette reconstruction à partir des reformulations SOCP (4.19) et (4.18). Ce n'est qu'un exemple : quelle que soit la reformulation utilisée pour résoudre (3.11) et (3.13), on peut raisonnablement penser que la solution duale du problème reformulé « contiendra » d'une manière ou d'une autre la solution complète $(v, \widetilde{u}, r)$ de (3.5).

4.7.4 Reconstruction à partir de la formulation SOCP

Considérons donc de nouveau le problème (4.18) dans lequel on introduit des multiplicateurs comme ci-dessous, dont on suppose que la valeur finale est donnée par le

solveur : [2] trouver une solution (t, z, x, y, r) dans $\mathbb{R}^{1+1+m+1+nd}$ de

$$\begin{cases} \min \frac{t}{2} + b^\top r & \\ z = \frac{t+1}{2} & \longleftarrow \delta \in \mathbb{R} \\ y = \frac{t-1}{2} & \longleftarrow \delta' \in \mathbb{R} \\ Cx = H^\top r & \longleftarrow \lambda \in \mathbb{R}^m \\ (z,(x,y)) \in K_{\mu=1} & \longleftarrow (\alpha,(\beta,\gamma)) \in K_{\mu=1} \\ r \in L & \longleftarrow \nu \in L^*. \end{cases} \tag{4.21}$$

On va montrer que la valeur de $\widetilde{u}$ est donnée sans aucun calcul par celle de la variable duale ν. Malheureusement, on n'obtient pas directement la valeur de v et il faut résoudre le système linéaire en M ; mais si l'on dispose d'une « bonne décomposition » $CC^\top$ de M, ce calcul est rapide.

Lemme 4.24. *Soient (t, z, x, y, r) et $(\delta, \delta', \lambda, \alpha, \beta, \gamma, \nu)$ un couple de solutions primal-dual de (4.21). Alors*

$$v := M^{-1}(H^\top r - f)$$

est tel que (v, ν, r) résout (3.5).

Démonstration. Par construction, on a $t = \|x\|^2$ à l'optimum. La contrainte de cône du second ordre sur $(z,(x,y))$ est donc saturée et (x,y) n'est pas nul (car $x = 0$ implique $y = -1/2$) donc $(z,(x,y))$ n'est pas au sommet du cône et le multiplicateur associé (α,β,γ) est défini à une constante multiplicative $\chi \geq 0$ près par

$$(\alpha,(\beta,\gamma)) = \chi(\|(x,y)\|, -z\frac{(x,y)}{\|(x,y)\|}).$$

Par construction, $\|(x,y)\| = z$ donc $(\alpha,(\beta,\gamma)) = \chi(z,(-x,-y))$. D'autre part, l'équation d'optimalité s'écrit

$$\begin{bmatrix} t \\ z \\ x \\ y \\ r \end{bmatrix} : \begin{bmatrix} 1/2 \\ 0 \\ 0 \\ 0 \\ b \end{bmatrix} + \begin{bmatrix} 1/2 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \delta + \begin{bmatrix} 1/2 \\ 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} \delta' + \begin{bmatrix} 0 \\ 0 \\ -C^\top \\ 0 \\ H \end{bmatrix} \lambda - \begin{bmatrix} 0 \\ \alpha \\ \beta \\ \gamma \\ \nu \end{bmatrix} = 0.$$

On en tire facilement $\delta = -\alpha$, $\delta' = -\gamma$ puis $\alpha + \gamma = 1$. Ceci fixe la valeur de χ : $1 = \alpha + \gamma = \chi(z - y) = \chi$ donc $\chi = 1$ et $\alpha = z$, $\beta = -x$ et $\gamma = -y$. D'autre part, l'équation d'optimalité donne directement $\lambda = -C^{-\top}\beta = C^{-\top}x$ et $b + H\lambda = \nu$; en

[2] Dans certains cas pathologiques, il peut ne pas y avoir de solution duale même quand le primal est réalisable et admet un minimum ; une hypothèse supplémentaire est requise, comme par exemple celle de Slater.

utilisant le fait que $x = C^{-1}H^\top r$, on a

$$\begin{aligned}\nu &= b + HC^{-\top}x \\ &= b + HC^{-\top}C^{-1}H^\top r \\ &= HM^{-1}H^\top r + w + Es - HM^{-1}f \\ &= HM^{-1}(H^\top r - f) + w + Es \\ &= Hv + w + Es \\ &= \widetilde{u}.\end{aligned} \tag{4.22}$$

■

On a donc bien directement la valeur de $\widetilde{u}$ en posant $\widetilde{u} := \nu$. Ce résultat est assez intuitif, car du point de vue mécanique on s'attend bien à ce que la variable duale associée à la force r soit la vitesse $\widetilde{u}$ (qui vérifie $r \perp \widetilde{u}$).

Considérons maintenant le problème (4.19), dans lequel on introduit des multiplicateurs comme ci-dessous, et on suppose toujours que leur valeur finale est fournie par le solveur : trouver une solution $(t, y, v, \widetilde{u})$ dans $\mathbb{R}^{1+m+m+nd}$ de

$$\left\{\begin{array}{ll} \min t & \\ (t,y) \in K_{\mu=1} & \longleftarrow (\alpha,\beta) \in K_{\mu=1} \\ \widetilde{u} \in L^* & \longleftarrow \nu \in L \\ -\widetilde{u} + Hv + w + Es = 0 & \longleftarrow \lambda_1 \in \mathbb{R}^{nd} \\ -y + C^\top v + C^{-1}f = 0 & \longleftarrow \lambda_2 \in \mathbb{R}^m. \end{array}\right. \tag{4.23}$$

Le résultat suivant est similaire au lemme 4.24, à ceci près qu'une constante multiplicative intervient : il dit que dans (3.11), dont la solution primale est v, on peut obtenir la valeur des forces de contact (*une* valeur possible, il n'y a pas unicité pour r) à partir de la solution duale associée à la contrainte $\widetilde{u} \in L^*$.

Lemme 4.25. *Soient $(t, y, \widetilde{u}, v)$ et $(\alpha, \beta, \nu, \lambda_1, \lambda_2)$ un couple de solutions primal-dual de (4.23). Alors $(v, \widetilde{u}, \|y\|\nu)$ est solution de (3.5).*

Démonstration. L'équation d'optimalité s'écrit

$$\begin{bmatrix} t \\ y \\ \widetilde{u} \\ v \end{bmatrix} : \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -\mathrm{I} \\ H^\top \end{bmatrix} \lambda_1 + \begin{bmatrix} 0 \\ -\mathrm{I} \\ 0 \\ C \end{bmatrix} \lambda_2 - \begin{bmatrix} \alpha \\ \beta \\ \nu \\ 0 \end{bmatrix} = 0.$$

On en tire directement $\alpha = 1$, $\lambda_2 + \beta = 0$, $\lambda_1 + \nu = 0$, $H^\top\lambda_1 + C\lambda_2 = 0$ donc $H^\top\nu = -C\beta$. Par construction, $t = \|y\|$ à l'optimum et la contrainte de cône $(t, y) \in K_{\mu=1}$ est saturée.

– Si $\|y\| \neq 0$, (t, y) ne se trouve pas au sommet du cône et la variable duale (α, β) est déterminée à une constante multiplicative près $\chi \geq 0$ par $(\alpha, \beta) = \chi(\|y\|, -t\frac{y}{\|y\|})$. Comme $\alpha = 1$, on a $\chi = 1/\|y\|$ et donc $\beta = -y/\|y\|$. On en déduit

$$H^\top\nu = -C\beta = C\frac{y}{\|y\|} = \frac{Mv + f}{\|y\|}.$$

– Si $\|y\| = 0$, alors $y = 0$ *i.e.* $Mv + f = 0$ donc $r = 0$ est une solution possible.
Autrement dit, une solution possible pour r est $r = \|y\|\nu$ dans les deux cas. ■

Conformément à l'intuition, la solution duale associée à la contrainte $\widetilde{u} \in L^*$ fournit bien la force r, via le facteur de proportionnalité connu $\|y\|$.

4.7.5 Différentiation de F

La fonction F étant définie par la formule $F^i(s) = \|u_{\mathrm{T}}^i(s)\|$, c'est-à-dire de manière plus explicite

$$F^i(s) := \| P_{\mathrm{T}}^i \left(Hv(s) + w\right) \|$$

où P_{T}^i est la matrice de la projection sur le plan tangent au i-ème contact, on voit que F est différentiable en s lorsque v est différentiable en s et que $P_{\mathrm{T}}^i \left(Hv(s) + w\right)$ est non nul pour tout i. Comme $v(s)$ est calculée comme solution d'un problème d'optimisation paramétrique qui dépend de s, on voudrait différentier v par rapport à s de la manière classique : en écrivant les conditions d'optimalité (3.5) au point s courant, en les linéarisant par rapport à s (ce qui donne (v, u, r) comme solution d'un système linéaire dépendant linéairement de s) puis en inversant ce système linéaire. C'est l'idée générale du théorème des fonctions implicites.

Malheureusement ici, on sait que la solution $r(s)$ n'est pas unique en général à s fixé, ce qui signifie que l'hypothèse d'inversibilité locale du théorème des fonctions implicites ne sera *pas satisfaite*. Malgré cela, le problème posé est raisonnable : on ne s'intéresse pas à la différentiation de (v, u, r) par rapport à s (qui n'aurait pas de sens, car $r(s)$ n'est pas défini de manière unique) mais seulement à la différentiation de $v(s)$ qui, lui, est défini de manière unique. Autrement dit, il nous faudrait une version élaborée du théorème des (multi-)fonctions implicites pour pouvoir justifier le calcul de la jacobienne de F. Pour implémenter la différentiation de F, on se contentera donc du calcul formel présenté ci-dessous, sans plus de justification théorique.

Linéarisation de la contrainte de complémentarité

Dans (3.5) que l'on veut linéariser, les deux premières lignes sont déjà linéaires et la troisième est la contrainte de complémentarité

$$K_i \ni r^i \perp \widetilde{u}^i \in K_i^*. \tag{4.24}$$

Pour pouvoir la linéariser, on la remplace par une contrainte fonctionnelle équivalente, comme

$$\mathrm{P}_K(r^i - \widetilde{u}^i) - r^i = 0$$

que l'on sait différentier explicitement (annexe E). Au voisinage de $(\bar{\widetilde{u}}, \bar{r})$, on linéarise donc (4.24) par

$$(\mathrm{I} - \frac{\partial \mathrm{P}_K}{\partial x})\, dr + \frac{\partial \mathrm{P}_K}{\partial x}\, d\widetilde{u} = 0. \tag{4.25}$$

On pourrait maintenant essayer de calculer explicitement la matrice jacobienne de $s \to v(s)$ en un certain point $\bar{s}$ pour lequel on connaît la solution $(v, u, r)(s)$ en utilisant l'idée du théorème des fonctions implicites : il suffirait de remplacer l'équation de complémentarité (4.24) par sa linéarisation (4.25) dans (3.5) pour obtenir le système suivant, où $X := (v, \widetilde{u}, r)$

$$A_X\, dX + A_s\, ds = 0, \quad \text{avec} \quad A_X = \begin{bmatrix} M & 0 & -H^\top \\ -H & \mathrm{I} & 0 \\ 0 & \dfrac{\partial \mathrm{P}_K}{\partial x} & (\mathrm{I} - \dfrac{\partial \mathrm{P}_K}{\partial x}) \end{bmatrix} \quad \text{et} \quad A_s = \begin{bmatrix} 0 \\ E \\ 0 \end{bmatrix}. \tag{4.26}$$

Si A_X était inversible, on aurait en éliminant X

$$dX = -A_X^{-1} A_s\, ds;$$

autrement dit, on aurait la matrice jacobienne de $s \to X(s)$ dont on pourrait extraire la matrice jacobienne de $s \to v(s)$. Cependant, il y a deux problèmes : d'une part la matrice A n'est pas inversible en général en raison du caractère multivoque de $s \to r(s)$, et d'autre part on ne veut pas calculer la matrice jacobienne de v : tout ce qui nous intéresse, *in fine*, c'est le calcul d'un pas de Newton pour l'équation $F(s) = s$. Ceci ne nécessite heureusement pas d'assembler la jacobienne de F, comme on va le voir.

Calcul d'un pas de Newton

On voudrait maintenant calculer un pas de Newton pour l'équation $F(s) = s$. On écrit le système linéarisé en $X = (v, \widetilde{u}, r)$ et s :

$$\left\{ \begin{array}{lcll} F + dF & = & s + ds & \text{(équation de Newton)} \\ A_X\, dX + A_s\, ds & = & 0 & \text{(KKT linéarisé)} \\ F^i\, dF^i & = & (\widetilde{u}^i_{\mathrm{T}})^\top d\widetilde{u}^i & \text{(différentiation de } F\text{).} \end{array} \right. \tag{4.27}$$

En éliminant dF grâce à la dernière équation et en rassemblant les inconnues à gauche et les constantes à droite, on obtient le système linéaire

$$\left\{ \begin{array}{lcl} -\mathrm{Diag}((\widetilde{u}^i_{\mathrm{T}})^\top / F^i)\, d\widetilde{u} + ds & = & F - s \\ A_X\, dX + A_s\, ds & = & 0. \end{array} \right. \tag{4.28}$$

En admettant qu'on trouve une solution $(dX,\, ds)$ au système linéaire (4.28), on obtient alors un nouvel itéré $s + ds$, ce qui achève une itération de la méthode de Newton.

Bien entendu, la solution r n'étant pas unique en général, le système (4.28) n'est en général pas inversible. On le résout donc au sens des moindres carrés et on vérifie *a posteriori* la qualité de la solution obtenue. D'après nos expériences, la solution trouvée est toujours excellente. Ceci est conforme avec l'intuition selon laquelle la non-unicité de r n'exerce pas d'influence néfaste grâce au fait que cette variable n'apparaît pas dans la définition de la fonction F (F ne dépend que de $\widetilde{u}$, qui est unique).

4.7.6 Interprétation géométrique

Pour schématiser, représentons la situation dans le plan (X, s). Le problème de complémentarité conique (3.5) peut être vu comme une contrainte $g(X, s) = 0$ et la contrainte $\|u_{\mathrm{T}}^i\| = s^i$ comme une contrainte $h(X, s) = 0$. Les solutions cherchées se trouvent à l'intersection de ces deux lignes de niveau (Figure 4.1). Notre méthode consiste à effectuer

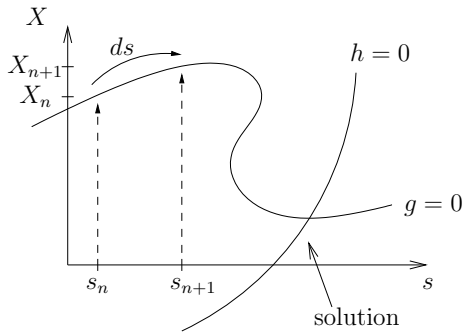


FIG. 4.1 – Vue schématique

la méthode de Newton en se restreignant au sous-ensemble $\{g(X, s) = 0\}$ du plan (X, s). A chaque fois qu'une nouvelle valeur de s est proposée par l'algorithme qui résout (3.24), que ce soit une itération de point fixe, de Newton ou autre, le problème interne ramène immédiatement le point courant sur la courbe $\{g(X, s) = 0\}$.

4.7.7 Cas bidimensionnel

Quand $d = 2$, toutes les contraintes sont linéaires et on peut théoriquement résoudre le problème en temps fini : il suffit de le reformuler comme un LCP et de le résoudre avec un solveur énumératif (ce qui prend en général un temps exponentiel). On pourrait croire que notre méthode itérative, qui *a priori* ne termine pas, est inadaptée pour résoudre un tel problème.

En réalité, en dimension $d = 2$ la solution $v(s)$ du problème (3.11) est une fonction affine par morceaux de s car les contraintes de ce problème quadratique sont linéaires et dépendent linéairement de s. Par conséquent, F est elle-même affine par morceaux ! Ceci signifie que si la méthode de Newton appliquée à l'équation de point fixe $F(s) = s$ trouve le « bon morceau », elle converge (de manière exacte, en théorie) en une seule itération. Autrement dit, si la méthode converge, elle le fait en un nombre fini d'itérations.

Ceci n'est plus vrai, par contre, si l'on essaie de résoudre l'équation $F(s) = s$ en effectuant des itérations de point fixe.

4.8 Approches locales et globales

Les méthodes présentées jusqu'à présent sont « globales », en ce sens qu'elles tentent de résoudre le problème incrémental d'un bloc, en effectuant des itérations qui modifient toutes les variables (les forces à tous les points de contact, par exemple) à la fois.

Moreau [Mor88, Mor03], Jourdan-Alart-Jean [JAJ98] utilisent un algorithme qui résout le problème du frottement « localement », c'est-à-dire contact par contact, et itère jusqu'à ce qu'une solution globale soit trouvée. Cette approche évoque la méthode de Gauss-Seidel pour résoudre les systèmes linéaires, dans laquelle on résout successivement des sous-systèmes plus petits, les autres variables étant fixées.

4.8.1 Inconvénients des méthodes fonctionnelles globales

Les méthodes fonctionnelles globales, visant à annuler une certaine fonction F qui dépend de toutes les variables, ne fonctionnent bien que sur certains problèmes ; elle sont très efficaces [CKPS98] en particulier sur les problèmes de corps déformables discrétisés par la méthode des éléments finis, ce que l'on peut expliquer en partie par les résultats théoriques du chapitre 3 (voir la sous-section 3.4.7). Cependant, sur d'autres systèmes mécaniques parfois très simples (en particulier lorsqu'on utilise des solides rigides), ces méthodes ne sont pas robustes : leurs faiblesses théoriques (jacobienne mal définie, non-inversible, direction non-descendante, cyclage) sont effectivement rencontrées en pratique.

Voici deux manières de contourner le problème.

- On peut conserver l'idée de minimisation du critère des moindres carrés $\|F\|^2$, en utilisant des algorithmes insensibles au caractère non-inversible de la jacobienne (contrairement à l'algorithme de Newton sur F) ; on a vu par exemple la méthode de Levenberg-Marquardt (sous-section 4.5.2).
- Résoudre le problème contact par contact, de manière itérative, en s'inspirant de la méthode de Gauss-Seidel pour la résolution des systèmes linéaires.

La seconde idée est l'objet de la sous-section suivante.

4.8.2 Approche contact-par-contact, dite « de Gauss-Seidel »

L'idée consiste à résoudre le problème incrémental de la manière suivante. On initialise r (avec la valeur des forces au pas de temps précédent ou en suivant une heuristique quelconque), puis on résout le problème du frottement au contact i seulement en maintenant les forces aux autres points de contact fixées ; si les variables u et v on été éliminées comme dans (1.34), ce problème est de petite taille et peut être résolu très rapidement. On itère ensuite sur i jusqu'à convergence, que l'on détecte en s'assurant que les variables ne changent plus d'une itération à la suivante (à une certaine tolérance près). Comme les forces ne sont *a priori* pas uniques, il est plus prudent de choisir comme critère d'arrêt la stabilisation des valeurs des vitesses. De nombreuses variantes de cet algorithme sont envisageables, par exemple celle qui suit.

1. Initialiser r
2. Tant que $N_{\text{iter}} \leq N_{\text{itermax}}$
   (a) $u_{\text{old}} \leftarrow Wr + q$
   (b) Pour $i = 1 \dots n$
      i. $u^i \leftarrow (Wr + q)^i$

ii. $\bar{q}^i \leftarrow u^i - W_{ii} r^i$

iii. Trouver r^i tel que $(W_{ii}\, r^i + \bar{q}^i, r^i) \in \mathcal{C}(e^i, \mu^i)$

(c) FinPour

(d) $u_{\text{new}} \leftarrow Wr + q$

(e) Si $\|u_{\text{old}} - u_{\text{new}}\| \leq \text{tol}$, sortir

3. FinTantQue

La boucle interne, au point 2-(b), met à jour la valeur de r en faisant une passe sur chacun des points de contacts. A chaque itération, un problème réduit à un seul contact est résolu au point 2-(b)-iii. On effectue cette résolution en utilisant n'importe laquelle des méthodes présentées jusqu'ici (comme la méthode d'Alart et Curnier), que l'on initialise avec la valeur courante de r^i. Cette résolution interne est très rapide (typiquement, quelques itérations) et peut être tronquée. On verra au chapitre 5 que cette approche est assez robuste, mais peut s'avérer très lente. De plus, elle manque de fondement théorique et la convergence n'en est pas garantie ; en pratique, elle est d'ailleurs régulièrement sujette au problème du cyclage.

Orientation

Avant de passer à l'implémentation, il faut encore spécifier plusieurs parties ou paramètres de chacun des algorithmes décrits dans ce chapitre : choix du solveur interne, type de recherche linéaire, nombre d'itérations internes maximum autorisées, initialisation pour le redémarrage à chaud, choix des variables à éliminer (u ou v et u) dans le système linéaire qui calcule le pas de Newton, etc. Nous sommes loin d'avoir exploré toutes les possibilités, qui sont fort nombreuses. Cependant, nous avons implémenté un certain nombre de méthodes que nous espérons représentatives ; afin d'évaluer les qualités de notre algorithme et de le comparer aux méthodes existantes, nous allons tester sa robustesse et sa rapidité dans le chapitre suivant sur des jeux de données variées.

Chapitre 5

Expériences numériques

A défaut de pouvoir comparer les qualités théoriques des algorithmes décrits dans le chapitre précédent, par manque de résultats de convergence, nous allons mener dans celui-ci une étude empirique consacrée essentiellement à leur robustesse d'une part, et leur vitesse d'autre part. Selon le système mécanique que l'on cherche à simuler, d'autres critères peuvent s'avérer pertinents ; par exemple, le caractère plus ou moins économe en mémoire d'une méthode est crucial lorsqu'on s'attaque à des problèmes de grande taille. Pour limiter notre champ d'investigation, nous restreindrons notre étude à deux familles de problèmes-tests : des instances aléatoires et des problèmes de tiges déformables. Tous les exemples considérés seront en dimension $d = 3$. Enfin, tous les algorithmes seront implémentés dans le même langage (Scilab) et les tests faits sur la même machine (un Dell Precision 380 à 3GHz doté de 2 Go de mémoire vive).

5.1 Problèmes-tests

5.1.1 Problèmes classiques

Dans les nombreux articles et ouvrages existants consacrés aux méthodes numériques pour les problèmes de contact et frottement, on retrouve très souvent deux types de problèmes-tests : les matériaux granulaires et les solides déformables en élasticité linéaire. Dans le premier cas, on considère un grand nombre de corps rigides (des sphères, dans le cas le plus simple, parfois des polyèdres) soumis à la gravité en présence d'objets extérieurs mobiles ou non, qui contraignent leur position. Dans le deuxième cas, on considère un objet discrétisé en espace par la méthode des éléments finis, qui repose sur le sol et est soumis à un chargement extérieur quelconque. Ces deux types de problèmes-tests ont l'avantage d'être assez simples à mettre en oeuvre. Pour les corps rigides, la modélisation est immédiate (il suffit d'écrire les équations d'Euler pour chaque solide) tandis que pour les déformables, on peut tirer directement les données du problème incrémental d'un logiciel existant de calcul par éléments finis. De même, la détection de collisions est très simple lorsqu'on considère uniquement des contacts point-plan (comme pour un objet déformable reposant sur le sol) ou entre deux sphères (il suffit de comparer

la distance des centres à la somme des rayons) ; en revanche elle devient plus délicate lorsqu'on considère un milieu granulaire formé de polyèdres.

Ces problèmes de matériaux granulaires et d'élasticité linéaire sont intéressants, mais d'une part ils ont déjà été beaucoup étudiés, et d'autre part ils sont très particuliers et présentent chacun une caractéristique qui les rend plus « faciles » (toutes proportions gardées) que le cas général : dans le cas des granulaires, la matrice de masse est généralement bien conditionnée (à moins de considérer des objets élémentaires très allongés, ou de tailles très différentes) ; dans le cas des objets élastiques, si on ne déclare de contact qu'au niveau des noeuds, on sait que le critère d'existence fort (H) du chapitre 3 est automatiquement vérifié ; le problème possède nécessairement une solution, et il est « loin » de l'infaisabilité. Enfin, la matrice de masse est constante dans les deux cas au cours des itérations et même diagonale pour les matériaux granulaires, ce qui rend le calcul de W (équation (1.34)) pratiquement gratuit.

Nous allons donc nous concentrer sur deux familles de problèmes moins étudiés jusqu'à présent dans la communauté de la mécanique du contact : les problèmes aléatoires et les problèmes de tiges.

5.1.2 Instances aléatoires

Les notations M, f, H, w, e^i et μ^i sont toujours celles du problème incrémental (1.33) du chapitre 1. On construit des problèmes aléatoires de la manière suivante. On choisit le nombre de sous-systèmes mécaniques et leur nombre de degrés de liberté. Pour chaque sous-système on génère aléatoirement une matrice de masse symétrique définie positive (dont on peut choisir le conditionnement pour rendre le problème plus ou moins difficile). Ceci permet de construire la matrice de masse globale M, à laquelle on associe un vecteur f aléatoire. Ensuite on choisit aléatoirement des paires de sous-systèmes, que l'on déclare en contact. On génère aléatoirement les blocs de H et w correspondants, ainsi que la direction normale e^i et le coefficient de frottement μ^i.

Il ne reste qu'à spécifier la loi de ces tirages aléatoires. En l'absence de phénomène sous-jacent réellement aléatoire, rien ne permet de préférer une loi à une autre ; la seule contrainte à respecter est que M soit définie positive. Dans nos expériences, nous avons utilisé la loi uniforme pour générer les variables à valeurs dans un segment (les coefficients de frottements, tirés entre $\mu_{min} = 0.2$ et $\mu_{max} = 0.8$) et des lois gaussiennes pour générer chacune des variables scalaires définissant les autres données.

La liste des instances aléatoires de petite taille que nous avons utilisées est rapportée dans la table 5.1. Le nom, de la forme « alea-m-n », contient la valeur m du nombre de degrés de liberté et celle n du nombre de points de contact. On donne le nombre m_{ss} de degrés de liberté des sous-systèmes, et leur nombre n_{ss}. La valeur du rapport $(nd)/m$ est aussi indiquée, car on soupçonne que le problème devienne plus difficile quand ce rapport augmente, en particulier s'il dépasse 1 (voir la sous-section 3.4.9). La densité de W donne une idée de la taille du problème en mémoire, et son conditionnement est aussi un indicateur de la difficulté du problème (en particulier pour notre approche, qui nécessite de résoudre un programme quadratique de matrice W). Enfin, lorsque nous avons pu vérifier le critère d'existence en utilisant l'une des méthodes proposées en 3.5,

Nom	m_{ss}	n_{ss}	$(nd)/m$	dens(W)	cond(W)	existence
alea-42-5	6	7	0.36	0.76	1.1e+02	oui
alea-48-8	6	8	0.50	0.50	5.2e+03	oui
alea-54-11	6	9	0.61	0.40	3.6e+05	oui
alea-60-14	6	10	0.70	0.47	5.9e+03	oui
alea-66-17	6	11	0.77	0.33	4.2e+03	oui
alea-72-20	6	12	0.83	0.31	7.3e+04	oui
alea-78-23	6	13	0.88	0.30	1.1e+04	oui
alea-84-26	6	14	0.93	0.25	4.1e+07	oui
alea-90-29	6	15	0.97	0.25	9.1e+08	oui
alea-96-32	6	16	1.00	0.25	1e+18	?
alea-102-35	6	17	1.03	0.20	2.7e+17	?
alea-108-38	6	18	1.06	0.19	7.6e+17	?
alea-114-41	6	19	1.08	0.19	9.3e+17	?
alea-120-44	6	20	1.10	0.18	3.5e+17	?
alea-126-47	6	21	1.12	0.18	6.6e+17	?

TAB. 5.1 – Instances aléatoires de petite taille

ceci est indiqué par un « oui » dans la dernière colonne de 5.1. On dispose alors d'un certificat qui prouve que l'hypothèse (H) est vérifiée. Inversement, si les tentatives de vérification du critère échouent, on ne dispose pas d'un certificat de non-validité du critère. On indique donc cet échec par « ? » plutôt que par « non » dans la dernière colonne de 5.1 (voir la remarque 3.23 du chapitre 3).

5.1.3 Problèmes de tiges

En collaboration avec F. Bertails, nous avons entrepris d'étendre le modèle de tiges des super-hélices [Ber06, BAC+06] pour y intégrer le frottement de Coulomb modélisé par (1.22). Une super-hélice est une courbe à une dimension de $\mathbb{R}^3$ formée par recollement C^1 d'une succession d'hélices. Chaque hélice est décrite par sa longueur, qui est fixée (ceci modélise donc une tige inextensible), et par ses trois degrés de liberté κ_1, κ_2 (en flexion) et τ (en torsion) comme sur la figure 5.1. Une super-hélice formée de α morceaux dispose donc de 3α degrés de liberté. Une description précise de ce modèle est donnée dans [Ber06], avec sa construction à partir du modèle de tige de Kirchhoff (qui modélise une tige inextensible et sans cisaillement). Ce système possède des carac-

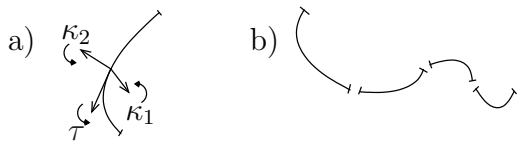


FIG. 5.1 – Une hélice et une super-hélice.

Nom	m_{ss}	n_{ss}	$(nd)/m$	dens(W)	cond(W)	existence
alea-750-100	6	125	0.40	0.04	1.4e+05	oui
alea-1002-200	6	167	0.60	0.03	5.3e+05	oui
alea-1122-300	6	187	0.80	0.02	2.4e+07	oui
alea-1200-400	6	200	1.00	0.02	8.4e+18	?
alea-1248-500	6	208	1.20	0.02	5.7e+19	?
alea-750-100	15	50	0.40	0.09	1.9e+04	oui
alea-1005-200	15	67	0.60	0.07	5.3e+04	oui
alea-1125-300	15	75	0.80	0.06	5.7e+05	oui
alea-1200-400	15	80	1.00	0.05	7.7e+18	?
alea-1245-500	15	83	1.20	0.05	7.6e+18	?
alea-756-100	36	21	0.40	0.18	1.1e+04	oui
alea-1008-200	36	28	0.60	0.14	5e+04	oui
alea-1116-300	36	31	0.81	0.13	6.2e+05	oui
alea-1188-400	36	33	1.01	0.12	5.3e+18	?
alea-1260-500	36	35	1.19	0.11	3.4e+19	?

TAB. 5.2 – Instances aléatoires de grande taille

téristiques « extrêmes » : les objets sont fins et allongés, ce qui impose l'usage de pas de temps assez petits pour empêcher qu'ils se traversent. Chaque tige est parfaitement indéformable dans le sens longitudinal et, au contraire, très souple en flexion et en torsion. Enfin, la matrice de masse de chaque objet élémentaire est pleine, H est souvent de rang faible, et W est mal conditionnée en général (table 5.3).

En revanche, on considère des problèmes dans lesquelles le seul objet extérieur (le support auquel sont attachées les tiges) est en mouvement de solide rigide. On a donc l'assurance qu'une solution existe au problème incrémental, comme on l'a vu à la section 3.4. Le coefficient de frottement utilisé est $\mu = 0.2$.

5.2 Méthodes

5.2.1 Algorithmes

Nous avons implémenté les méthodes suivantes parmi celles passées en revue au chapitre 4 : la méthode d'Alart et Curnier, celle « d'Uzawa » sur le bipotentiel de De Saxcé, la méthode de « Gauss-Seidel » (où les problèmes réduits à un seul contact sont résolus à l'aide de la méthode d'Alart et Curnier) et celle de Haslinger (où les programmes quadratiques internes sont résolus par la méthode du gradient projeté, et le problème externe par itérations de point fixe). Nous testons aussi une méthode moins connue : la méthode de quasi-Newton sur la fonction de Fischer-Burmeister, qui représente les techniques par fonctions de mérite. Enfin notre approche consistant à résoudre l'équation de point fixe (3.24) est implémentée de trois manières : par la méthode de Newton, celle

Nom	m_{ss}	n_{ss}	$(nd)/m$	dens(W)	cond(W)	existence
hair-54-4	18	3	0.22	1.00	3.3e+09	oui
hair-54-5	18	3	0.28	1.00	1.6e+09	·
hair-36-3	18	2	0.25	1.00	5.7e+05	·
hair-54-9	18	3	0.50	1.00	2.4e+17	·
hair-54-9	18	3	0.50	1.00	6e+17	·
hair-90-8	18	5	0.27	0.75	1e+09	·
hair-90-13	18	5	0.43	0.76	2.8e+11	·
hair-90-13	18	5	0.43	0.76	1.3e+11	·
hair-90-12	18	5	0.40	0.62	1.1e+09	·
hair-108-17	18	6	0.47	0.75	3.8e+18	·
hair-108-18	18	6	0.50	0.52	2.6e+19	·
hair-108-16	18	6	0.44	0.55	9.7e+11	·
hair-90-20	18	5	0.67	0.78	1.5e+19	·
hair-108-15	18	6	0.42	0.64	5.2e+17	·
hair-126-22	18	7	0.52	0.48	3.5e+18	oui

TAB. 5.3 – Problèmes de tige de petite taille

Nom	m_{ss}	n_{ss}	$(nd)/m$	dens(W)	cond(W)	existence
hair-468-150	18	26	0.96	0.18	1.5e+20	oui
hair-540-127	18	30	0.71	0.16	7.8e+19	·
hair-522-158	18	29	0.91	0.15	5.1e+20	·
hair-702-106	18	39	0.45	0.11	3.8e+19	·
hair-576-148	18	32	0.77	0.16	1.7e+20	·
hair-738-129	18	41	0.52	0.12	3.5e+19	·
hair-666-148	18	37	0.67	0.12	3.1e+20	·
hair-792-178	18	44	0.67	0.13	1.5e+20	·
hair-792-174	18	44	0.66	0.11	1.3e+20	·
hair-756-184	18	32	0.73	0.11	8.4e+19	·
hair-738-209	18	41	0.85	0.12	7.6e+19	·
hair-810-194	18	45	0.72	0.10	6.5e+20	·
hair-954-229	18	53	0.72	0.10	1.1e+20	·
hair-882-294	18	49	1.00	0.10	1.4e+21	·
hair-972-369	18	54	1.14	0.09	?	oui

TAB. 5.4 – Problèmes de tige de grande taille

Abréviation	Méthode
AC	Méthode de Newton sur la formulation d'Alart et Curnier
DS	Méthode du point fixe sur la formulation de De Saxcé
GS	Méthode de Gauss-Seidel en résolvant les problèmes internes par la méthode d'Alart et Curnier
FB	Méthode de quasi-Newton (BFGS) sur la fonction de Fischer-Burmeister
H-pf	Méthode de point fixe sur la formulation de Haslinger ; QP internes résolus par gradient projeté
NA-Newton	Notre approche, en utilisant la méthode de Newton ; QP internes résolus par gradient projeté
NA-Broyden	Notre approche, en utilisant la méthode de quasi-Newton ; QP internes résolus par gradient projeté
NA-pf	Notre approche, en utilisant la méthode du point fixe ; QP internes résolus par gradient projeté

TAB. 5.5 – Liste des méthodes évaluées

de quasi-Newton (méthode de Broyden) et celle du point fixe. L'ensemble des méthodes évaluées est résumé dans la table 5.5, avec des abréviations servant à les identifier dans les tableaux de résultats à venir.

Enfin, pour tous les algorithmes, on élimine la variable v en calculant explicitement W et q (équation (1.34) du chapitre 1). La sous-section 5.3.7 discute de l'intérêt qu'il peut y avoir, au contraire, à conserver cette variable.

On peut combiner les différentes formulations du problème incrémental et les diverses techniques de résolution de bien d'autres manières ; par exemple, on peut facilement appliquer la méthode de Newton et celle de Broyden à la formulation de Haslinger. Cependant, les performances de ces deux variantes sont déjà rapportées pour notre approche (NA-Newton et NA-Broyden) et nos expériences ont montré que leur comportement était sensiblement le même lorsqu'on les applique à la méthode de Haslinger et à notre approche. On pourrait aussi utiliser la méthode de Newton sur la fonction de De Saxcé au lieu de celle d'Alart et Curnier ; selon nos expériences, on obtient dans les deux cas des résultats très similaires. On se permet donc, pour ne pas alourdir la présentation des résultats, de se limiter à l'étude des huit méthodes décrites par la table 5.5.

5.2.2 Paramètres

Tous les algorithmes nécessitent de régler des paramètres internes. Les valeurs retenues sont indiquées dans la table 5.6. Dans la formulation de De Saxcé, on fixe $\rho = 1$. Ce choix arbitraire explique peut-être les performances décevantes observées pour cette méthode, mais même en essayant empiriquement d'autres valeurs nous n'avons pas réussi à trouver un réglage satisfaisant. On notera d'ailleurs que ρ n'est pas sans dimension. Pour la résolution interne du problème à un seul contact dans la méthode de Gauss-Seidel, on

Méthode	Paramètre	Valeur
AC	Recherche linéaire Goldstein-Price : pas initial	1
	Recherche linéaire Goldstein-Price : nb. iter. max.	25
DS	Paramètre ρ	1
GS	Tolérance sur le critère d'AC interne	10^{-8}
	Nombre d'itérations internes de Newton	15
FB	Recherche linéaire de Wolfe : pas initial	1
	Recherche linéaire de Wolfe : nb. iter. max.	12
H-pf	QP interne (gradient projeté) : nb. iter. max.	250
	QP interne (gradient projeté) : tol. sur $\|x_n - x_{n+1}\|$	10^{-10}
NA-Newton	QP interne (gradient projeté) : nb. iter. max.	250
	QP interne (gradient projeté) : tol. sur $\|x_n - x_{n+1}\|$	10^{-10}
	Recherche linéaire d'Armijo : pas initial	1
	Recherche linéaire d'Armijo : nb. iter. max.	12
NA-Broyden	QP interne (gradient projeté) : nb. iter. max.	250
	QP interne (gradient projeté) : tol. sur $\|x_n - x_{n+1}\|$	10^{-10}
	(pas de recherche linéaire)	
NA-pf	QP interne (gradient projeté) : nb. iter. max.	250
	QP interne (gradient projeté) : tol. sur $\|x_n - x_{n+1}\|$	10^{-10}

TAB. 5.6 – Paramètres des différentes méthodes

fixe la tolérance sur le critère d'Alart et Curnier à 10^{-8} pour le problème interne (elle est à 10^{-6} pour le problème externe). Pour la résolution interne des programmes quadratiques par gradient projeté, on autorise au maximum 250 itérations internes (de gradient projeté) et on arrête prématurément l'algorithme interne si la différence en norme entre deux itérés successifs devient inférieure à 10^{-10}. Ces valeurs ont été choisies par essais et erreurs sur les jeux de problèmes dont nous disposons, et nous ne prétendons pas qu'elles sont judicieuses pour d'autres instances.

5.2.3 Initialisation

Pour les problèmes de tiges, issus d'une simulation, on initialise les inconnues avec leur valeur au pas de temps précédent – si on en dispose. En revanche, si le contact n'était pas actif au pas de temps précédent, on initialise à zéro. Pour les problèmes aléatoires, on initialise toutes les inconnues à zéro.

5.2.4 Critère d'arrêt

Les critères d'arrêt naturels des différents algorithmes ne sont pas les mêmes. Par exemple dans notre méthode qui consiste à résoudre l'équation de point fixe $F(s) = s$, on a envie d'arrêter l'algorithme lorsque $\|F(s) - s\|$ devient petit ; en revanche, dans la méthode d'Alart et Curnier qui consiste à annuler $f_{AC}(u, r)$, on s'intéresse plutôt à la

valeur de $\|f_{AC}(u,r)\|$. Pour uniformiser les différents critères d'arrêt et pouvoir comparer la performance des différents algorithmes en fonction d'une même exigence de précision, on choisit arbitrairement d'utiliser le critère d'Alart et Curnier suivant pour *tous les algorithmes* : on stoppe les itérations lorsque

$$\frac{1}{2nd}\|f_{AC}(u,r)\|^2 < \epsilon_{AC}. \tag{5.1}$$

Noter la présence du nombre de contacts n au dénominateur pour normaliser par rapport à la taille du vecteur $f_{AC}(u,r) \in \mathbb{R}^{nd}$. Il est possible que ce choix favorise la méthode d'Alart et Curnier, qui vise précisément à satisfaire ce critère ; par contraste, dans les autres méthodes la diminution de $\|f_{AC}(u,r)\|$ n'est qu'un sous-produit d'opérations qui visent à satisfaire un autre critère.

Dans toute la suite, la tolérance sur le critère d'Alart et Curnier est fixée à $\epsilon_{AC} = 10^{-6}$. De plus, on limite le temps de calcul à une valeur T_{max} indiquée pour les différentes expériences.

5.3 Expériences

5.3.1 Problèmes aléatoires de petite taille

Considérons d'abord les problèmes de petite taille. On compare tous les algorithmes en temps et en précision (table 5.7). On rappelle que le critère d'arrêt est d'atteindre une précision de $\epsilon_{AC} = 10^{-6}$ dans (5.1), et on fixe la limite de temps à $T_{max} = 30s$. Les temps de calculs doivent être considérés avec une certaine méfiance en raison de l'usage d'un langage interprété comme Scilab. Il vaut mieux considérer les temps de calcul les uns par rapport aux autres plutôt que s'intéresser à leur valeur elle-même.

Les résultats rapportés dans la table 5.7 montrent clairement l'inefficacité de la technique de point fixe sur la formulation de De Saxcé (colonne 3, notée DS). Quelques expériences suffisent à se convaincre que les résultats sont les mêmes en utilisant à la place la fonction d'Alart et Curnier. Cette technique, utilisée avec un certain succès dans la littérature pour résoudre les problèmes réduits à un seul contact au sein d'une boucle de Gauss-Seidel, ne réussit manifestement pas lorsqu'on considère des problèmes un peu plus gros. La technique des fonctions de mérite (colonne 5, notée FB) est elle aussi prise en défaut très rapidement : elle ne parvient à résoudre que les deux premiers problèmes. L'expérience montre qu'en laissant suffisamment de temps à l'algorithme, il parvient très souvent à trouver une solution ; mais avec une extrême lenteur.

La méthode de Gauss-Seidel rend son plus mauvais résultat sur alea-48-8, avec un résidu de l'ordre de l'unité. Elle parvient en général à s'approcher ou à atteindre la tolérance de 10^{-6}. Cependant, elle atteint plus d'une fois sur deux la limite de temps et s'avère donc lente, même sur des problèmes assez petits. Typiquement, elle réduit rapidement le résidu jusqu'à une valeur de 10^{-2} à 10^{-4}, puis il faut attendre beaucoup plus longtemps pour gagner les décimales suivantes. Il faut noter, pour la défense de cette technique, qu'elle est particulièrement pénalisée par le langage utilisé puisqu'elle effectue de très nombreuses itérations (peu coûteuses) alors que la méthode d'Alart et Curnier,

Problème	AC		DS		GS		FB	
	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.
alea-42-5	0.13	5.2e-08	1.79	5.7e+121	0.50	1.0e-08	8.31	8.9e-17
alea-48-8	30.00	1.2	2.48	2.7e+159	30.00	1.1	29.01	1.0e-15
alea-54-11	0.47	4.9e-13	3.53	1.3e+193	1.10	6.7e-07	30.00	42
alea-60-14	0.25	5.8e-09	3.95	1.7e+260	5.02	8.0e-07	30.00	5.3e+07
alea-66-17	0.53	7.7e-11	4.91	4.1e+167	1.73	7.1e-07	30.00	1.6e+21
alea-72-20	0.79	7.6e-09	5.40	5.8e+121	2.39	2.7e-07	30.00	2.0e+07
alea-78-23	1.22	1.4e-13	6.06	1.5e+168	30.00	4.1e-03	30.00	8.1e+23
alea-84-26	30.01	2.8e-02	7.79	Inf	30.00	6.2e-02	30.00	3.5e+08
alea-90-29	4.49	2.7e-10	7.49	6.9e+161	30.00	8.0e-01	30.00	2.7e+13
alea-96-32	0.92	1.6e-09	8.31	1.7e+71	4.19	6.2e-07	30.00	3.7e+26
alea-102-35	1.40	6.1e-10	10.48	2.3e+282	3.03	1.2e-07	30.00	6.9e+34
alea-108-38	30.01	1.8e-01	9.64	Nan	30.00	2.4e-05	30.00	8.8e+16
alea-114-41	1.53	5.3e-08	11.01	Inf	30.00	7.9e-06	30.00	5.8e+11
alea-120-44	4.39	1.8e-13	11.16	1.4e+71	30.00	7.7e-06	30.00	5.5e+12
alea-126-47	3.64	7.2e-07	12.42	4.3e+190	30.00	1.0e-05	30.00	3.8e+12
	H-pf		NA-Newton		NA-Broyden		NA-pf	
alea-42-5	0.32	2.0e-07	1.50	1.6e-12	0.46	2.0e-07	2.34	2.4e-07
alea-48-8	6.23	7.5e-07	30.02	1.2e-03	11.56	1.0e-07	30.00	7.3e-05
alea-54-11	1.33	7.0e-07	27.74	2.5e-07	2.34	1.5e-07	7.53	9.5e-07
alea-60-14	0.70	2.7e-07	4.22	9.5e-09	1.41	7.9e-08	4.9	2.5e-07
alea-66-17	0.25	3.7e-07	1.62	6.0e-11	0.25	2.6e-07	1.83	2.5e-07
alea-72-20	0.49	6.3e-07	13.40	2.4e-8	1.04	4.7e-07	6.21	7.3e-07
alea-78-23	0.94	1.3e-07	4.55	3.0e-10	4.15	5.5e-07	18.51	7.2e-07
alea-84-26	1.19	5.3e-07	4.38	1.3e-09	13.54	5.5e-07	6.31	5.4e-07
alea-90-29	3.12	8.1e-07	25.72	2.8e-09	12.45	4.7e-07	14.15	6.1e-07
alea-96-32	0.76	8.5e-07	3.34	4.8e-10	1.24	3.7e-07	6.88	4.4e-07
alea-102-35	0.67	2.7e-07	2.82	3.0e-07	3.95	2.0e-07	7.82	9.7e-07
alea-108-38	1.68	3.4e-07	8.06	2.7e-07	2.51	5.6e-07	7.69	6.0e-07
alea-114-41	0.75	5.6e-07	4.56	1.1e-11	2.51	4.3e-07	9.32	7.1e-07
alea-120-44	0.70	5.1e-07	5.77	2.7e-11	2.35	2.6e-07	8.63	5.6e-07
alea-126-47	1.34	7.3e-07	6.56	1.6e-09	5.55	6.4e-07	8.23	2.9e-07

TAB. 5.7 – Performances sur les problèmes aléatoires de petite taille

par exemple, ne fait qu'une dizaine d'itérations (très coûteuses) en général. Dans un langage interprété comme Scilab, chaque passage dans la boucle provoque un surcoût qui pénalise donc particulièrement la méthode de Gauss-Seidel. Cependant, il ne faut pas voir dans cette explication la raison unique de sa lenteur : la convergence de cette méthode (quand elle a lieu) est intrinsèquement lente. Par exemple sur alea-114-41, on atteint en 30s un résidu de $7.93 * 10^{-6}$ mais même en patientant dix fois plus longtemps (300s), on ne parvient pas à passer sous la barre de 10^{-6} !

La méthode d'Alart et Curnier s'avère un peu moins robuste que la précédente, car elle échoue sur trois des quinze problèmes ; même en lui laissant plus de temps (quelques minutes), elle ne parvient pas à les résoudre. En revanche, on voit tout l'intérêt qu'il y a à utiliser la méthode de Newton : quand tout se passe bien, la convergence est très rapide : d'une part les temps de calcul sont assez petits par rapport aux autres méthodes, d'autre part on atteint une précision très élevée. Le résidu est de l'ordre de 10^{-13} sur certains problèmes, alors que la précision requise est 10^{-6} : cela signifie que le résidu est passé de plus de 10^{-6} à environ 10^{-13} en une seule itération (ce qui est une belle illustration de la convergence quadratique de l'algorithme de Newton).

Notre approche, en utilisant des itérations de Newton (sixième colonne, notée NA-Newton) est assez robuste : elle n'échoue que sur alea-48-8 (et elle parvient à résoudre ce problème à 10^{-6} près si on lui laisse quelques secondes de plus). Avec des itérations de point fixe, les résultats sont similaires : on n'échoue qu'une seule fois, de nouveau sur le problème alea-48-8, en atteignant cependant une précision raisonnable de $7.31 * 10^{-5}$ (et on peut atteindre la précision requise de 10^{-6} en environ 45 secondes).

Enfin l'approche de Haslinger avec des itérations de point fixe (H-pf) est supérieure à toutes les autres sur ce jeu de problèmes. Elle résout toutes les instances, en un temps inférieur à la seconde sur 9 des 15 problèmes et sans excéder 6.23s dans le pire des cas (sur alea-48-8, qui pose aussi problème à toutes les autres méthodes).

5.3.2 Problèmes de tiges de petite taille

Considérons maintenant les problèmes de tige de petite taille. Ces problèmes ont été sélectionnés au cours d'une simulation dans laquelle deux tiges emportées par leur élan s'enroulent l'une contre l'autre (figure 5.2). On a conservé uniquement les problèmes sur lesquels le solveur utilisé pour la simulation n'est pas parvenu à la tolérance requise dans le temps imparti ; il s'agissait d'une implémentation C++ de l'algorithme d'Alart et Curnier. On ne considère plus les méthodes DS et FB, dont l'inefficacité est démontrée

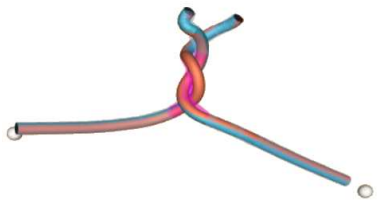


FIG. 5.2 – Deux super-hélices emportées par leur élan

par la table 5.7. Le critère d'arrêt est toujours d'atteindre une précision de $\epsilon_{AC} = 10^{-6}$ dans (5.1), et on fixe la limite de temps à $T_{max} = 10s$. La méthode AC obtient d'assez mauvais résultats, ce qui est normal compte tenu de la manière dont les problèmes ont été sélectionnés : elle produit dans la moitié des cas un résidu de l'ordre de 10^{-2} à 1, avec un pic à 76 pour hair-108-16. La méthode GS se comporte de manière similaire. Sur ces problèmes, la robustesse des approches par point fixe est manifeste. Les approches NA-* atteignent systématiquement la précision requise en mois de 5 secondes. La méthode H-pf dépasse parfois ce temps (6.68 sur hair-90-20, 9.52 sur hair-126-22) mais la précision atteinte est stupéfiante : on obtient régulièrement un résidu entre 10^{-15} et 10^{-13} alors que la tolérance est à 10^{-6} ! Ceci serait moins étonnant si l'on employait une méthode d'ordre 2 ; mais avec des itérations de point fixe, ce résultat est très impressionnant. Le nombre d'itérations externes nécessaires pour y parvenir est typiquement de 4 ou 5 seulement.

5.3.3 Problèmes aléatoires de grande taille

Considérons maintenant les problèmes de grande taille, en commençant par les problèmes aléatoires. On passe la limite de temps à $T_{max} = 100s$, sans changer le critère d'arrêt. La dimension est telle que la méthode de Newton (utilisée par AC et NA-Newton) n'est pas compétitive : l'assemblage et surtout la résolution des systèmes linéaires sont tellement coûteux que chaque itération prend plusieurs minutes. Au contraire, les approches GS et par point fixe (H-pf, NA-*) ne font que des itérations peu coûteuses et il reste possible de faire plusieurs milliers ou dizaines de milliers d'itérations dans un temps raisonnable. On ne considérera que les approches H-pf et NA-*, car la méthode GS atteint systématiquement la limite de temps sur ces gros problèmes. Elle parvient souvent à ramener le résidu à une valeur assez faible, mais reste en dessous des performances de ses concurrentes.

La table 5.9 accorde une certaine supériorité à l'approche NA-pf, qui n'atteint la limite de temps que sur cinq des quinze problèmes-tests, suivie de H-pf qui atteint la limite de temps sur sept d'entre eux. La méthode de Broyden est ralentie par les calculs supplémentaires dûs aux mises à jour de la matrice de quasi-Newton ; comme on n'observe pas la convergence super-linéaire qui motivait son introduction, elle n'est pas compétitive.

5.3.4 Problèmes de tiges de grande taille

Les problèmes de tiges de grande taille ont été sélectionnés de la même manière que ceux de petite taille, au cours de simulations comme celle de la figure 5.3 où l'on attache un certain nombre de tiges à un support fixe auquel on impose un mouvement rapide. Comme pour les problèmes aléatoires de grande taille, la méthode de Broyden ne produit pas le gain de précision qui motive son utilisation. Ses performances sont pratiquement identiques à celles de la méthode NA-pf, qui lui est donc préférable. Quant à la méthode H-pf, elle fournit des résultats d'une précision raisonnable, mais s'avère cette fois plus lente que la méthode NA-pf.

Problème	AC		GS		H-pf	
	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.
hair-54-4	10.01	2.6e-04	10.00	1.8e-03	0.30	4.0e-15
hair-54-5	5.23	1.5e-08	0.70	4.0e-07	0.21	5.8e-08
hair-36-3	10.01	4.4e-03	0.11	2.7e-21	0.24	3.2e-15
hair-54-9	10.05	1.0e+00	10.00	3.7e-02	0.88	4.8e-14
hair-54-9	10.05	1.2e+00	10.00	4.1e-02	0.85	1.7e-14
hair-90-8	0.19	4.0e-08	0.27	5.6e-08	0.89	1.0e-13
hair-90-13	4.06	4.2e-07	0.69	5.9e-07	0.82	1.3e-13
hair-90-13	4.43	1.6e-07	0.32	5.3e-08	0.82	1.0e-13
hair-90-12	0.12	2.2e-07	0.14	8.6e-07	0.48	1.8e-14
hair-108-17	10.01	1.3e+00	10.00	2.3e-01	3.73	1.1e-12
hair-108-18	1.00	2.7e-08	10.00	6.1e-02	2.05	3.7e-12
hair-108-16	10.05	7.6e+01	10.00	8.7e-02	3.07	1.3e-12
hair-90-20	10.01	1.9e-02	10.00	1.1e-02	6.68	8.3e-07
hair-108-15	10.04	1.2e-02	10.00	2.1e+00	5.37	7.4e-13
hair-126-22	10.03	2.6e-02	10.00	6.9e+00	9.52	9.3e-07
Problème	NA-Newton		NA-Broyden		NA-pf	
	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.
hair-54-4	0.36	9.6e-08	0.16	6.1e-08	0.42	6.1e-08
hair-54-5	0.26	1.9e-08	0.16	1.0e-07	0.20	1.0e-07
hair-36-3	0.06	3.3e-07	0.05	7.8e-10	0.07	3.3e-07
hair-54-9	0.76	1.9e-06	0.42	5.2e-09	0.56	5.2e-09
hair-54-9	0.33	1.7e-08	0.39	6.4e-09	0.52	6.4e-09
hair-90-8	0.32	3.6e-07	0.25	2.8e-07	0.38	2.8e-07
hair-90-13	0.23	1.7e-09	0.26	1.5e-07	0.39	1.5e-07
hair-90-13	0.22	7.4e-08	0.26	4.2e-08	0.39	4.2e-08
hair-90-12	0.33	9.2e-10	0.26	2.4e-08	0.39	2.4e-08
hair-108-17	2.41	3.5e-07	1.63	9.3e-07	1.76	6.4e-07
hair-108-18	2.39	6.6e-07	1.10	3.0e-07	1.24	3.0e-07
hair-108-16	1.41	2.7e-07	0.94	6.2e-07	1.08	6.2e-07
hair-90-20	0.37	9.7e-07	0.14	8.4e-07	0.28	8.4e-07
hair-108-15	1.28	8.7e-07	1.35	7.4e-07	1.49	7.4e-07
hair-126-22	4.26	4.9e-07	4.38	7.2e-07	4.25	7.4e-07

TAB. 5.8 – Performances sur les problèmes de tiges de petite taille

Problème	H-pf		NA-Broyden		NA-pf	
	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.
alea-750-100	60.04	3.2e-03	53.08	1.0e-06	6.22	1.0e-06
alea-1002-200	60.08	3.5e-03	19.42	9.9e-07	3.54	9.9e-07
alea-1122-300	10.08	9.7e-07	60.31	2.7e-03	11.73	1.0e-06
alea-1200-400	60.15	9.0e-04	60.52	7.5e-02	36.54	9.9e-07
alea-1248-500	60.57	9.6e-03	60.54	3.2e-01	60.04	6.7e-03
alea-750-100	4.17	1.0e-06	26.73	9.5e-07	3.58	9.6e-07
alea-1005-200	6.43	9.9e-07	60.30	3.3e-04	6.19	1.0e-06
alea-1125-300	14.90	9.8e-07	60.65	2.7e-02	18.49	1.0e-06
alea-1200-400	47.65	9.8e-07	60.11	1.5e-01	60.72	1.2e-06
alea-1245-500	60.69	2.0e-05	60.96	8.8e-02	60.60	3.5e-05
alea-756-100	2.92	9.9e-07	46.64	1.0e-06	4.03	1.0e-06
alea-1008-200	10.14	1.0e-06	60.29	5.7e-04	14.51	9.8e-07
alea-1116-300	41.13	1.0e-06	60.53	1.6e-02	51.61	1.0e-06
alea-1188-400	60.21	5.0e-05	61.79	2.2e-01	60.96	5.7e-04
alea-1260-500	60.25	3.3e-02	62.83	3.4e-01	62.70	9.2e-02

TAB. 5.9 – Performances sur les problèmes aléatoires de grande taille

Problème	H-pf		NA-Broyden		NA-pf	
	T(s)	Préc.	T(s)	Préc.	T(s)	Préc.
hair-108-20	11.21	9.9e-07	4.36	9.9e-07	3.41	8.4e-07
hair-360-63	60.11	1.2e-03	5.61	9.4e-07	4.07	9.1e-07
hair-396-108	60.16	2.1e-05	1.24	1.0e-06	1.65	8.6e-07
hair-270-86	14.69	1.0e-06	1.95	7.9e-07	2.21	7.1e-07
hair-144-18	0.88	5.4e-07	0.47	6.8e-07	0.54	6.6e-07
hair-126-15	14.69	4.8e-07	5.21	1.0e-06	6.46	9.4e-07
hair-270-54	4.55	9.8e-07	1.83	1.0e-06	2.03	1.0e-06
hair-378-90	6.08	8.9e-07	1.10	5.5e-07	1.45	8.2e-07
hair-270-58	25.18	1.0e-06	1.62	8.0e-07	1.77	5.6e-07
hair-90-23	1.31	5.9e-07	0.51	4.4e-07	0.58	4.2e-07
hair-360-43	20.18	9.9e-07	3.36	1.0e-06	1.91	8.0e-07
hair-126-21	2.26	9.6e-07	0.86	2.2e-07	0.92	6.2e-07
hair-360-81	7.23	8.1e-07	2.01	8.7e-07	2.26	7.1e-07
hair-378-94	18.20	9.7e-07	0.65	9.0e-08	0.97	9.0e-07
hair-360-89	60.32	1.1e-04	3.23	1.0e-06	3.54	1.0e-06

TAB. 5.10 – Performances sur les problèmes de tiges de grande taille

FIG. 5.3 – Simulation d'une chevelure

5.3.5 Vitesse de convergence

Les inconvénients de la méthode d'Alart et Curnier sont clairs : elle n'est pas très robuste (il arrive que la recherche linéaire se bloque dans un minimum local ou parce que la direction de Newton n'est pas descendante) et son coût en mémoire et en temps devient prohibitif sur les gros problèmes. Cependant, lorsque leur taille est limitée et que « tout se passe bien » au cours des itérations (la recherche linéaire ne se bloque pas), on observe bien la convergence super-linéaire attendue de la méthode de Newton. La figure 5.4 montre l'évolution du résidu en fonction du temps sur un tel problème ; à la dernière itération, la méthode AC fait décroître le résidu de 10^{-6} à 10^{-14} ! La figure 5.5 montre au contraire un exemple où la recherche linéaire se bloque et la méthode AC n'atteint qu'un résidu d'environ 10^{-4}. Sur ce problème, seule la méthode NA-pf réussit à atteindre la tolérance de 10^{-6} ; la méthode NA-Broyden se comporte quant à elle de manière particulièrement instable. D'autre part, on observe bien une évolution monotone décroissante du résidu avec la méthode AC.

5.3.6 Temps de calcul selon la taille du problème

Lorsqu'on s'intéresse à la simulation de problèmes de taille importante, il est naturel de vouloir estimer la complexité du calcul à effectuer pour résoudre le problème incrémental. Quand $d = 2$, on peut résoudre le problème en temps fini et on voudrait estimer le nombre d'opérations élémentaires nécessaires pour le faire. Quand $d = 3$, on ne dispose pas d'algorithme qui termine, mais on peut se demander combien d'opérations sont nécessaires pour arriver à une tolérance ϵ_{AC} donnée. Donner une réponse théorique à cette question de complexité est malheureusement hors de portée : même dans le cas le plus facile, en 2D, le problème est NP-complet. On se contente donc de quelques considérations empiriques.

La taille du problème dépend essentiellement des deux paramètres m (le nombre de degrés de liberté) et n (le nombre de points de contacts). D'autre part, la difficulté du problème (et ses « chances » d'avoir une solution) dépend de la valeur du rapport $m/(nd)$

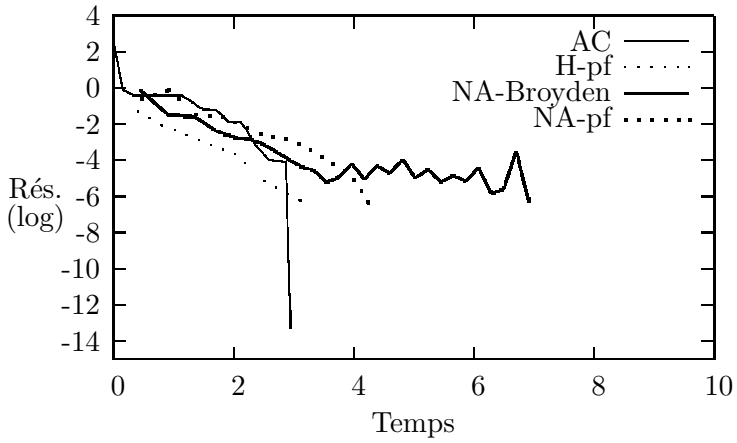


FIG. 5.4 – Évolution du résidu ($m = 2250$, $n = 100$)

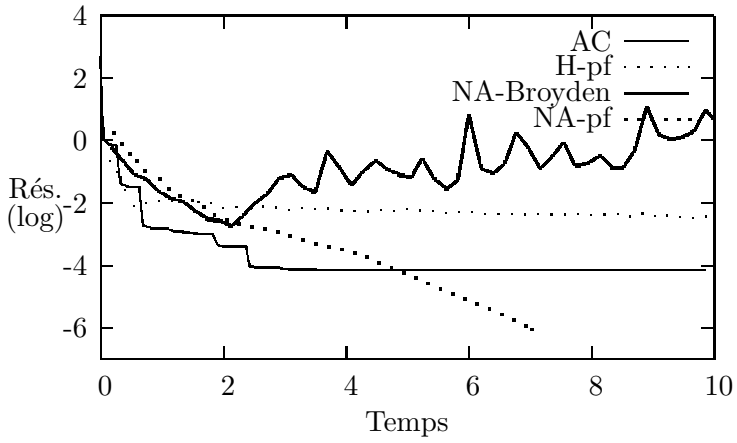


FIG. 5.5 – Évolution du résidu ($m = 2250$, $n = 100$)

(sous-section 3.4.9). Sur les figures 5.6 et 5.7, on fait évoluer les deux paramètres m et n ensemble en maintenant constant le rapport $(nd)/m$ à une valeur de 1/3 (problèmes plutôt faciles) et 1 (problèmes difficiles) respectivement. Sur la figure 5.8 en revanche, $m = 1000$ est fixé et n varie jusqu'à 500 (comme $d = 3$, on arrive alors à $(nd)/m = 1.5$). Pour tous les problèmes, la limite de temps était fixée à $T_{max} = 100s$. Il est manifeste que les problèmes sont plus faciles lorsque $(nd)/m = 1/3$ que lorsque $(nd)/m = 1$: les temps de calculs sont de l'ordre de 10 secondes dans le pire des cas sur la figure 5.6, loin sous la limite de temps, alors qu'ils atteignent régulièrement cette limite sur la figure 5.7. De la même manière sur la figure 5.8, on observe une nette augmentation du temps de calcul (sans atteindre T_{max} cependant) lorsque $(nd)/m$ atteint 1 pour $n \approx 333$.

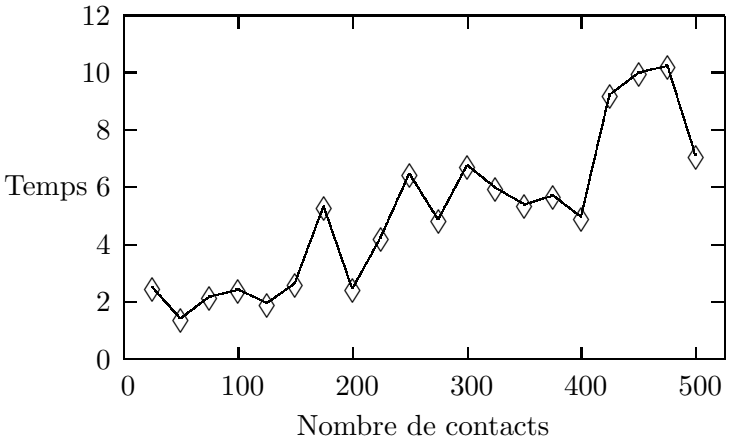


FIG. 5.6 – Temps de calcul (sec.) pour $(nd)/m = 1/3$

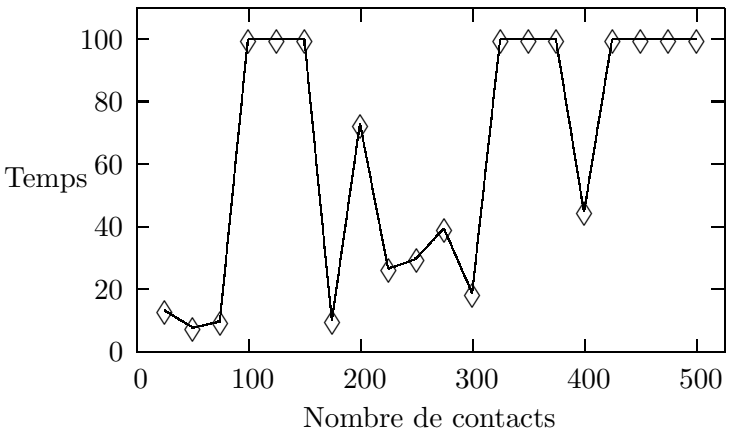


FIG. 5.7 – Temps de calcul (sec.) pour $(nd)/m = 1$

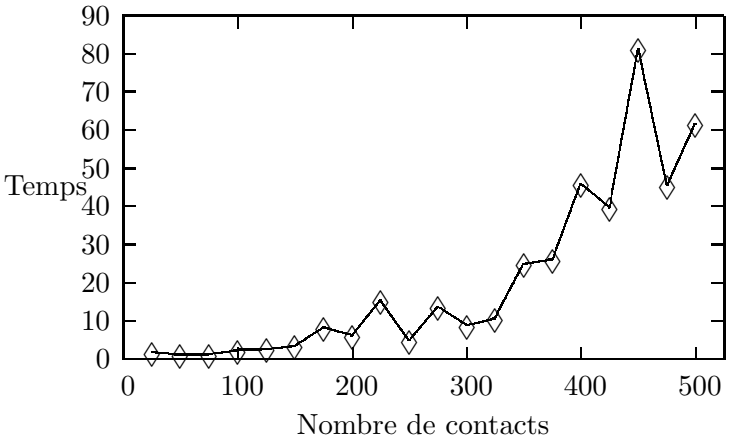


FIG. 5.8 – Temps de calcul (sec.) pour $m = 1000$

5.3.7 Élimination des vitesses généralisées

Dans toutes les expériences ci-dessus, nous avons calculé explicitement W et q dans (1.34). Cependant, assembler la matrice W prend du temps, de l'espace et augmente certainement les erreurs numériques. Pour toutes ces raisons, il peut être préférable de ne pas procéder à ce calcul.

Dans les approches H-pf et NA-*, la matrice W n'intervient que dans le problème interne : il s'agit de minimiser une forme quadratique de matrice W sur un certain ensemble convexe. Lorsque ce problème est résolu par l'algorithme du gradient projeté, W n'intervient que dans l'évaluation du gradient : il faut être capable d'évaluer le produit Wr pour r quelconque.

Au lieu d'assembler W, effectuons seulement la décomposition de Choleski de M ; pour tout vecteur r, on peut alors évaluer le produit Wr en calculant successivement $y := H^\top r$, puis $z := M^{-1}y$ grâce à la décomposition de Choleski, et finalement $Wr = Hz$. On se passe ainsi complètement du calcul et du stockage de W. Sur nos problèmes de grande taille, chaque itération de gradient projeté avec cette technique est entre 2 et 5 fois plus lente que lorsque W est calculée explicitement.

5.3.8 Influence de l'initialisation

Une idée répandue parmi les mécaniciens du contact est que les algorithmes itératifs « choisissent » parmi les multiples solutions éventuelles du problème incrémental celle qui est la plus proche, en un certain sens, de la solution du pas de temps précédent (utilisée pour initialiser l'algorithme). On observe cependant que certaines solutions sont instables : lorsqu'on initialise l'algorithme avec une valeur très proche de cette solution, il converge vers une autre. C'est le cas sur le problème de la barre de la sous-section 1.4.4, lorsqu'il existe deux solutions en force : $r_d = (0, 0)$ (décollement) et r_g (glisse-

ment). La solution r_d est stable pour l'algorithme AC, mais pas la solution r_g : si on initialise l'algorithme avec $r_g + (0, \epsilon)$ ($\epsilon > 0$) on retourne vers r_g, tandis que si on l'initialise avec $r_g + (0, -\epsilon)$ on converge vers la solution éloignée r_d. Ceci peut être observé expérimentalement et montré par le calcul.

Cette observation est assez intuitive : la solution en glissement n'est pas stable, car si on perturbe le système pour rompre l'adhérence due au glissement, alors le système passe en décollement et la force extérieure (dirigée vers le haut) éloigne irrémédiablement la barre du sol, empêchant la reprise du contact.

5.4 Discussion

On tire de ce chapitre les conclusions empiriques suivantes.

- La méthode de point fixe sur la fonction d'Alart et Curnier ou sur celle de De Saxcé ne fonctionne presque jamais sur des problèmes à plusieurs contacts. Étrangement, elle est pourtant parfois suffisante pour être utilisée au sein d'une boucle de Gauss-Seidel sur les problèmes à un seul contact.
- Notre implémentation de la méthode des fonctions de mérite n'est pas efficace du tout, sauf sur de très petits problèmes. Cette technique reste cependant peu étudiée et il est probable qu'on puisse faire mieux.
- Notre approche avec des itérations de Newton (NA-Newton) présente à la fois les inconvénients de la méthode d'Alart et Curnier (coût de l'assemblage et de la résolution du système linéaire sur les gros problèmes, difficultés d'implémentation liées aux matrices creuses) et en partie ceux de la méthode NA-pf (coût élevé de l'évaluation interne). Elle n'est donc pas très séduisante.
- La méthode de Gauss-Seidel est sans doute la plus étudiée et la plus utilisée dans la pratique. Ses avantages sont connus : économie en mémoire, relative robustesse, facilité d'implémentation. Son inconvénient majeur est la lenteur de la convergence, qui oblige souvent à se satisfaire d'une précision moyenne.
- La méthode d'Alart et Curnier globale a été moins souvent étudiée dans la littérature, peut-être parce que son implémentation efficace est non triviale : elle nécessite de faire les calculs d'algèbre linéaire en utilisant des structures de données et des algorithmes adaptés aux matrices creuses. Sur les problèmes de petite et moyenne taille, tant que l'assemblage et la résolution du système de Newton ne sont pas excessivement coûteux, cette méthode est indubitablement la plus rapide : soit elle trouve très rapidement une excellente solution (grâce à la convergence super-linéaire de la méthode de Newton), soit elle se bloque tout aussi rapidement sans trouver de direction de descente. La situation désagréable où l'on ne trouve pas de solution arrive parfois, mais le résidu est souvent assez faible et permet d'envisager de continuer la simulation.
- Notre approche avec des itérations de point fixe (NA-pf) présente les mêmes avantages que la méthode de Gauss-Seidel (économie en mémoire, simplicité d'implémentation, robustesse) tout en étant en général plus rapide ; il faudrait cependant confirmer ce fait en raffinant l'implémentation de ces deux méthodes.

- La méthode de quasi-Newton (méthode de Broyden) ne semble pas plus efficace que celle du point fixe, ce qui n'est pas surprenant vu le caractère non-régulier de la fonction que l'on cherche à annuler.
- L'approche de Haslinger avec des itérations de point fixe est d'une efficacité comparable à la nôtre (NA-pf).

Notre nouvelle approche NA-pf fait donc à peu près aussi bien que la technique existante H-pf, et en constitue une sorte de complément : en effet, les problèmes sur lesquels elle échoue ne sont en général pas les mêmes que ceux sur lesquels H-pf échoue. En disposant de ces deux méthodes, auxquelles on pourrait ajouter GS, on peut donc attaquer les problèmes de grande taille en espérant qu'au moins l'une des trois atteindra la tolérance requise, et en ayant en tout cas de grandes chances d'atteindre un résidu « raisonnable ». Au cours d'une simulation, il est nécessaire de résoudre typiquement plusieurs milliers de fois le problème incrémental, et il suffit d'échouer sérieusement sur l'un d'entre eux pour compromettre toute la suite du calcul. Selon nos expériences, l'usage des méthodes H-pf, NA-pf et GS offre une certaine garantie empirique que ce cas défavorable n'arrivera pas trop souvent.

Conclusion de la première partie

La mécanique du contact est un domaine de recherche très actif (avec les mots-clés « contact mechanics », on obtient 942 000 réponses sur le moteur de recherche scientifique Google Scholar, et « Coulomb friction » livre 64 500 résultats) dans lequel les méthodes numériques ont démontré leur pertinence grâce aux travaux des mécaniciens académiques. Cependant, leur diffusion dans le milieu industriel est encore faible, ce que l'on peut certainement imputer à leur manque de robustesse et à la simplicité ou la taille réduite des instances que l'on sait traiter par rapport à la complexité et la dimension des problèmes couramment rencontrés en ingénierie. Ce défaut de fiabilité est une conséquence directe du choix du modèle de frottement, qui mène à un problème incrémental dont on a observé le caractère NP-complet et dont la résolution efficace est très probablement hors de portée dans le cas général.

Face à cette situation, il est tentant de changer de modèle pour en adopter une variante simplifiée. Cependant, nous avons observé que les méthodes par optimisation paramétrique couplée avec un problème de point fixe (H-pf et NA-pf) parviennent presque toujours réduire le résidu jusqu'à une valeur « faible ». Parfois il décroît linéairement sans problème de convergence apparent ; parfois il se stabilise à une valeur non nulle. Ce dernier cas peut arriver soit parce que l'on s'est approché d'un point fixe mais qu'il n'est pas localement contractant comme sur la figure 5.9, soit parce qu'on est proche d'une solution approximative mais pas nécessairement d'un « vrai » point fixe. Dans

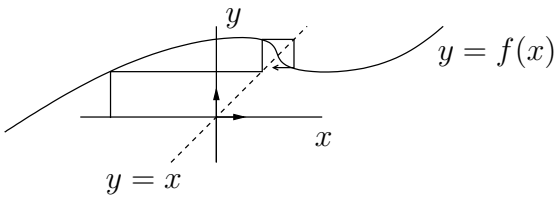


FIG. 5.9 – La fonction f n'est pas localement contractante

tous les cas, il est remarquable que le résidu laissé par les méthodes de point fixe soit toujours assez faible en pratique, même lorsque la convergence linéaire n'a pas lieu et que le cyclage apparaît ; la divergence brutale n'est quant à elle jamais observée, alors que c'est généralement le cas lorsqu'on effectue des itérations de point fixe sur d'autres formulations comme celle d'Alart et Curnier ou celle de De Saxcé. De plus, l'efficacité de ces méthodes pourrait encore être améliorée par un travail sérieux d'implémenta-

tion, par des techniques plus raffinées pour le réglage automatique des paramètres des algorithmes, etc. Puisque le problème incrémental semble plus facile lorsque $nd < m$, on pourrait imaginer des techniques de « splitting » (comme celle de Gauss-Seidel, qui divise le problème en sous-problèmes à un seul contact) qui se ramènent à cette situation en traitant les contacts par groupes. On peut aussi envisager de combiner les approches entre elles, par exemple en utilisant la technique des fonctions de mérite pour « finir le travail » lorsque le résidu laissé par les approches par point fixe converge vers une valeur non nulle... Les possibilités (et, malheureusement, la quantité de travail correspondante) sont presque sans limite.

Si l'on tient absolument à résoudre le problème incrémental de manière très précise, il faut être prêt à en payer le prix : celui d'un problème NP-complet. Mais on peut aussi se satisfaire d'une solution approximative dotée d'un résidu raisonnable, en se disant que le modèle de Coulomb étant lui-même approximatif, les solutions exactes du problème incrémental n'ont pas de raison d'être mécaniquement plus pertinentes que des solutions (suffisamment) approchées. De plus, les solutions approchées vérifient des critères physiques importants :

- elles vérifient un certain principe de dissipation maximum, au sens où elles sont issues – *via* le problème interne – de la minimisation d'une fonction qui a la dimension d'une énergie ;
- on a $u_{\mathrm{N}} \geq 0$ pour H-pf, autrement dit on évite la pénétration ;
- on a $r \in L$ pour NA-pf, et r appartient bien au cône de frottement.

Pour toutes ces raisons, nous considérons que les solutions approchées auxquelles mènent les approches H-pf et NA-pf sont de bons candidats pour la simulation, par essence approximative, des systèmes mécaniques réels. Bien entendu, seule la confrontation avec l'expérience permettra de savoir si ce point de vue est justifié.

La méthode que nous proposons pour ce problème possède des avantages théoriques ; en particulier, elle élimine les contacts sans frottement du problème externe, et élimine aussi l'indétermination due au coincement en faisant disparaître les forces r (non-uniques à s fixé) dans le problème interne. Autrement dit lorsqu'un continuum de solutions en forces existe, il correspond à une solution isolée de l'équation de point fixe (3.24) qui ne porte que sur la norme des vitesses tangentielles. Par comparaison, l'approche de Haslinger ou celle de Gauss-Seidel ne possèdent pas cette qualité car elles itèrent sur les forces. Dans leur ensemble, les différentes méthodes que nous avons implémentées et testées se sont montrées capables de résoudre approximativement des problèmes variés.

Dans le cas où l'on tient absolument à trouver une solution de très bonne qualité du problème incrémental, ou si les techniques usuelles laissent un résidu inacceptable, alors il est tentant – et même nécessaire, au vu du caractère NP-complet – d'utiliser des techniques inspirées par l'optimisation combinatoire. En dimension $d = 2$, le problème est un LCP et il existe des méthodes de recherche arborescente pour le résoudre [JFMR02]. On pourrait envisager de généraliser cette idée au cas non linéaire ($d = 3$). Une telle méthode serait certainement lente ; mais comme les méthodes usuelles fonctionnent souvent, on peut espérer que l'approche énumérative ne serait que rarement utilisée et ne servirait, en quelque sorte, qu'en dernier recours. Cette direction de recherche n'a jamais été ex-

plorée à notre connaissance, et constitue donc un champ d'investigation complètement vierge.

Deuxième partie

Méthodes de coupes

Chapitre 6

Problème de séparation

Ce chapitre et le suivant constituent essentiellement la traduction en français de l'article [Cad08].

Le problème de séparation est de trouver un hyperplan affine, ou « coupe », qui se trouve entre l'origine O et un ensemble convexe donné Q dans un espace euclidien. Typiquement, les coupes servent à raffiner une relaxation donnée d'un problème d'optimisation en ajoutant à ses contraintes une inégalité qui est valide pour l'ensemble réalisable de ce problème, mais violée par la solution relaxée courante. Une coupe est d'autant plus intéressante qu'elle raffine efficacement la relaxation : on s'intéresse donc à celles qui sont profondes, en un sens géométrique précis, et qui exposent une facette. Les cas où la coupe la plus profonde peut être décomposée comme combinaison convexe de coupes qui exposent une facette sont caractérisés en utilisant le polaire inverse. Une description explicite du polaire inverse est donnée dans le cas particulier, courant en pratique, où Q est un polyèdre disjonctif. Cette description est liée au *cut generating linear program* (CGLP, « programme linéaire de génération de coupes ») des techniques dites *lift-and-project*.

Introduction

Le problème de *séparation* est essentiel en optimisation combinatoire, il apparaît par exemple dans le contexte suivant :

- une fonction doit être optimisée sur un ensemble « compliqué » dans $\mathbb{R}^n$, typiquement un énorme ensemble de points entiers
- ce problème étant trop difficile, on résout une relaxation plus simple
- on veut ensuite séparer la solution relaxée de l'ensemble compliqué, afin d'améliorer la relaxation.

Les hyperplans de séparation, appelés coupes dans la communauté de l'optimisation combinatoire, sont utilisés dans de nombreuses méthodes pratiques de résolution de programmes linéaires en nombres entiers, souvent à l'intérieur d'un algorithme de type séparation-évaluation (« branch-and-bound »). On s'intéresse aux coupes générales, qui n'exploitent pas la structure particulière des contraintes à part leur linéarité. Un aperçu

général des techniques existantes pour générer de telles coupes peut être trouvé dans [Cor08]. On s'intéresse aux coupes dites disjonctives [Bal79], dont un exemple important sont les coupes lift-and-project. La méthode présentée ici est similaire à la méthode lift-and-project (avec des variables entières générales, pas nécessairement 0-1, et des disjonctions générales) de plusieurs façons. La différence principale est que, au lieu d'utiliser le CGLP pour générer une seule coupe, on l'utilise pour générer plusieurs coupes qui, ensemble, maximisent la profondeur. Pour certains auteurs, la profondeur est synonyme de violation. Ici, la profondeur est la distance géométrique entre l'hyperplan de coupe et le point à séparer, au sens d'une certaine norme. On s'intéresse principalement à la norme euclidienne, mais il est aussi possible en pratique d'utiliser les normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$.

L'idée de maximiser la profondeur a été utilisée par E.A. Boyd [Boy93, Boy94, Boy95], qui suggère dans ce but une méthode de sous-gradients à pas fixe. D'autre part, Balas et Perregaard [PB01] proposent une procédure pour générer une coupe lift-and-project qui expose une facette en utilisant le CGLP. Ici, on reprend ces deux idées et on propose de générer plusieurs facettes qui, ensemble, maximisent la profondeur au sens où elles impliquent la coupe la plus profonde. L'algorithme proposé dans ce but n'a besoin que d'un oracle qui résout le CGLP pour différentes fonctions objectif, et il n'est pas basé sur une méthode de sous-gradient mais sur un mécanisme de génération de colonnes.

La programmation disjonctive et les coupes lift-and-project ont souvent été étudiées ([Bal79, CL06, RS02]...) du point de vue de l'analyse convexe, en utilisant les notions d'ensembles polaires, de fonction d'appui et de conjugaison. Nous utiliserons aussi ces notions, car elles offrent une vision géométrique de la situation. L'idée essentielle pour trouver des facettes d'un polyèdre convexe Q qui le séparent de l'origine, est d'utiliser une correspondance précise entre les facettes de Q et les points extrêmes d'un autre ensemble convexe dans l'espace dual : le *polaire inverse* de Q, noté Q^-, introduit par Balas [Bal79]. De plus, une facette est profonde lorsque le point extrême correspondant est proche de l'origine dans l'espace dual. Ainsi, notre algorithme pour calculer des facettes profondes construit un ensemble de points extrêmes du polaire inverse dont l'enveloppe convexe contient la projection de l'origine sur Q^-, au sens d'une certaine norme. Comme on s'intéresse à la norme euclidienne, le problème de projection est un programme quadratique (QP). L'étude est ensuite particularisée au cas où Q est un polyèdre disjonctif.

Ce chapitre est organisé de la manière suivante. La section 6.1 introduit l'objet principal, le polaire inverse d'un ensemble convexe. La section 6.2 suggère l'utilisation de coupes profondes et exposant des facettes. Une caractérisation de telles coupes est donnée en utilisant la projection de l'origine sur le polaire inverse Q^- dans l'espace dual, et sa décomposition en points extrêmes de Q^-.

6.1 Polaire inverse et cône normal d'un ensemble convexe

Rappelons d'abord quelques définitions et propriétés. Soit $Q \subsetneq \mathbb{R}^n$ un ensemble convexe fermé non vide, tel que $O \notin Q$. On veut séparer O de Q. Les éléments de $\mathbb{R}^n$

sont identifiés à des vecteurs colonnes, et le produit scalaire euclidien des vecteurs x et y est noté $x \cdot y$. On va utiliser la *fonction d'appui* $\sigma_S : \mathbb{R} \to \mathbb{R} \cup +\infty$ d'un ensemble convexe $S \subset \mathbb{R}^n$:

$$\sigma_S(d) := \sup_{x \in S} d \cdot x. \tag{6.1}$$

Pour un exposé détaillé de cette notion, et d'autres, dans la cadre de l'analyse convexe, on renvoie à [HUL01].

Définition 6.1 (Polaire inverse)**.** Soit $Q \subsetneq \mathbb{R}^n$ un ensemble convexe fermé non vide tel que $O \notin Q$. L'ensemble

$$Q^- := \{d \in \mathbb{R}^n \ ; \ \sigma_Q(d) \leq -1\} = \{d \in \mathbb{R}^n \ ; \ \forall x \in Q \ : \ d \cdot x \leq -1\} \tag{6.2}$$

est appelé le *polaire inverse* de Q.

Cet ensemble, introduit par Balas [Bal79], est très pratique pour générer des coupes : en effet, d sépare O et Q si et seulement si $d \in Q^-$ (à multiplication près par une constante strictement positive), par l'hyperplan d'équation

$$d \cdot x = \sigma_Q(d). \tag{6.3}$$

On remarque que le membre de droite $\sigma_Q(d)$ est la valeur de la violation de l'équation par l'origine, que l'on cherche à séparer.

Définition 6.2 (Cône normal)**.** Soit $Q \subsetneq \mathbb{R}^n$ un ensemble convexe fermé non vide tel que $O \notin Q$. L'ensemble

$$N_Q := \{d \in \mathbb{R}^n \ ; \ \sigma_Q(d) \leq 0\} = \{d \in \mathbb{R}^n \ ; \ \forall x \in Q \ : \ d \cdot x \leq 0\} \tag{6.4}$$

est appelé le *cône normal* de Q.

Il s'agit d'une petite généralisation de la notion de cône normal (habituellement, on demande $O \in Q$ dans cette définition). On remarque que cette définition est très similaire à celle du polaire inverse Q^- : le -1 du membre de droite a seulement été remplacé par un 0. Le cône normal sera utile dans la description de Q^- grâce au lemme suivant :

Lemme 6.3 (Cône de récession)**.** *Soit $Q \subsetneq \mathbb{R}^n$ un ensemble fermé non vide tel que $\notin Q$. Le cône de récession de Q^- est exactement N_Q :*

$$(Q^-)_\infty = N_Q. \tag{6.5}$$

Démonstration. Par définition de Q^- et N_Q on a $Q^- \subset N_Q$, donc le cône de récession Q^- est inclus dans $(N_Q)_\infty = N_Q$. Réciproquement, $\forall d_1 \in Q^-$ et $\forall d_2 \in N_Q$, on a : $\forall t \geq 0$, $\forall x \in Q$, $(d_1 + t d_2) \cdot (x - q) \leq -1$ donc $d_1 + t d_2 \in Q^-$, ce qui implique que $d_2 \in (Q^-)_\infty$. ■

La figure 6.1 illustre les objets ci-dessus. Le cas où Q est (l'enveloppe convexe fermée de) l'union de deux ensembles convexes va apparaître dans la suite, on rappelle donc le résultat suivant [CL06].

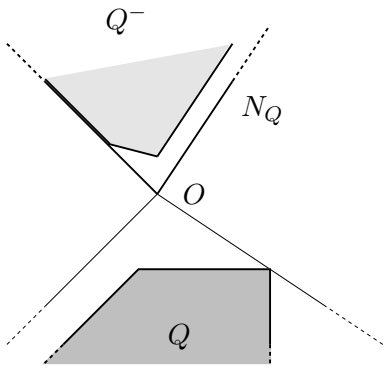


FIG. 6.1 – Cône normal et polaire inverse d'un polyèdre

Lemme 6.4 (Polaire inverse d'une union)**.** *Soit Q_0 et Q_1 deux ensembles convexes fermés non vides de $\mathbb{R}^n$, avec $O \notin Q := \overline{conv}(Q_0 \cup Q_1)$. On a*

$$Q^- = Q_0^- \cap Q_1^- . \tag{6.6}$$

6.2 « Bonnes » coupes

Il y a une infinité de coupes possibles, et cette section est consacrée à la manière de les choisir. Un premier critère naturel est que la coupe touche Q, ce qui fixe le membre de droite comme dans (6.3). Cette section introduit deux nouvelles exigences : être profonde et exposer une facette.

6.2.1 Coupes exposant des facettes

Le cas particulier où Q est un polyèdre mérite une attention particulière. On rappelle que, par définition, une *facette* d'un polyèdre est une face (c'est-à-dire l'intersection de Q et d'un hyperplan d'appui) de dimension maximale, distincte de Q. Quand Q est un polyèdre, ses facettes fournissent en un certain sens la description la plus concise de cet ensemble, comme l'explique le théorème suivant ([Sch86], chapitre 8).

Théorème 6.5 (Représentation minimale d'un polyèdre)**.** *Soit $Q \subset \mathbb{R}^n$ un polyèdre de dimension pleine décrit par des inégalités non-redondantes $Ax \leq b$. Alors $Ax \leq b$ est l'unique représentation minimale de Q, à multiplication près des inégalités par des scalaires strictement positifs. De plus, il existe une correspondance biunivoque entre les facettes de Q et les inégalités dans $Ax \leq b$.*

Par conséquent, quand Q est un polyèdre, on s'intéresse particulièrement aux coupes qui exposent une facette de Q. Ceci motive le résultat suivant (théorème 6.2 de [CL06]).

Théorème 6.6 (Points extrêmes de Q^-)**.** *Soit $Q \subsetneq \mathbb{R}^n$ un polyèdre tel que $lin(Q) = \mathbb{R}^n$ et $O \notin Q$. Alors Q^- est un polyèdre, et la face de Q exposée par $d \in Q^-$ est une facette de Q si et seulement si*

$$d^* := -\frac{d}{\sigma_Q(d)} \tag{6.7}$$

est un point extrême de Q^-.

Remarque 6.7. Le théorème 6.6 n'est pas vrai sans l'hypothèse $lin(Q) = \mathbb{R}^n$. Par exemple sur la figure 6.2, en deux dimensions, l'enveloppe affine $lin(Q)$ de Q est l'axe des x, qui n'est pas de dimension pleine. La direction orthogonale $Q^\perp$est l'axe des y et Q^- est invariant par translation selon l'axe des y, et il n'a donc pas de point extrême. En théorie, cette difficulté peut être contournée en travaillant dans $lin(Q)$ plutôt que dans $\mathbb{R}^n$, mais en pratique on peut très bien ne rien savoir de $lin(Q)$. Ce problème sera examiné dans la section 7.3.

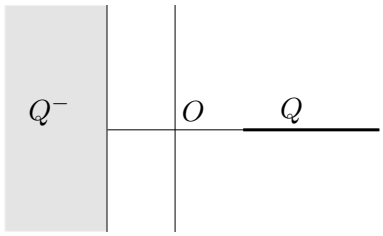


FIG. 6.2 – Quand $lin(Q) \subsetneq \mathbb{R}^n$ ...

6.2.2 Coupes profondes

Une idée pour sélectionner des coupes est de considérer en plus leur *profondeur*, définie comme la distance, pour une norme donnée, entre O est l'hyperplan de coupe. Cette idée est exploitée par Boyd [Boy93, Boy94, Boy95] qui a introduit les coupes de Fenchel et réalisé des expériences numériques en maximisant la profondeur au sens des normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$. Tant les arguments théoriques que les expériences numériques dans [Boy95] incitent à utiliser la norme euclidienne. Nos expériences, pourtant, tendent à montrer que même la profondeur euclidienne est une mauvaise mesure de la qualité d'une coupe, au sens où une coupe profonde au sens de la distance euclidienne n'est pas nécessairement une coupe qui fait remonter largement la valeur de la fonction objectif (voir la section 7.4).

Remarque 6.8 (Forme du polyèdre)**.** L'approche de Boyd, comme la nôtre, utilise seulement la possibilité de maximiser une fonction linéaire sur le polyèdre Q, et aucune autre

information concernant ce polyèdre. La différence principale est que Boyd considère le polyèdre d'un problème de sac-à-dos (« knapsack ») pour Q, tandis que nous considèrons un polyèdre disjonctif. Ceci change la nature de l'oracle qui optimise sur Q : dans [Boy93], chaque appel de l'oracle nécessite de résoudre un problème de sac-à-dos (qui est souvent de petite taille, si le problème est suffisamment creux) tandis que dans notre approche, on doit résoudre à chaque appel de l'oracle un programme linéaire classique.

Le lemme suivant montre que maximiser la profondeur d'une coupe qui sépare l'origine d'un ensemble convexe Q, au sens d'une norme $N(\cdot)$, revient à trouver un point de Q de N-norme minimale. c'est-à-dire à résoudre : $\min_{x\in Q} N(x)$.

Lemme 6.9 (Existence d'une coupe la plus profonde)**.** *Soit $Q \subsetneq \mathbb{R}^n$ un ensemble convexe non-vide tel que $O \notin Q$, $N(\cdot)$ une norme quelconque, et q^* la projection de O sur Q pour $N(\cdot)$. Alors :*

- *il existe une coupe qui sépare O de Q de N-profondeur maximale,*
- *sa profondeur est $N(q^*)$,*
- *la direction de coupe est (à multiplication près par une constance strictement positive) un sous-gradient de $N(\cdot)$ en q^*.*

Démonstration. Par définition, une coupe de profondeur au moins δ est un hyperplan affine qui sépare Q et la boule $B_N(0,\delta)$. Pour tout $\delta > N(q^*)$, $Q \cap int(B_N(0,\delta)) \neq \emptyset$ donc les deux ensembles ne peuvent pas être séparés. Pour $\delta = N(q^*)$, on a $Q \cap int(B_N(0,\delta)) = \emptyset$ donc les deux ensembles peuvent être séparés. Tout hyperplan de séparation est une coupe de profondeur maximale. Comme la direction de coupe est normale à l'ensemble de niveau de $N(\cdot)$ en q^* (qui est exactement $B_N(0, N(q^*))$), c'est un sous-gradient de $N(\cdot)$ en q^* à multiplication près par une constante strictement positive ([HUL01, thm. D.1.3.5]). ∎

Remarque 6.10 (Unicité dans le cas euclidien)**.** Le lemme 6.9 affirme qu'une coupe de profondeur maximale existe, mais elle n'est pas unique en général. Cependant, dans le cas particulier de la distance euclidienne ($N(\cdot) = \|\cdot\|_2$), on observe que la projection est unique puisque $\|\cdot\|_2^2$ est strictement convexe, et le sous-gradient est également unique puisque $\|\cdot\|_2^2$ est une fonction différentiable. Ainsi, la coupe euclidienne la plus profonde est unique.

Un premier résultat de dualité ([Boy95] prop 2.3) affirme que projeter O sur Q est équivalent à maximiser la violation de la direction de coupe sous contrainte de norme (duale) $N^*(\cdot)$.

Théorème 6.11 (Fenchel)**.** *Soit $N(\cdot)$ une norme quelconque sur $\mathbb{R}^n$, $N^*(\cdot)$ sa norme duale, et Q un ensemble convexe fermé. On considère*

$$(P)\ \min_{x\in Q} N(x) \quad \text{et} \quad (D)\ \min_{N^*(d)\leq 1} \sigma_Q(d). \tag{6.8}$$

Alors (P) et (D) sont duaux :

– *si x est réalisable pour (P) et d est réalisable pour (D), alors $N(x) \geq \sigma_Q(d)$,*
– *si x est optimal pour (P) et d est optimal pour (D), alors $N(x) = \sigma_Q(d)$.*

Le théorème de Fenchel 6.11 affirme que minimiser $N(\cdot)$ sur Q est équivalent à maximiser la violation de l'origine $\sigma_Q(d)$ sous une contrainte de normalisation sur $N^*(d)$. Par exemple, la technique de Boyd [Boy93] ainsi que la méthode lift-and-project [BCC93] consistent à maximiser la *violation* de la contrainte générée, sous une contrainte de normalisation $N^*(d) \leq 1$ où $N^*(\cdot)$ est une norme (ou plus généralement, une *semi-norme* [CS97]).

Remarque 6.12 (Normalisation linéaire)**.** Quand la normalisation est polyédrale, le problème d'optimisation résultant est linéaire : c'est le « cut generating linear program » (CGLP) [Cor08]. Différentes contraintes de normalisation ont été testées, avec un succès variable. Dans [CS97], la relation entre la normalisation choisie dans le programme linéaire dual et la relaxation correspondante du programme linéaire primal est expliquée en détails pour différentes normalisations.

6.2.3 Lien avec la projection sur Q^-

Le lemme suivant [CL06] affirme que, par homogénéité positive, le problème de maximisation de la violation sous contrainte de normalisation $(\min_{N^*(d)\leq 1} \sigma_Q(d))$ est à son tour équivalent à la minimisation de la norme duale $N^*(d)$ de la direction de coupe sous contrainte que la violation soit suffisante, de sorte que le polaire inverse réapparaît naturellement.

Lemme 6.13. *Soit $Q \subsetneq \mathbb{R}^n$ un ensemble convexe non-vide tel que $O \notin Q$. On considère*

$$\lambda_1 := \min_{N^*(d)\leq 1} \sigma_Q(d) \quad \text{et} \quad \lambda_2 := \min_{\sigma_Q(d)\leq -1} N^*(d). \tag{6.9}$$

Alors $\lambda_1\lambda_2 = -1$, et les deux problèmes ont le même argmin (à multiplication près par une constante strictement positive).

Démonstration. Puisque $O \notin Q$, $\lambda_1 < 0$ et $\lambda_2 > 0$. Soit d_2 dans l'argmin de P_2 : $N(d_2) = \lambda_2$ et $\sigma_Q(d_2) = -1$ (la contrainte est nécessairement active, sinon d_2 pourrait être amélioré par homothétie). Alors $N(d_2\lambda_2^{-1}) = 1$, donc $d_2\lambda_2^{-1}$ est réalisable pour (P_1), et $\sigma_Q(d_2\lambda_2^{-1}) = -\lambda_2^{-1}$, ce qui prouve que $\lambda_1 \leq -\lambda_2^{-1}$. Le même argument montre que $\lambda_2 \leq -\lambda_1^{-1}$. Finalement, $-1 \leq \lambda_1\lambda_2 \leq -1$, et d_2 est une solution de (P_2) si et seulement si $d_2\lambda_2^{-1}$ est une solution de (P_1). ∎

Bien qu'il ait été plus courant de maximiser la violation $(\min_{N^*(d)\leq 1} \sigma_Q(d))$ dans les études précédentes, nous avons choisi l'autre option $(\min_{\sigma_Q(d)\leq -1} N^*(d))$. En effet, dans notre application à la norme euclidienne, il semble plus naturel de minimiser une fonction quadratique (la norme euclidienne) sur un ensemble polyédral, plutôt que de minimiser une fonction polyédrale (la fonction d'appui $\sigma_Q(\cdot)$) sur un ensemble défini par une contrainte quadratique.

6.2.4 Distance euclidienne

Le cas de la distance euclidienne est notre application principale et jouit de propriétés particulières. La remarque 6.10 montre qu'il existe une unique coupe la plus profonde pour la distance euclidienne, et le lemme 6.9 révèle l'importance de la projection sur un ensemble convexe fermé. On rappelle la caractérisation suivante d'une projection, qui sera utilisée plus loin dans la section 7.1.

Lemme 6.14 (Caractérisation d'une projection)**.** *Soit C un ensemble convexe fermé et $c^* \in C$. Alors c^* est la projection euclidienne de l'origine sur C si et seulement si $c^* \cdot c \geqslant \|c^*\|^2$ pour tout $c \in C$.*

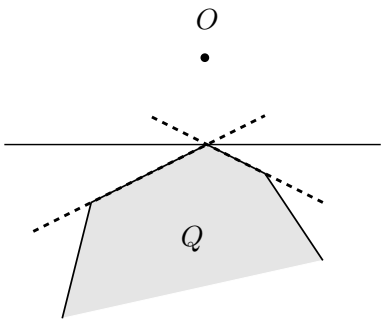


FIG. 6.3 – Coupe la plus profonde et facettes de Q

Comme la norme euclidienne est auto-duale, le théorème de Fenchel 6.11 et le lemme 6.13 se résument au lemme suivant, pour lequel on propose une démonstration simple.

Lemme 6.15 (Dualité, cas euclidien)**.** *Soit q^* la projection de O sur Q et d^* sa projection sur Q^-. On a*

$$d^* = -\frac{q^*}{|q^*|^2}. \tag{6.10}$$

Démonstration. Par définition, pour tout $d \in Q^-$,

$$-1 \geq \sigma_Q(d) \geq d \cdot q^* \geq -\|d\|\,\|q^*\| \;\Rightarrow \|d\| \geq \frac{1}{\|q^*\|}. \tag{6.11}$$

Posons $d^* := -\frac{q^*}{|q^*|^2}$: pour tout $x \in Q$,

$$d^* \cdot x = -\frac{q^* \cdot (x - q^* + q^*)}{\|q^*\|^2} = -\frac{q^* \cdot (x - q^*)}{\|q^*\|^2} - 1;$$

mais $q^* \cdot (x - q^*) \geq 0$ (lemme 6.14). On a donc prouvé que $d^* \cdot x \leq -1$ pour tout $x \in Q$, donc $d^* \in Q^-$. Comme $|d^*| = \frac{1}{|q^*|}$, (6.11) prouve que d^* est de norme minimale dans Q^-. ■

6.2.5 Coupes profondes exposant des facettes

Les deux propriétés introduites ci-dessus pour une coupe, être profonde et exposer une facette, sont antagonistes. Par exemple, la figure 6.3 montre que la coupe la plus profonde (en trait plein) n'expose pas de facette ; et les deux coupes exposant des facettes (en trait pointillé) ne sont quant à elles pas très profondes. Néanmoins, à elles deux, elles impliquent la coupe la plus profonde.

On suggère donc de rechercher des coupes qui, d'une part exposent des facettes, et d'autre part impliquent la coupe la plus profonde. Comme il est équivalent de projeter O sur Q pour $N(\cdot)$ et de projeter O sur Q^- pour $N^*(\cdot)$, calculer des facettes impliquant la coupe la plus profonde au sens de N revient à calculer un ensemble de points extrêmes de Q^- dont l'enveloppe convexe contient la N^*-projection de l'origine d^*. Ceci est l'objet du chapitre suivant.

Chapitre 7

Méthode par décomposition en facettes

Le chapitre précédent a permis d'introduire le problème de séparation ainsi que les objets mathématiques nécessaires à son étude. Dans ce chapitre, on propose une nouvelle méthode de génération de coupes qui s'appuie sur l'idée de décomposition en facettes.

La section 7.1 est consacrée au problème du calcul la projection d'un point sur un ensemble convexe fermé, ainsi que la décomposition en point extrêmes. Un algorithme de projections successives sur le polaire inverse est proposé, qui permet de calculer la décomposition en facettes de la coupe la plus profonde. La section 7.2 donne une caractérisation de Q^- quand Q est un polyèdre décrit explicitement ; la projection et la décomposition peuvent être effectivement calculés dans ce cas. La section 7.2.2 introduit la programmation disjonctive et la section 7.3 généralise la méthode de la section 7.2 au cas où Q est un polyèdre disjonctif. Finalement, on propose un algorithme pour calculer des coupes disjonctives profondes et qui exposent des facettes pour la programmation entière. Des expériences numériques simples sont données dans la section 7.4. Elles montrent comment ces coupes se comportent en pratique par rapport à la coupe la plus profonde et la coupe la plus violée.

7.1 Algorithme de projection

Le chapitre précédent a montré l'importance de la minimisation de la fonction quadratique $\|\cdot\|_2^2$ (le carré de la distance euclidienne) pour calculer des coupes profondes au sens de cette distance. Elle a aussi montré l'importance de la décomposition de la projection en points extrêmes. Dans cette section, on présente un algorithme pour résoudre ce problème (projeter et décomposer) où les points extrêmes sont générés par une suite de programmes linéaires. De plus, la section 7.3 présentera une application dans laquelle ces programmes linéaires peuvent être effectivement résolus.

7.1.1 Algorithme

Soit C un ensemble convexe fermé borné. L'algorithme suivant calcule la projection de l'origine sur C par un mécanisme de génération de colonnes, ou algorithme de projections successives, illustré par la figure 7.1. On remarque que, C étant borné, il possède un point extrême, donc l'initialisation a un sens.

Algorithme 7.1 (Projection par génération de colonnes)**.** *On choisit un point extrême d_1 de C ; on pose $k = 1$.*

ÉTAPE 1 *(*Problème maître*). On calcule la projection d_k^* de l'origine sur l'enveloppe convexe de $d_1, \ldots, d_k$.*

ÉTAPE 2 *(*Appel à l'oracle*). On calcule un point extrême $d_{k+1} \in C$ en minimisant sur C la fonction linéaire $d \mapsto d_k^* \cdot d$.*

ÉTAPE 3 *(*Test d'arrêt*). Si $d_k^* \cdot d_{k+1} < \|d_k^*\|^2$, on incrémente k et on retourne à l'étape 1. Sinon on s'arrête.*

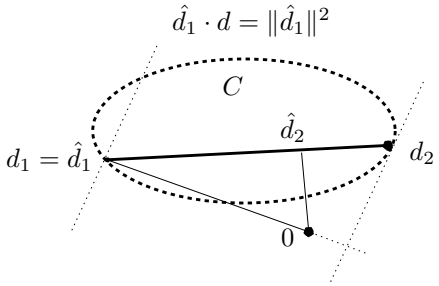


FIG. 7.1 – Les deux premières itérations de l'algorithme 7.1

L'étape 1 consiste à résoudre pour $\alpha \in \mathbb{R}^k$

$$\min \frac{1}{2}\Big\|\sum_{i=1}^{k} \alpha_i d_i\Big\|^2, \quad \sum_{i=1}^{k} \alpha_i = 1\,, \quad \alpha_i \geqslant 0,\ i = 1, \ldots, k\,. \tag{7.1}$$

N'importe quel algorithme de programmation quadratique peut être utilisé pour ce problème ; cependant, Ph. Wolfe en décrit un dans [Wol76], adapté à la forme (7.1). Le résultat d'un tel algorithme est un ensemble de α_i ; les d_i correspondants sont des points extrêmes de C (par construction à l'étape 2).

L'algorithme 7.1 est utilisable en pratique lorsqu'on dispose d'un oracle pour minimiser les fonctions linéaires sur C. Cet oracle doit retourner des points extrêmes.

Remarque 7.2 (Variante sans QP)**.** Dans sa formulation originale [Wol76], l'algorithme de Wolfe utilise explicitement la liste des points (ici les d_i) sur lesquels on projette. Cette liste est parcourue lorsque l'algorithme a besoin de savoir quel est le plus petit des $X \cdot d_1, \ldots, X \cdot d_k$ (pour un certain X). Cette étape peut donc être remplacée, dans notre algorithme, par un appel à l'oracle pour minimiser $X \cdot d$ sur C. Autrement, dit il n'est pas nécessaire d'utiliser un solveur de QP : on peut implémenter directement la

variante ci-dessus de l'algorithme de Wolfe, qui n'utilise qu'un solveur LP. C'est ce que nous avons fait dans nos expériences numériques.

Si l'algorithme s'arrête à l'itération K, alors les étapes 2 et 3 impliquent que $d_K^* \in C$ vérifie $d_K^* \cdot d \geqslant \|d_K^*\|^2$ pour tout $d \in C$; grâce au théorème 6.14, ceci signifie que d_K^* est la projection de l'origine sur C. Supposons maintenant que l'algorithme 7.1 soit appliqué à $C = Q^-$ et s'arrête à l'itération K. Dans le langage du chapitre 6, d_K^* est la coupe la plus profonde (théorème 6.15). De plus, un certain nombre de coupes d_i avec $\alpha_i > 0$ seraient produites, et elles exposeraient des facettes (théorème 6.6) ; comme $d_K^* = \sum_i \alpha_i d_i$, elles impliqueraient la coupe la plus profonde. En un mot, 7.1 atteindrait les objectifs de la section 6.2. Cependant, l'identification $C = Q^-$ viole le caractère intrinsèquement non borné de Q^- ; ce problème sera l'objet de la sous-section 7.1.3.

7.1.2 Convergence

Les d_{k+1} produits par l'étape 2 de l'algorithme 7.1 ne peuvent pas être l'un des d_i précédents ; l'algorithme termine donc si C a un nombre fini de points extrêmes, ce qui est le cas dans notre application à la programmation entière. Cependant, on donne la démonstration de convergence suivante, plus intrinsèque, et qui montre le rôle de l'hypothèse de compacité de C.

Lemme 7.3. *En posant* $d(t) := d_k^* + t(d_{k+1} - d_k^*)$, *on a pour tout* $t \geqslant 0$

$$\|d(t)\|^2 \leqslant \|d_k^*\|^2 - 2t\|d^* - d_k^*\|^2 + t^2\|d_{k+1} - d_k^*\|^2 .$$

Démonstration. Dans le développement

$$\|d(t)\|^2 = \|d_k^*\|^2 + 2td_k^* \cdot (d_{k+1} - d_k^*) + t^2\|d_{k+1} - d_k^*\|^2 ,$$

on sort le coefficient de $t \geqslant 0$: d'après la définition de d_{k+1} à l'étape 2 ($d^* \in C$!) :

$$\begin{array}{rcl} d_k^* \cdot (d_{k+1} - d_k^*) & \leqslant & d_k^* \cdot (d^* - d_k^*) \\ & = & d^* \cdot (d^* - d_k^*) - \|d_k^* - d^*\|^2 \\ & \leqslant & -\|d_k^* - d^*\|^2 , \end{array}$$

où la dernière inégalité vient du théorème 6.14 ($d_k^* \in C$!). ∎

Théorème 7.4 (Convergence)**.** *La suite* $\{d_k^*\}$ *converge vers* d^*.

Démonstration. Comme d_{k+1} et d_k^* appartiennent au borné C, $\|d_{k+1} - d_k^*\|^2 \leqslant M$ pour un certain M. Soit $\delta > 0$ et soit K_δ l'ensemble de k tels que $\|d^* - d_k^*\|^2 > \delta$. Quitte à diminuer δ, on peut supposer $\delta/M \leqslant 1$; alors $d(\delta/M)$ appartient au segment $[d_k^*, d_{k+1}]$ et, d'après l'étape 1 :

$$\|d_{k+1}^*\|^2 \leqslant \|d(t)\|^2 \leqslant \|d_k^*\|^2 - 2t\delta + t^2 M$$

pour tout $t \in [0, \delta/M]$. En particulier, pour $t = \delta/M$:

$$\|d^*_{k+1}\|^2 \leqslant \|d^*_k\|^2 - \frac{\delta^2}{M}\,, \quad \text{pour tout } k \in K_\delta\ .$$

La suite $\{\|d^*_k\|^2\}$ est décroissante, donc l'inégalité ci-dessus ne peut avoir lieu que pour un nombre fini de k : K_δ est un ensemble fini.

En d'autres termes, pour δ arbitrairement petit, $\|d^* - d^*_k\|^2 \leqslant \delta$ si k est suffisamment grand ; ceci signifie que $\|d^*_k - d^*\|^2 \to 0$. ■

7.1.3 L'hypothèse de compacité

Lorsque C n'est pas borné, la démonstration du théorème 7.4 suggère que l'algorithme 7.1 ne converge pas en général. De plus, si C n'est pas borné, les fonctions linéaires peuvent tendre vers $-\infty$ sur C : l'étape 2 peut échouer à produire un d_{k+1}.

De fait, la projection d^* n'est pas nécessairement une combinaison convexe de points extrêmes : des rayons extrêmes peuvent être nécessaires pour la décomposer. Dans notre contexte, cela correspond aux cas où la coupe la plus profonde ne peut pas être décomposée en facettes, comme sur la figure 7.2. Les objets que l'algorithme 7.1 est censé calculer peuvent ne pas exister.

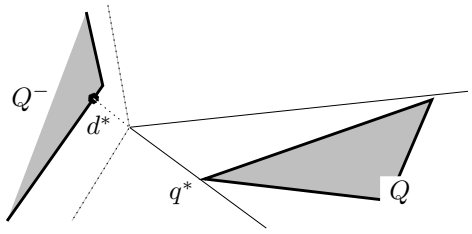


FIG. 7.2 – La coupe la plus profonde n'est pas décomposable en coupes de facettes

Il semble possible de contourner ces problèmes en utilisant le caractère polyédral de C : on peut alors écrire

$$C = \sum_{i=1}^{p} \alpha_i d_i + \sum_{j=1}^{r} \lambda_j r_j\,, \tag{7.2}$$

où α varie dans le simplexe unité et $\lambda \geqslant 0$. La projection d^* peut ainsi être décrite comme une combinaison de points extrêmes d_i et de rayons extrêmes r_j. Avec des modifications appropriées, l'algorithme 7.1 pourrait encore s'appliquer dans le cas non borné mais polyédral, avec un oracle qui renverrait soit un point extrême, soit un rayon extrême comme résultat de la minimisation d'une fonction linéaire.

Cependant, les rayons extrêmes de $C = Q^-$ ne jouent pas un rôle aussi crucial que les points extrêmes, qui correspondent à des facettes. De plus, des difficultés numériques existent, comme l'illustre la figure 7.3 :

– A la première itération, l'oracle renvoie le rayon horizontal r_2.
– Ensuite d_2^* est la projection de 0 sur $d_1 + r_2$.
– Numériquement, d_2^* n'est pas exactement orthogonal à r_2 et l'oracle appelé en d_2^* peut très bien renvoyer r_2 de nouveau (au lieu du point extrême d_2).
– A partir de là, l'algorithme 7.1 boucle et ne termine pas.

Pour contourner cette difficulté, une description explicite des éléments de C dans (7.2) semble nécessaire.

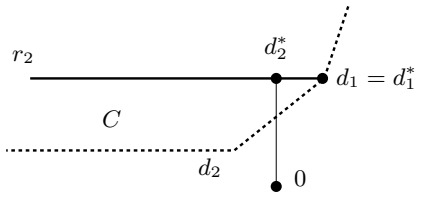


FIG. 7.3 – Instabilité numérique dans le cas non borné

Les raisons ci-dessus expliquent pourquoi, dans notre application où $C = Q^-$ est toujours non borné, nous allons borner artificiellement C en ajoutant une contrainte de normalisation. Remarquer que cette contrainte de normalisation est très différente de celle utilisée dans les méthodes lift-and-project. Dans notre cas, la borne est seulement un paramètre de sécurité pour assurer la convergence de l'algorithme, dont le choix ne doit pas « trop » (voir la remarque 7.22 ci-dessous) changer le résultat de l'algorithme.

7.1.4 Utilisation d'un solveur QP

Un inconvénient de cette approche, du point de vue pratique, est le grand nombre de programmes linéaires qu'il peut être nécessaire de résoudre au cours de l'algorithme de Wolfe. En supposant que C est polyédral et qu'une description appropriée est disponible (comme c'est le cas dans notre application à la programmation disjonctive, voir la section 7.2.2), une solution possible pour accélérer les calculs est de calculer directement d^* à l'aide d'un solveur QP. Ensuite, en initialisant l'algorithme de Wolfe avec la solution d^* (éventuellement en ajoutant la contrainte $d \cdot d^* \leq \|d^*\|^2$), seul un petit nombre d'appels à l'oracle LP serait nécessaire puisque la projection serait déjà calculée et que seule la décomposition en points extrêmes resterait à déterminer.

7.2 Cas d'un polyèdre décrit par des inégalités

On suppose, dans cette section, que Q est un polyèdre explicitement décrit par des inégalités. Nous allons caractériser le polaire inverse Q^- comme l'enveloppe de ses points extrêmes et de ses rayons extrêmes.

7.2.1 Caractérisation du cône normal

Le lemme suivant ([CL06], théorème 6.7), qui est une formulation du lemme de Farkas, caractérise le cône normal de Q.

Lemme 7.5 (Farkas). *Soit $Q = \{x \in \mathbb{R}^n \ ; \ Ax \leq b\}$ où $m \in \mathbb{N}$, $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$. On suppose que $Q \neq \emptyset$ et $O \notin Q$. Alors le cône normal du polyèdre Q est*

$$N_Q = \{d = A^\top u \ ; \ u \geq 0 \ ; \ b \cdot u \leq 0\}. \tag{7.3}$$

Autrement dit, N_Q est l'image par $A^\top$ du cône

$$K = \{u \geq 0 \ ; \ b \cdot u \leq 0\}. \tag{7.4}$$

Remarque 7.6. Le lemme 7.5 n'est pas vrai sans l'hypothèse selon laquelle Q est non vide. En pratique, cependant, on peut ignorer si Q est vide ou non. Ce problème est considéré dans la section 7.3 et un contre-exemple au lemme 7.5 sera donné lorsque $Q = \emptyset$.

Soit

$$\begin{aligned} I_- &= \{i \in 1...m \ : \ b_i < 0\}, \\ I_0 &= \{i \in 1...m \ : \ b_i = 0\}, \\ I_+ &= \{i \in 1...m \ : \ b_i > 0\}. \end{aligned} \tag{7.5}$$

Le résultat suivant donne les rayons extrêmes de K dans (7.4) et des générateurs de N_Q.

Lemme 7.7 (Générateurs du cône K). *Soient I_-, I_0, I_+ comme dans (7.5).*

1. *Soit $(e_1, \ldots, e_m)$ la base canonique de $\mathbb{R}^m$. Les rayons extrêmes du cône K sont exactement (à multiplication près par une constante positive) les éléments de*

$$E = \{e_i \ ; \ i \in I_- \cup I_0\} \cup \{b_j e_i - b_i e_j \ ; \ (i,j) \in I_- \times I_+\}. \tag{7.6}$$

2. *L'image par $A^\top$ de E génère le cône normal N_Q :*

$$\operatorname{cone} A^\top E = N_Q. \tag{7.7}$$

Démonstration. L'équation (7.4) montre que les rayons extrêmes de $K \subset \mathbb{R}^m$ sont obtenus de la manière suivante :

– on extrait de $\{e_1, \ldots, e_m, b\}$ tous les sous-ensembles de $m-1$ vecteurs linéairement indépendants.

– on résout le système linéaire correspondant, produisant une droite de $\mathbb{R}^m$. Cette droite ne peut pas être contenue dans K car $K \subset \mathbb{R}^m_+$.

– Si cette droite ne contient que 0 comme point réalisable dans (7.4), le point extrême $0 \in K$ est produit, mais pas de rayon extrême.

– Sinon, la demi-droite faisable est un rayon extrême de K.

Pour extraire un $(m-1)$-uplet, c'est-à-dire éliminer un couple parmi les $\{e_1, \ldots, e_m, b\}$, on a deux possibilités.

– Soit on élimine e_i ($i \in [1, m]$) et b, alors les $m-1$ vecteurs restants e_k avec $k \neq i$ sont automatiquement linéairement indépendants, et l'intersection de K avec la demi-droite $\mathbb{R}_+ e_i$ n'est pas le singleton $\{0\}$ si et seulement si $e_i \cdot b \leq 0$.
– Soit on élimine e_i et e_j ($i \neq j$). Dans ce cas, la famille est linéairement indépendante si et seulement si $(e_i \cdot b, e_j \cdot b) \neq (0, 0)$ (on peut même supposer que $e_i \cdot b \neq 0$ et $e_j \cdot b \neq 0$ sinon on est ramené au cas précédent où la droite est dirigée par l'un des vecteurs de base), et la droite déterminée par ces $(m-1)$ vecteurs est

$$\{u = \alpha e_i + \beta e_j \ ; \ u \cdot b = 0\}$$

Un vecteur directeur de cette droite est $(b \cdot e_j)e_i - (b \cdot e_i)e_j$. L'intersection de K avec la droite est non triviale si et seulement si $b \cdot e_i$ et $b \cdot e_j$ ont des signes opposés (on a supposé qu'ils étaient tous deux non nuls).

L'équation (7.7) vient de $N_Q = A^\top K = A^\top \operatorname{cone} E = \operatorname{cone} A^\top E$. ■

Remarque 7.8 (Problèmes creux)**.** Comme les vecteurs de E ne contiennent qu'un ou deux coefficients non nuls, les générateurs $A^\top E$ de N_Q calculés de cette manière sont des lignes de A ou des combinaisons linéaires de deux lignes de A (transposées). Si A est creuse, les générateurs le seront aussi.

Remarque 7.9 (Nombre de générateurs)**.** On considère le produit cartésien de I_- et I_+ et E possède *a priori* $O(m^2)$ éléments. Plus précisément, le nombre de générateurs calculés est

$$|E| = |I_-| + |I_0| + (|I_-|)(|I_+|). \tag{7.8}$$

Cette technique nous fournit un ensemble de vecteurs qui contient les rayons extrêmes de N_Q, mais aussi éventuellement des vecteurs inutiles qui ne sont pas extrêmes.

Exemple 7.10. Considérons pour Q le carré dont les quatre coins sont les points (-3,-1), (-3,1), (-1,-1) et (-1,1) comme sur la figure 7.4. Soit

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 0 \end{pmatrix} \text{ et } b = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 3 \end{pmatrix},$$

de sorte que $I_+ = \{1, 3, 4\}$ and $I_- = \{2\}$. Les rayons extrêmes de K sont les vecteurs $e_2 = (0, 1, 0, 0)$, $a_{21} = (1, 1, 0, 0)$, $a_{23} = (0, 1, 1, 0)$ et $a_{24} = (0, 3, 0, 1)$. Leurs images par $A^\top$ sont respectivement $(1, 0)$, $(1, 1)$, $(1, -1)$ et $(2, 0)$. On voit que les rayons extrêmes de N_Q sont donnés par le second et le troisième vecteurs ($(1, 1)$ et $(1, -1)$) tandis que les deux autres ne sont pas extrêmes.

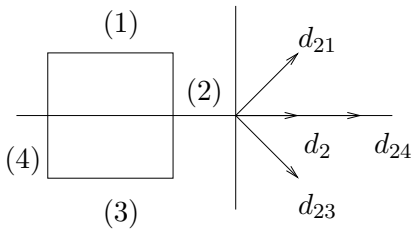


FIG. 7.4 – Générateurs du cône normal

7.2.2 Caractérisation du polaire inverse

Quand le polyèdre Q est de dimension pleine et qu'une description par inégalités de Q est disponible, les points extrêmes de Q^- peuvent être trouvés facilement.

Lemme 7.11 (Points extrêmes de Q^-)**.** *Soit $Q \subsetneq \mathbb{R}^n$ un polyèdre de dimension pleine explicitement décrit pas un ensemble d'inégalités $Ax \leq b$, avec $O \notin Q$. On note a_i les lignes de A.*

1. *Si la description $Ax \leq b$ est minimale, alors les points extrêmes de Q^- sont exactement les*
$$P_i := -\frac{a_i^\top}{b_i} \text{ avec } i \in I_-. \tag{7.9}$$
2. *Si la description $Ax \leq b$ n'est pas minimale, les P_i définis à l'équation (7.9) contiennent tous les points extrêmes de Q^-, plus des points qui appartiennent à Q^- mais ne sont pas extrêmes.*

Démonstration. Quand la description $Ax \leq b$ de Q est minimale, on connaît les facettes de Q. Une application du théorème 6.6 fournit le résultat. Quand la description n'est pas minimale, on peut en extraire un sous-ensemble minimal d'inégalités qui fournissent les points extrêmes de Q^-. ■

Remarque 7.12 (Hypothèse de dimension pleine)**.** En pratique, Q peut ne pas être de dimension pleine. On verra dans la section 7.3 une manière de contourner ce problème.

On peut maintenant établir un important théorème qui caractérise Q^- quand Q est un polyèdre défini par des inégalités.

Théorème 7.13 (Caractérisation de Q^-)**.** *Soit $Q \subsetneq \mathbb{R}^n$ un polyèdre de dimension pleine décrit explicitement par un ensemble d'inégalités $Ax \leq b$, avec $0 \notin Q$. Soient I_-, I_0, I_+ définis par (7.5), l'ensemble E défini par (7.6) et les P_i pour $i \in I_-$ définis au théorème 7.11. On a :*
$$Q^- = conv(P_i \ ; \ i \in I_-) + \operatorname{cone} A^\top E. \tag{7.10}$$

Démonstration. Comme Q^- est un polyèdre, il est la somme de l'enveloppe convexe de ses points extrêmes et de son cône de récession $(Q^-)_\infty$. Le théorème 7.11 donne les points extrêmes ; le théorème 6.3 affirme que $(Q^-)_\infty = N_Q$ et le théorème 7.7 dit que $N_Q = \operatorname{cone} A^\top E$. ■

Les résultats établis jusqu'ici s'appliquent à la programmation disjonctive, définie dans [Bal79].

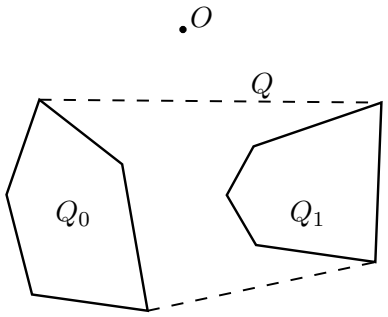


FIG. 7.5 – Programmation disjonctive

Définition 7.14 (Programmation disjonctive)**.** Soient Q_0 et Q_1 deux polyèdres non vides de $\mathbb{R}^n$, tels que $O \notin \overline{conv}(Q_0 \cup Q_1)$. L'ensemble

$$Q = \overline{conv}(Q_0 \cup Q_1)$$

est un polyèdre [Bal79], appelé *polyèdre disjonctif* (figure 7.5). Une coupe qui sépare O de Q est appelée une coupe disjonctive.

La sous-section suivante illustre l'idée de la programmation disjonctive. Elle est consacrée à un type particulier de disjonctions, appelées *disjonctions split*.

7.2.3 Disjonctions split

Considérons un problème de programmation linéaire en nombres entiers mixte, de la forme

$$\begin{cases} \min\ c \cdot x \\ (x_1, \ldots, x_n) \in \mathbb{Z}^n \\ (x_{n+1}, \ldots, x_{n+p}) \in \mathbb{R}^p \\ Ax \leq b \end{cases} \tag{7.11}$$

où $n, p, m \in \mathbb{N}$, $A \in \mathcal{M}_{m,n+p}(\mathbb{R})$, $c \in \mathbb{R}^n$ et $b \in \mathbb{R}^m$. Les algorithmes utilisés en pratique pour résoudre de tels problèmes utilisent souvent la *relaxation linéaire* suivante de (7.11) :

$$\begin{cases} \min\ c \cdot x \\ (x_1, \ldots, x_{n+p}) \in \mathbb{R}^{n+p} \\ Ax \leq b. \end{cases} \tag{7.12}$$

qui est beaucoup plus facile à résoudre. Le problème de la génération de coupes est de séparer une solution optimale $\overline{x}$ de (7.12) de l'ensemble réalisable dans (7.11). Soit R le polyèdre relaxé :

$$R = \{x \in \mathbb{R}^{n+p} \ ; \ Ax \leq b\}$$

et soit $j \in 1 \dots n$ tel que $\overline{x}_j \notin \mathbb{Z}$ (si un tel j n'existe pas, (7.11) est résolu). L'ensemble réalisable de (7.11) est entièrement contenu dans l'union des deux polyèdres suivants

$$\begin{array}{l} Q_0 = R \cap \{x \in \mathbb{R}^n \ ; \ x_j \leq \lfloor \overline{x}_j \rfloor\} \\ Q_1 = R \cap \{x \in \mathbb{R}^n \ ; \ x_j \geq \lfloor \overline{x}_j \rfloor + 1\} \end{array} \tag{7.13}$$

alors que $\overline{x}$ ne l'est pas (figure 7.6). En supposant que $\overline{x}$ est un point extrême de R, il ne peut pas appartenir à $\overline{conv}(Q_0 \cup Q_1)$. Pour le séparer de l'ensemble réalisable dans (7.11), il suffit de le séparer de $\overline{conv}(Q_0 \cup Q_1)$.

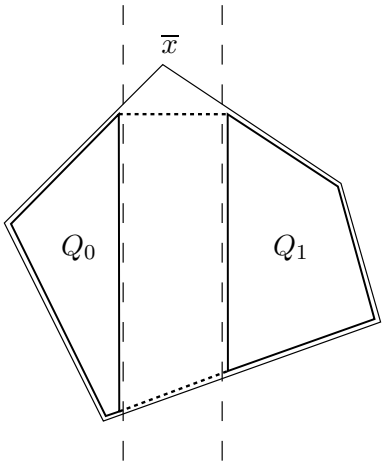


FIG. 7.6 – Cas particulier d'une disjonction split

7.2.4 Polaire inverse d'un polyèdre split

Revenant au cas étudié jusqu'ici où le point à séparer est l'origine O, considérons le problème suivant : soit $R \subsetneq \mathbb{R}^n$ un polyèdre de dimension pleine décrit explicitement par m inégalités $a_i \cdot x \leq b_i$ $(i = 1 \dots m)$ ou, sous forme matricielle, $Ax \leq b$, tel que O soit un point extrême de R (en particulier, $b \geq 0$ puisque O est réalisable). Soient $\pi \in \mathbb{R}^n$ et $\pi_0 < 0 < \pi_1$.

Remarque 7.15. En relation avec la sous-section précédente, O peut être vu comme la translation de la solution relaxée courante $\overline{x}$, R comme la translation du polyèdre relaxé, et π comme le j-ème vecteur de base.

On introduit la disjonction

$$\begin{array}{l} Q_0 = R \cap \{x \ : \ \pi \cdot x \leq \pi_0\} \\ Q_1 = R \cap \{x \ : \ \pi \cdot x \geq \pi_1\} \end{array}$$

et soit $Q = \overline{conv}(Q_0 \cup Q_1)$. Trouver une coupe disjonctive pour $\overline{x}$ dans le problème original (7.12) revient à séparer O et Q dans le problème translaté. Le théorème 7.16 particularise le théorème 7.13 lorsque Q est un polyèdre split, et donne une description explicite compacte de Q^-.

Théorème 7.16. *On suppose que $Q_0 \neq \emptyset$, $Q_1 \neq \emptyset$ et on définit*

$$\begin{aligned} P_0 &= \tfrac{\pi}{|\pi_0|}, \;\; K_0 = \operatorname{cone} A^\top + \tfrac{\pi b^\top}{|\pi_0|}, \\ P_1 &= -\tfrac{\pi}{\pi_1}, \;\; K_1 = \operatorname{cone} A^\top - \tfrac{\pi b^\top}{\pi_1}. \end{aligned}$$

Alors Q_0^- et Q_1^- sont des cônes polyédraux translatés :

$$\begin{aligned} Q_0^- &= P_0 + K_0 + \pi\mathbb{R}_+, \\ Q_1^- &= P_1 + K_1 - \pi\mathbb{R}_+. \end{aligned}$$

Démonstration. On applique le théorème 7.13 à Q_0 et Q_1 en remarquant que O satisfait toutes les contraintes qui définissent les Q_i, sauf l'inégalité de split : I_- dans (7.5) est un singleton, chaque Q_i^- a seulement un point extrême décrit par le théorème 7.11. Le nombre de générateurs donnés par le théorème 7.7 est

$$|E| = |I_-| + |I_0| + |I_-|\,|I_+| = 1 + |I_0| + |I_+| = 1 + m.$$

■

Ainsi la description de Q_0 et Q_1, quand ils viennent d'une disjonction, est particulièrement simple puisque I_- est un singleton. D'autre part, l'hypothèse $Q_0 \neq \emptyset$ et $Q_1 \neq \emptyset$ peut être difficile à vérifier en pratique. Ce problème est considéré dans la sous-section 7.3.3. Explicitement, on a

$$\left\{ \begin{array}{l} d \in Q^- \iff \exists\ (u_0, \lambda_0, u_1, \lambda_1) \in \mathbb{R}_+^{2m+2} \text{ tel que} \\ d = (A^\top + \tfrac{\pi b^\top}{|\pi_0|})u_0 + (\tfrac{1}{|\pi_0|} + \lambda_0)\pi = (A^\top - \tfrac{\pi b^\top}{\pi_1})u_1 - (\tfrac{1}{\pi_1} + \lambda_1)\pi\}. \end{array} \right. \tag{7.14}$$

Notre objectif est d'implémenter un oracle qui optimise des fonctions linéaires $d \to d_0 \cdot d$ sur $d \in Q^-$, et la formulation (7.14) convient. Il suffit de résoudre le programme linéaire

$$\left\{ \begin{array}{l} \min\ d_0 \cdot d \\ d = (A^\top + \tfrac{\pi b^\top}{|\pi_0|})u_0 + (\tfrac{1}{|\pi_0|} + \lambda_0)\pi = (A^\top - \tfrac{\pi b^\top}{\pi_1})u_1 - (\tfrac{1}{\pi_1} + \lambda_1)\pi, \\ u_i, \lambda_i \geq 0. \end{array} \right. \tag{7.15}$$

Cette description linéaire a $2m + 2$ variables dans $\mathbb{R}^+$ et n contraintes d'égalité. Dans la sous-section suivante, on explique comment ce programme linéaire est relié au *cut generating LP* de la méthode lift-and-project.

Remarque 7.17. La formulation (7.14) peut être interprétée de la manière suivante. Soit $(a_i)_{i=1\dots m}$ les lignes de A. En combinant les équations $a_i x \leq b_i$ et $\pi \cdot x \leq \pi_0$ (avec $\pi_0 < 0$), on voit que tout $x \in Q_0$ satisfait

$$(a_i^\top + \frac{b_i}{|\pi_0|}\pi) \cdot x \leq 0 \quad \text{et} \quad \pi \cdot x \leq \pi_0. \tag{7.16}$$

Pour tous multiplicateurs $(u_0, \lambda_0) \geq 0$, l'inégalité

$$\left((A^\top + \frac{\pi b^\top}{|\pi_0|})u_0 + (\frac{1}{|\pi_0|} + \lambda_0)\pi\right) \cdot x \leq -1 - \lambda_0|\pi_0| \leq -1 \tag{7.17}$$

est valide pour Q_0 : si d est le vecteur qui multiplie x dans (7.17), $\sigma_{Q_0}(d) \leq -1$. Le même argument est valable pour Q_1. Finalement, tout d satisfaisant (7.14) satisfait aussi $d \cdot x \leq -1$ pour tout x dans $Q_0 \cup Q_1$, donc est une inégalité valide pour $\overline{conv}(Q_0 \cup Q_1)$. En un mot, $d \in Q^-$.

Remarque 7.18. On voit aussi dans (7.17) que $1 + \lambda_0|\pi_0|$ est la quantité par laquelle l'origine viole la contrainte correspondant à (u_0, λ_0) ; de même pour $1 + \lambda_1\pi_1$ et (u_1, λ_1). Pour obtenir une « coupe la plus profonde » au sens de lift-and-project, c'est-à-dire une coupe la plus violée, il suffit d'ajouter la contrainte

$$\nu = -1 - \lambda_0|\pi_0| = -1 - \lambda_1\pi_1$$

(plus une contrainte de normalisation) et de minimiser ν.

Le lien entre le CGLP de lift-and-project et notre oracle est expliqué plus précisément dans la sous-section suivante.

7.2.5 Lien avec lift-and-project

La description suivante de la méthode lift-and-project pour les programmes 0-1 est adaptée de [Cor08]. Soit $P := \{x \in [0,1]^n : \ Ax \leq b\}$ et $S := \{x \in \{0,1\}^n : \ Ax \leq b\}$, avec des notations usuelles, et supposons sans perte de généralité que les contraintes $Ax \leq b$ contiennent $0 \leq x \leq 1$. On choisit $j \in \{1, \dots n\}$ et on considère l'ensemble (décrit par des équations non linéaires) : $\{x \in [0,1]^n : \ x_j(Ax - b) \leq 0, \ (1 - x_j)(Ax - b) \leq 0\}$. Ensuite, on linéarise les équations en substituant les nouvelles variables y_i pour $x_i x_j$, $i \neq j$ et x_j pour x_j^2 : en notant A_j la matrice A privée de sa j-ième colonne a_j, cette substitution définit le polyèdre

$$\left\{ \begin{array}{ll} M_j := \{ & (x,y) \in [0,1]^{2n-1} : \\ & A_j y + (a_j - b)x_j \leq 0, \\ & Ax - A_j y - (a_j - b)x_j \leq b\}. \end{array} \right. \tag{7.18}$$

Soit Q la projection de M_j sur l'espace des x : on a $S \subset Q \subset P$, donc Q est une relaxation de l'ensemble réalisable S qui est au moins aussi bonne que P. Plus précisément, Q est exactement le polyèdre disjonctif associé à la disjonction split $[x_j = 0] \vee [x_j = 1]$, comme l'affirme le lemme suivant.

Lemme 7.19. *Soit* $P_0 := \{x \in P \ : \ x_j = 0\}$ *et* $P_1 := \{x \in P \ : \ x_j = 1\}$. *Alors*

$$Q = \overline{conv}(P_0 \cup P_1).$$

Démonstration. On introduit les nouvelles variables y' et z', et (7.18) se récrit

$$\left\{ \begin{array}{ll} Q = \{ & (x,y) \in [0,1]^{2n-1} : \ \exists y' \in P_1, \ z' \in P_0 \\ & x = x_j y' + (1-x_j) z', \\ & y_i = x_j y'_i \ (i \neq j)\} \end{array} \right. \tag{7.19}$$

de sorte que x est explicitement écrit comme une combinaison convexe de deux éléments $y' \in P_1$ et $z' \in P_0$. La variable x_j et son complément $1 - x_j$ jouent le rôle de multiplicateurs convexes. ∎

En utilisant le lemme de Farkas pour projeter M_j sur l'espace des x, on a

$$Q = \{x \in [0,1]^n : \ d \cdot x \leq e, \ (d,e) \in C\} \tag{7.20}$$

où

$$\begin{array}{ll} C = \{ & (d,e) \in \mathbb{R}^{n+1} : \exists u_0, u_1 \geq 0, \\ & d = A^\top u_0 + (a_j - b) \cdot (u_1 - u_0) e_j, \\ & e = b \cdot u_0, \\ & A_j^\top u_0 = A_j^\top u_1\}. \end{array} \tag{7.21}$$

L'ensemble C contient exactement les inégalités valides pour P. La méthode lift-and-project consiste à chercher une inégalité dans C qui est violée le plus possible par un certain point à séparer $\overline{x}$:

$$\max_{(d,e) \in C} d \cdot \overline{x} - e.$$

Atteindre une valeur positive dans ce problème de maximisation signifie que l'inégalité (d,e) sépare $\overline{x}$ et Q. Comme (u_0, u_1) dans (7.21) appartiennent à un cône, il est nécessaire de tronquer ce cône par une contrainte du type $f(u_0, u_1) \leq N$ (où f est une fonction, typiquement linéaire, et N une constante) afin que le problème d'optimisation n'atteigne pas $+\infty$. Soit $\tilde{C}$ l'ensemble obtenu en tronquant C par la contrainte de normalisation ; la génération d'une coupe lift-and-project consiste à résoudre

$$\max_{(d,e) \in \tilde{C}} d \cdot \overline{x} - e \tag{7.22}$$

tandis qu'un appel à notre oracle qui minimise des fonctions linéaires de la forme $d \rightarrow d_0 \cdot d$ pour $d \in Q^-$ consiste essentiellement à résoudre

$$\min_{(d,e) \in \tilde{C}, \ d \cdot \overline{x} - e \geq 1} d_0 \cdot d \tag{7.23}$$

Autrement dit, au lieu de maximiser la violation, on impose une violation minimum de 1 et on minimise le produit scalaire de la direction de coupe avec un vecteur donné d_0 (déterminé par la génération de colonnes). On remarque que la contrainte de normalisation

$f(u_0, u_1) \leq N$ dans (7.23) ne doit pas être trop forte, sinon aucune coupe n'atteindra la violation minimum de 1 et l'ensemble réalisable sera vide. Dans la sous-section 7.3.2, on propose une manière de résoudre ce problème en choisissant une contrainte de normalisation qui assure que l'ensemble faisable contienne au moins une coupe (la coupe d'intersection générée depuis la base courante, décrite dans la sous-section 7.2.6).

Remarque 7.20. Finalement, il n'y a pas beaucoup de différences entre le CGLP de lift-and-project et notre oracle qui optimise sur le polaire inverse (d'ailleurs, les deux programmes linéaires sont obtenus en utilisant le même argument basé sur le lemme de Farkas). Notre description est un peu plus géométrique, puisqu'elle fait apparaître Q^- comme l'intersection des deux cônes translatés Q_0^- et Q_1^- qui sont les polaires inverses des polyèdres utilisés dans la disjonction.

La différence principale est qu'en général, chaque polyèdre disjonctif possède $m+1$ contraintes (m étant le nombre de contraintes dans le problème original) à cause de la contrainte disjonctive supplémentaire, de sorte que $m+1$ multiplicateurs sont nécessaires pour chaque polyèdre. C'est pourquoi (7.15) a $2m+2$ variables. Dans le cas particulier d'une disjonction split pour un programme 0-1, cependant, les polyèdres disjonctifs sont seulement soumis aux m contraintes originales (la disjonction split est obtenue en fixant une variable à 0 ou 1, pas en ajoutant une contrainte). C'est pourquoi (7.23) a $2m$ variables.

Une méthode due à Balas et Perregaard [PB03] permet de travailler avec seulement m variables au lieu de $2m$ dans (7.22), réduisant ainsi le temps de calcul. Elle pourrait être utilisée aussi dans (7.23).

7.2.6 Une coupe particulière

La description (7.14) de Q^- permet de trouver un point particulier de Q^-, c'est-à-dire une coupe. Soit $B \subset \{1, \dots m\}$ la base courante de A, correspondant au point que l'on veut séparer. Autrement dit, on choisit n contraintes indépendantes qui sont actives à l'origine ; leur membre de droite est donc nul : pour tout $i \in B$, $b_i = 0$. On fixe $\lambda_0 = \lambda_1 = 0$ et $(u_0)_i = (u_1)_i = 0$ pour tout $i \notin B$, on a $u_0 \cdot b = u_1 \cdot b = 0$. L'équation (7.14) s'écrit :

$$A^\top (u_0 - u_1) = -(\frac{1}{|\pi_0|} + \frac{1}{\pi_1})\pi, \ u_i \geq 0. \tag{7.24}$$

On note A_B la matrice A privée de ses lignes hors-base $i \notin B$, donc A_B est une matrice carrée inversible. Une solution du système ci-dessus peut être obtenue en résolvant

$$A_B^\top u = -(\frac{1}{|\pi_0|} + \frac{1}{\pi_1})\pi, \ u \in \mathbb{R}^n \tag{7.25}$$

qui a une solution unique, puisque B est une base ; en posant

$$\overline{u}_0 = \max(u, 0) \ \text{ et } \ \overline{u}_1 = -\min(u, 0) \tag{7.26}$$

(et en complétant avec des zéros pour les lignes hors-base de A) on obtient un point réalisable. Cette coupe est appelée la coupe d'intersection [Bal71] générée à partir de la base courante.

Remarque 7.21. Cette coupe simple peut être calculée assez rapidement (il suffit de résoudre un système linéaire), et donne une idée de la qualité de la disjonction choisie [KC05]. Elle peut être utilisée pour choisir une direction prometteuse avant de démarrer l'algorithme.

7.3 Algorithme

Les sections précédentes suggèrent de calculer des coupes splits en projetant l'origine sur le polaire inverse, dans l'espace dual, tout en décomposant cette projection comme une combinaison convexe de points extrêmes, et d'utiliser ces points extrêmes de Q^- comme directions de coupe. Mais avant d'utiliser cette approche, considérons les difficultés suivantes.

- Le théorème 6.6 n'est pas vrai lorsque $lin(Q) \neq \mathbb{R}^n$.
- La caractérisation du cône normal dans le théorème 7.7 est perdue si le polyèdre est vide.
- L'algorithme 7.1 ne peut être utilisé que sur un polyèdre borné (un polytope) ; sinon, le programme linéaire de l'étape 2 peut être non borné.
- Selon (7.14), Q^- est décrit *via* les variables $(u_0, u_1) \in \mathbb{R}^{2m+2}$. Si le LP à l'étape 2 de l'algorithme 7.1 a plusieurs solutions, une solution extrémale dans l'espace étendu peut ne pas être extrémale dans $\mathbb{R}^n$.

7.3.1 Quand $lin(Q) \neq \mathbb{R}^n$

Quand $lin(Q) \neq \mathbb{R}^n$, le théorème 6.6 n'est pas vrai : les facettes de Q ne correspondent plus à des points extrêmes de Q^-. Théoriquement, ceci peut être évité en travaillant dans l'espace euclidien $lin(Q)$ au lieu de $\mathbb{R}^n$. Dans les applications, nous allons simplement ignorer ce problème et faire comme si $lin(Q)$ était égal à l'espace entier[1] $\mathbb{R}^n$. Pour cette raison, Q^- peut ne pas avoir de point extrême du tout : cependant, grâce à la contrainte de normalisation présentée dans la section suivante, ce cas ne sera jamais rencontré car Q^- sera remplacé par un ensemble borné $\tilde{Q}^-$.

7.3.2 Contrainte de normalisation

Il est clair d'après la définition de Q^- que cet ensemble est non borné. Malheureusement, l'algorithme de projections successives ne fonctionne que si le polyèdre est un polytope, et nous devons donc borner Q^- artificiellement d'une manière ou d'une autre pour utiliser notre méthode. La description (7.14) suggère de borner Q^- en le remplaçant par $\tilde{Q}^-$, défini en bornant la somme des multiplicateurs (comme cela est souvent

[1] Ou $lin(Q) = \{A_e x = b_e\}$ si le problème a des contraintes d'égalité $A_e x = b_e$.

fait dans la méthode lift-and-project) :

$$\sum_{i=1}^{m}(u_0)_i + \lambda_0 + \sum_{i=1}^{m}(u_1)_i + \lambda_1 \leq N \tag{7.27}$$

où N est une constante de normalisation suffisamment grande. Il suffit de calculer la coupe particulière décrite dans la sous-section 7.2.6 pour obtenir une valeur de N qui assure $\tilde{Q}^- \neq \emptyset$. Plus précisément, il suffit de poser $N := \sum_i(\overline{u}_0)_i + (\overline{u}_1)_i$ avec les notations de l'équation (7.26).

Remarque 7.22 (Normalisation)**.** En remplaçant Q^- par $\tilde{Q}^-$, on introduit de nouveaux points extrêmes qui ne correspondent pas à des facettes de Q. Les coupes correspondantes n'exposeront donc pas de facette. Ceci semble inévitable, puisqu'on a vu sur la figure 7.2 que la coupe la plus profonde n'était pas nécessairement impliquée par des coupes de facettes. Si la décomposition en coupes de facettes est possible, alors le choix de la normalisation est complètement arbitraire (on obtiendra toujours des facettes qui impliquent la coupe la plus profonde, quelle que soit la normalisation) pourvu que N soit assez grand. Si, au contraire, des rayons extrêmes de Q^- sont nécessaires pour décomposer la coupe la plus profonde, alors le choix de la normalisation importe : des normalisations différentes fourniront des coupes différentes, selon le point d'intersection de la contrainte de normalisation et du rayon extrême.

La figure 7.7 illustre le problème posé par la situation où $lin(Q) \neq \mathbb{R}^n$ et le rôle de la contrainte de normalisation. Q est défini par les contraintes $a_i \cdot x \leq b_i$ $(i = 1, \ldots, 3)$ et $lin(Q)$ est l'axe horizontal. La projection de O sur $\tilde{Q}^-$ est décomposée en une combinaison convexe des points A et B, et les deux coupes générées exposent la même facette F du polyèdre Q.

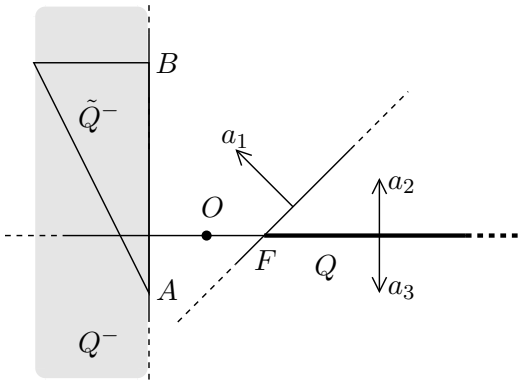


FIG. 7.7 – Un exemple où $lin(Q) \neq \mathbb{R}^n$

7.3.3 Cône normal d'un polyèdre vide

Dans cette sous-section, on montre que le lemme 7.5 n'est pas vrai sans l'hypothèse que le polyèdre est non vide. Soit $n = 2$ et

$$A = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

Alors $Q := \{(x = (x_1, x_2) \in \mathbb{R}^2 \ ; \ Ax \leq b\} = \emptyset$. L'ensemble des $u \in \mathbb{R}^2_+$ tels que $b \cdot u \leq 0$ est simplement $\mathbb{R}^2_+$ et l'ensemble $A^\top u$ pour de tels u est $\mathbb{R} \times \{0\}$. Si le lemme 7.5 était vrai dans ce cas, le cône normal de Q serait $\mathbb{R} \times \{0\}$, mais comme Q est vide, son cône normal est $\mathbb{R}^2$ tout entier.

Dans notre situation, $Q = \overline{conv}(Q_0 \cup Q_1)$ peut être supposé non vide (sinon le problème de séparation n'a pas beaucoup de sens). Cependant, Q_0 ou Q_1 (mais pas les deux) peuvent être vides, ce qui est coûteux à vérifier en pratique. Cette difficulté peut sans doute être ignorée, car si par exemple $Q_0 = \emptyset$, de sorte que $Q^- = Q_1^-$, peu importe quel ensemble (appelons-le $\bar{Q}_0^-$) est obtenu par le lemme 7.5 à la place de Q_0^- : les éléments de $\bar{Q}_0^- \cap Q_1^-$ appartiendront à $Q_1^- = Q^-$. Autrement dit, ces éléments seront des inégalités valides pour Q, et dans notre application à la programmation entière, il n'y a pas de risque de couper une solution réalisable.

7.3.4 L'oracle peut retourner un point non extrême

Dans l'algorithme 7.1, l'oracle appelé à l'étape 2 doit retourner un point extrême de Q^-. Cependant, dans notre application à la programmation disjonctive, ceci ne peut pas être garanti. La situation de la figure 7.8, où Q^- est l'intersection de deux cônes comme dans (7.14), peut avoir lieu. Tout le segment AB est optimal pour la minimisation de $x \mapsto c \cdot x$ sur Q^-, et les points A, B et C (qui n'est pas extrême dans Q^-) peuvent tous les trois être retournés par le solveur qui minimise dans l'espace étendu, avec la description (7.14) de Q^-. Ce problème, cependant, n'apparaît que dans des cas très particuliers et s'il a lieu, on court seulement le risque d'avoir dans la décomposition de la coupe la plus profonde une coupe qui n'expose pas de facette, mais au moins est valide.

7.3.5 Algorithme

Pour résumer, on propose la méthode suivante pour générer des coupes disjonctives.

– On considère un programme linéaire en nombres entiers mixte et on obtient un point extrême $\overline{x}$ de sa relaxation linéaire (7.12). On choisit une coordonnée i telle que $\overline{x}_i$ n'est pas entier alors que x_i est contraint à être entier dans (7.11). On peut aussi choisir une disjonction plus générale.
– On calcule la coupe d'intersection associée à la base courante (sous-section 7.2.6).
– A partir de cette coupe, on calcule la coupe la plus violée (sous-section 7.2.5).

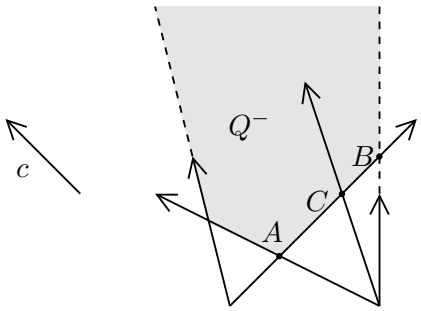


FIG. 7.8 – Le point retourné par l'oracle peut n'être pas extrême

– A partir de la coupe la plus violée, on effectue l'algorithme de projections successives (sous-section 7.1) pour calculer la projection de l'origine O sur $\tilde{Q}^-$ et sa décomposition en points extrêmes $d_1, \ldots, d_k$ de Q^-.

Pour finir, on utilise les inégalités $d_i \cdot x \leq -1$, $i = 1, \ldots, k$ comme coupes. Cet algorithme de génération de coupes est testé dans la section suivante sur des instances aléatoires de programmes linéaires en nombres entiers.

7.4 Expériences numériques

On teste l'algorithme sur des instances aléatoires de programmes linéaires en nombres entiers, en dimension n allant de 20 à 90 avec des matrices pleines. On résout la relaxation LP pour obtenir la valeur optimale relaxée $\overline{f}$, et on choisit la variable la plus fractionnelle comme direction de split. On obtient (i) une coupe la plus violée (coupe lift-and-project), (ii) la coupe la plus profonde au sens de la norme euclidienne, et (iii) la décomposition de la coupe la plus profonde en coupes de facettes. On ajoute ces coupes à la matrice, et on obtient de nouvelles valeurs optimales relaxées $\overline{f}_{(i)}$, $\overline{f}_{(ii)}$ et $\overline{f}_{(iii)}$. On reporte

– l'amélioration relative de la fonction objectif avec nos coupes par rapport à la coupe la plus profonde en % $(= 100\ (\overline{f}_{(iii)} - \overline{f})/(\overline{f}_{(ii)} - \overline{f})$,
– l'amélioration relative de la fonction objectif entre la coupe la plus violée et la coupe la plus profonde en % $(= 100\ (\overline{f}_{(i)} - \overline{f})/(\overline{f}_{(ii)} - \overline{f}))$,
– la profondeur relative entre la coupe la plus violée et la coupe la plus profonde en % $(= 100\ \text{depth}(i)/\text{depth}(ii))$,
– le nombre de coupes générées par notre algorithme,
– le nombre d'appels à l'oracle LP.

Les résultats sont rassemblés dans la table 7.1, où le nom du problème contient le nombre de variables et le nombre de contraintes (m) en plus de la positivité des variables.

On n'utilise pas de technique de renforcement (e.g. Balas-Jeroslow [BJ80]) : le but de ces expériences est seulement d'évaluer la qualité relative de la coupe la plus violée, de la coupe la plus profonde, et de nos coupes.

Probl eme	Wolfe/+prof. (obj. en %)	+viol./+prof. (obj. en %)	+viol./+prof (prof. en %)	#coupes	#appels
n20-m20-s0	474.35	409.85	51.15	14	18
n20-m20-s1	859.17	1037.70	60.70	14	15
n20-m20-s2	100.00	100.00	77.91	14	15
n30-m30-s0	865.46	866.85	52.57	25	26
n30-m30-s1	374.86	457.82	63.42	20	22
n30-m30-s2	107.81	107.81	6.11	22	23
n40-m40-s0	100.00	100.50	61.61	29	30
n40-m40-s1	333.48	147.95	60.60	29	57
n40-m40-s2	2690.61	1565.30	47.39	29	35
n50-m50-s0	362.31	362.31	60.95	37	38
n50-m50-s1	350.37	427.96	58.50	39	40
n60-m60-s0	100.76	100.00	59.84	42	43
n70-m70-s0	100.00	100.01	52.95	55	86
n80-m80-s0	8877.09	10593.63	50.03	65	70
n90-m90-s0	460.21	434.84	49.35	67	135

TAB. 7.1 – Résultats expérimentaux

A partir du tableau 7.1, on observe que la coupe la plus violée et nos coupes apportent une amélioration très importante de la fonction objectif par rapport à la coupe la plus profonde, qui semble en général assez faible de ce point de vue. De manière intéressante, les performances de la coupe la plus violée et de nos coupes sont comparables en terme d'amélioration de la fonction objectif (bien sûr, ceci est biaisé par le fait qu'on ajoute une seule coupe d'un côté, et plusieurs de l'autre). On observe aussi que la coupe la plus violée est en gros deux fois moins profonde que la coupe la plus profonde en général, et que le nombre d'appels à l'oracle nécessaires n'est pas beaucoup plus grand que le nombre de facettes dans la décomposition.

Finalement, nos coupes semblent apporter la même amélioration de la fonction objectif que la coupe la plus violée, tout en étant environ deux fois plus profondes. Le prix à payer est le coût de calcul pour ré-optimiser plusieurs fois le CGLP, et l'ajout d'un grand nombre de coupes à la matrice ; de plus, ces coupes sont pleines (*i.e.* pas creuses) en général. Ceci suggère qu'une procédure est nécessaire pour sélectionner seulement quelques coupes parmi les nombreuses qui sont générées. Une meilleure implémentation est aussi nécessaire pour évaluer le compromis entre le coût de la procédure et son efficacité ; cependant, on peut déjà tirer les conclusions suivantes.

– Contrairement à une idée répandue, la coupe la plus profonde au sens de la norme euclidienne n'améliore pas tellement la fonction objectif. La coupe lift-and-project, qui est aussi une coupe la plus profonde au sens d'une norme ou semi-norme « imprévisible », donne de bien meilleurs résultats. Il semble préférable d'utiliser une (semi-)norme adaptée à chaque problème, comme dans la méthode lift-and-

project, plutôt qu'une norme décidée à l'avance (la norme euclidienne).
- L'approche proposée n'est pas attrayante du point de vue pratique. Le gain de profondeur par rapport à la coupe lift-and-project, observé dans les expériences, est insuffisant pour compenser le surcoût en calcul lors des ré-optimisations successives du CGLP, et l'ajout de nombreuses coupes à la matrice.

Pour rendre cette approche plus pratique, on pourrait décomposer une autre coupe donnée – et non pas la plus profonde – en facettes : par exemple la coupe la plus violée, ou celle qui améliore le plus la fonction objectif. Malheureusement, on ne sait pour l'instant pas tirer parti des idées développées dans ce chapitre pour générer des coupes efficaces dans ce contexte.

Conclusion de la seconde partie

Cette partie, plus abstraite que la première, concerne le problème de séparation en analyse convexe. On veut trouver de bons hyperplans qui séparent O du convexe Q. Le polaire inverse Q^- de Q détermine les directions de coupe, et de plus, quand Q est un polyèdre, ses points extrêmes (à intersection près avec un sous-espace vectoriel, si Q n'est pas de dimension pleine) correspondent aux coupes qui exposent des facettes de Q. De plus, le problème de projection de O sur Q, pour déterminer la coupe la plus profonde, est équivalent au problème de projection de O sur Q^- au sens de la norme duale. On suggère donc de projeter O sur Q^- (au sens de la norme euclidienne) et la décomposition de cette projection en points extrêmes de Q^-. On utilise ensuite ces points extrêmes pour générer des coupes qui séparent O et Q.

Cette méthode s'applique à la programmation disjonctive, qui permet de générer des coupes pour la programmation entière. Son importance théorique dans ce contexte peut être défendue, mais son utilité numérique est pour l'instant douteuse, même si des pistes d'améliorations sont envisageables. Une autre application de la programmation disjonctive, plus directement reliée aux préoccupations de la première partie de cette thèse sur la mécanique non régulière, est l'étude des problèmes de complémentarité. En effet, ceux-ci se présentent souvent naturellement sous la forme d'une disjonction. Un programme de recherche intéressant serait donc d'appliquer les idées de l'optimisation combinatoire (« brancher, borner et couper » – pour les synthétiser à l'extrême) aux problèmes de dynamique non-régulière. Nous espérons que ces idées perdureront dans l'équipe Bipop et seront utilisées par ses membres pour l'étude des problèmes de mécanique et d'électronique qui les motivent. Nous leur en laissons le soin.

Annexe A

Pendule double

Cet exemple très simple est destiné à illustrer la méthode de mise en équation lagrangienne. Il comporte deux tiges rigides articulées par des liaisons pivot parfaites et portant à leur extrémité des masses m_1 et m_2, comme sur la figure 2.1 du chapitre 2. Ce système à deux degrés de liberté est paramétré par $q := (\theta_1, \theta_2)$. La position de M_1 et M_2 est donnée par

$$M_1 = \begin{bmatrix} l_1 \sin(\theta_1) \\ -l_1 \cos(\theta_1) \end{bmatrix}, \quad M_2 = \begin{bmatrix} l_1 \sin(\theta_1) + l_2 \sin(\theta_2) \\ -l_1 \cos(\theta_1) - l_2 \cos(\theta_2) \end{bmatrix}. \tag{A.1}$$

Les matrices jacobiennes $\frac{\partial}{\partial q}(M_1)$ et $\frac{\partial}{\partial q}(M_2)$ sont

$$\frac{\partial}{\partial q}(M_1) = \begin{bmatrix} l_1 \cos(\theta_1) & 0 \\ l_1 \sin(\theta_1) & 0 \end{bmatrix}, \quad \frac{\partial}{\partial q}(M_2) = \begin{bmatrix} l_1 \cos(\theta_1) & l_2 \cos(\theta_2) \\ l_1 \sin(\theta_1) & l_2 \sin(\theta_2) \end{bmatrix} \tag{A.2}$$

et l'énergie cinétique E_c du système est

$$\begin{aligned} E_c &:= \frac{1}{2} m_1 \dot{M}_1^2 + \frac{1}{2} m_2 \dot{M}_2^2 \\ &= \frac{1}{2}(m_1 + m_2)\, l_1^2\, \dot{\theta}_1^2 + \frac{1}{2} m_2 (l_2^2\, \dot{\theta}_2^2 + 2 l_1 l_2 \cos(\theta_1 - \theta_2) \dot{\theta}_1 \dot{\theta}_2). \end{aligned} \tag{A.3}$$

On calcule ensuite

$$\frac{\partial}{\partial \dot{q}}(E_c) = \begin{bmatrix} (m_1 + m_2) l_1^2 \dot{\theta}_1 + m_2 l_1 l_2 \cos(\theta_1 - \theta_2) \dot{\theta}_2 \\ m_2 l_1 l_2 \cos(\theta_1 - \theta_2) \dot{\theta}_1 + m_2 l_2^2 \dot{\theta}_2 \end{bmatrix} \tag{A.4}$$

puis

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial}{\partial \dot{q}}(E_c)\right) = \begin{bmatrix} (m_1 + m_2) l_1^2 & m_2 l_1 l_2 \cos(\theta_1 - \theta_2) \\ m_2 l_1 l_2 \cos(\theta_1 - \theta_2) & m_2 l_2^2 \end{bmatrix} \ddot{q} \\ - m_2 l_1 l_2 \sin(\theta_1 - \theta_2)(\dot{\theta}_1 - \dot{\theta}_2) \begin{bmatrix} \dot{\theta}_2 \\ \dot{\theta}_1 \end{bmatrix} \end{aligned} \tag{A.5}$$

et

$$\frac{\partial}{\partial q}(E_c) = m_2 l_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_1\dot{\theta}_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{A.6}$$

Autrement dit, les différents termes de (2.2) sont

$$\mathbf{M} = \begin{bmatrix} (m_1 + m_2)l_1^2 & m_2 l_1 l_2 \cos(\theta_1 - \theta_2) \\ m_2 l_1 l_2 \cos(\theta_1 - \theta_2) & m_2 l_2^2 \end{bmatrix},$$
$$\mathbf{N} = m_2 l_1 l_2 \sin(\theta_1 - \theta_2) \begin{bmatrix} \dot{\theta}_2^2 \\ \dot{\theta}_1^2 \end{bmatrix} \quad \text{et} \quad \mathbf{F}_{\text{int}} = 0. \tag{A.7}$$

Il reste à calculer le travail virtuel des forces extérieures :

$$\begin{aligned} \mathbf{F}_{\text{ext}} &= \left(\frac{\partial}{\partial q}(M_1)\right)^\top (m_1 g) + \left(\frac{\partial}{\partial q}(M_2)\right)^\top (m_2 g) \\ &= \begin{bmatrix} l_1(m_1 + m_2)(\cos(\theta_1)g_x + \sin(\theta_1)g_y) \\ l_2 m_2(\cos(\theta_2)g_x + \sin(\theta_2)g_y) \end{bmatrix}. \end{aligned} \tag{A.8}$$

Connaissant tous les termes $\mathbf{M}$, $\mathbf{N}$, $\mathbf{F}_{\text{int}}$ et $\mathbf{F}_{\text{ext}}$, on peut ensuite discrétiser l'équation de Lagrange comme à la section 2.2.

Annexe B

Angles d'Euler

Les angles d'Euler permettent de paramétrer l'orientation d'un repère par rapport à un autre dans $\mathbb{R}^3$. On rappelle ici leur définition, et on présente le problème du "gimbal lock" ou perte d'un degré de liberté, qui motive l'introduction des quaternions dans l'annexe C.

B.1 Définition

On considère un repère $Oxyz$ où la base $B = (x, y, z)$ est une base orthonormale. On effectue successivement trois rotations.

- La première rotation $R(\psi, z)$, d'angle ψ autour de l'axe Oz, est appelée *précession*. L'image du vecteur x par cette rotation est appelée u. Le vecteur rotation correspondant est $\dot{\psi}z$.
- La deuxième rotation $R(\theta, u)$, d'angle θ autour de u, est appelée *nutation*. L'image du vecteur z par cette rotation est appelée z'. Le vecteur rotation correspondant est $\dot{\theta}u$.

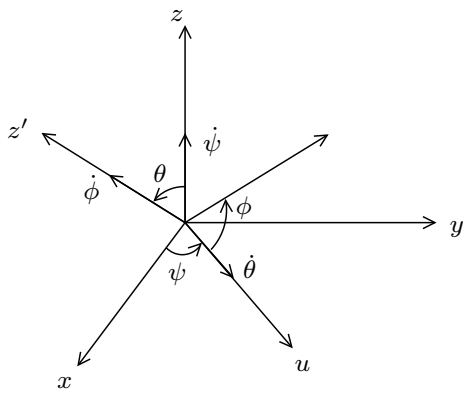


FIG. B.1 – Les angles d'Euler (ψ, θ, ϕ)

– La troisième rotation $R(\phi, z')$, d'angle ϕ autour de z', est appelée *rotation propre*. Le vecteur rotation correspondant est $\dot{\phi}z'$.

Le produit R de ces trois rotations (dans cet ordre), défini par

$$R := R(\phi, z') \circ R(\theta, u) \circ R(\psi, z) \tag{B.1}$$

permet de passer de la base B à la base d'Euler $B' = (x', y', z')$. Le vecteur rotation total de B' par rapport à B exprimé dans la base B est

$$\omega_{R'/R} = \dot{\psi}z + \dot{\theta}u + \dot{\phi}z'. \tag{B.2}$$

B.2 Perte d'un degré de liberté

Comme $z = [0, 0, 1]^\top$, $u = [\cos\psi, \sin\psi, 0]^\top$ et $z' = [\sin\theta\sin\psi, -\sin\theta\cos\psi, \cos\theta]^\top$, la formule (B.2) se récrit

$$\omega_{R'/R} = M \begin{bmatrix} \dot{\psi} \\ \dot{\theta} \\ \dot{\phi} \end{bmatrix} \text{ avec } M := \begin{bmatrix} 0 & \cos\psi & \sin\theta\sin\psi \\ 0 & \sin\psi & -\sin\theta\cos\psi \\ 1 & 0 & \cos\theta \end{bmatrix} \tag{B.3}$$

et le déterminant de la matrice M définie dans (B.3) vaut $\sin\theta$: lorsque $\theta = 0[\pi]$, elle n'est donc pas inversible. Dans ce cas, puisque $\mathrm{Im}(M) \neq \mathbb{R}^3$, il existe des vecteurs $\omega_{R'/R}$ dans $\mathbb{R}^3$ qui ne correspondent à aucune valeur de $(\dot{\psi}, \dot{\theta}, \dot{\phi})$. Ceci n'est pas physique (toutes les valeurs de $\omega_{R'/R}$ dans $\mathbb{R}^3$ sont mécaniquement possibles), c'est un problème artificiel introduit par le choix des angles d'Euler pour paramétrer l'orientation de la base B'.

B.3 Analogie mécanique

Le problème de singularité de la matrice M dans (B.3) est appelé *gimbal lock* ou "perte d'un degré de liberté". On peut le comprendre par une analogie avec le fonctionnement d'un joint de Cardan : lorsque deux des axes du cardan sont alignés (l'hypothèse $\theta = 0[\pi]$ ci-dessus signifie précisément que les axes z et z' de précession et de rotation propre sont alignés), ils autorisent tous les deux le même degré de liberté et il ne reste plus au cardan que deux degrés de liberté indépendants. Il cesse donc de jouer son rôle, et l'anneau intérieur ne peut plus adopter n'importe quel vecteur rotation par rapport à l'anneau extérieur.

Il existe plusieurs possibilités pour lutter contre le problème du gimbal lock ; ce dernier se traduit en pratique lors des simulations par l'apparition de matrices de masse non-inversibles lorsque les angles d'Euler sont choisis comme coordonnées généralisées et leur dérivée temporelle comme vitesse généralisée. Une solution possible est d'utiliser une matrice de rotation ; l'annexe C en présente une autre, consistant à paramétrer l'orientation de B' par rapport à B grâce à un quaternion unité.

Annexe C

Quaternions et représentation des rotations

Les quaternions ont été introduits par Hamilton en 1843, avec l'intention de généraliser la notion de nombre complexe. Il réalisa ensuite que, tout comme la multiplication par un nombre complexe de norme 1 peut être vue comme une rotation plane dans $\mathbb{C}$, la conjugaison par un quaternion peut être vue comme une rotation en dimension 3. En mécanique du solide, on peut donc paramétrer l'orientation d'un repère dans l'espace grâce à la donnée d'un quaternion. Par rapport à l'utilisation des angles d'Euler, on introduit une variable supplémentaire (paramétrage dans $\mathbb{R}^4$ au lieu de $\mathbb{R}^3$). En contrepartie, on élimine le problème de singularité artificielle ("gimbal lock") que posent les angles d'Euler ; cela est nécessaire lorsque la cinématique ou les propriétés du système n'assurent pas que les configurations singulières seront évitées. Cet appendice, inspiré de [CR99], introduit la notion de quaternions et rassemble les propriétés essentielles qui permettent de les utiliser en mécanique à la place des angles d'Euler.

C.1 Définition et premières propriétés

C.1.1 Définition

Définition C.1 (Quaternions)**.** Un quaternion q est un couple formé d'un réel a et d'un vecteur $\underline{b}$ de $\mathbb{R}^3$:

$$q := (a, \underline{b}) \in \mathbb{R} \times \mathbb{R}^3.$$

On note indifféremment $q = (a, \underline{b})$ comme ci-dessus, ou $q = a + \underline{b}$. Pour expliciter les composantes b_i du triplet $\underline{b}$, on note $q = ae_0 + b_1\, e_1 + b_2\, e_2 + b_3\, e_3$.

L'ensemble des quaternions est noté $\mathbb{Q}$ (tant qu'il n'y a pas d'ambiguïté avec l'ensemble des nombres rationnels). Le réel a est appelé la *partie réelle* de q et le vecteur $\underline{b}$ sa *partie imaginaire*. Lorsque $a = 0$, on dit que q est un *quaternion pur* et on l'identifie à un vecteur de $\mathbb{R}^3$. L'ensemble des quaternions purs est noté $\mathbb{Q}_0$. Lorsque $\underline{b} = 0$, on identifie q à un réel. Un quaternion quelconque est identifié à un élément de $\mathbb{R}^4$, et la

base $(e_0, \dots e_3)$ est identifiée à la base canonique de $\mathbb{R}^4$. On définit ensuite les opérations suivantes sur $\mathbb{Q}$.

Définition C.2 (Opérations)**.** On définit comme suit les opérations d'addition, de multiplication, de conjugaison et la norme sur $\mathbb{Q}$. Soient $q = (a, \underline{b})$, $q_1 = (a_1, \underline{b}_1)$ et $q_2 = (a_2, \underline{b}_2)$ dans $\mathbb{Q}$, on pose

$$q_1 + q_2 := (a_1 + a_2,\ \underline{b}_1 + \underline{b}_2) \tag{C.1}$$

$$q_1 q_2 := (a_1 a_2 - \underline{b}_1 \cdot \underline{b}_2,\ a_1 \underline{b}_2 + a_2 \underline{b}_1 + \underline{b}_1 \wedge \underline{b}_2) \tag{C.2}$$

$$q^c := (a, -\underline{b}) \tag{C.3}$$

$$\|q\|^2 := qq^c = q^c q = (a^2 + \|\underline{b}\|^2, \underline{0}). \tag{C.4}$$

On remarque que la définition de l'addition dans (C.1) est cohérente avec la notation $q = a + \underline{b}$ de la définition C.1, avec l'identification du réel a au quaternion $(a, 0)$ et du vecteur $\underline{b}$ au quaternion $(0, \underline{b})$. Cette identification permet aussi de considérer la norme de q comme un réel, et elle coïncide avec la norme euclidienne sur $\mathbb{R}^4$. On note $\mathbb{Q}_1$ l'ensemble des quaternions de norme 1, appelés *quaternions-unité*.

C.1.2 Propriétés

Propriété C.1. On vérifie facilement les propriétés suivantes.

- L'addition est associative et commutative.
- Le quaternion nul $(0, \underline{0})$ est élément neutre pour l'addition.
- La multiplication est associative et non-commutative.
- Le quaternion $1 = (1, \underline{0})$ est élément neutre à droite et à gauche pour la multiplication.
- La multiplication est distributive par rapport à l'addition.
- Pour les éléments de base e_i, les règles de multiplications sont

$$e_1 e_2 = e_3,\ e_2 e_3 = e_1,\ e_3 e_1 = e_2; \quad e_i^2 = -1; \quad e_i e_j = -e_j e_i \quad (i \neq j).$$

- Tout quaternion non-nul possède un unique inverse à gauche et un unique inverse à droite, qui sont égaux. On note cet élément q^{-1} et on a

$$q^{-1} = \frac{(a, -\underline{b})}{qq^c}.$$

- La norme du produit est égale au produit des normes :

$$\|q_1 q_2\| = \|q_1\|\,\|q_2\|. \tag{C.5}$$

- Le conjugué d'un produit et égal au produit commuté des conjugués :

$$(q_1 q_2)^c = q_2^c q_1^c. \tag{C.6}$$

- L'ensemble $\mathbb{Q}$ muni de l'addition et de la multiplication forme un corps non-commutatif.

Les propriétés ci-dessus suffisent pour effectuer tous les calculs élémentaires dont nous aurons besoin dans la suite.

C.1.3 Opérateur de multiplication

La multiplication (à droite ou à gauche) par un quaternion q fixé est une opération linéaire qui apparaît dans les équations d'Euler formulées avec des quaternions (sous-section C.3.2 ci-dessous). On va exprimer sa matrice dans la base canonique de $\mathbb{R}^4$. Soit T_q^g l'opérateur de multiplication à gauche par q, défini par

$$T_q^g : \begin{cases} \mathbb{Q} \longrightarrow \mathbb{Q} \\ v \longmapsto qv \end{cases}$$

et T_q^d l'opérateur de multiplication à droite par q, défini par

$$T_q^d : \begin{cases} \mathbb{Q} \longrightarrow \mathbb{Q} \\ v \longmapsto vq \end{cases} .$$

Ces deux opérateurs sont identifiés à des endomorphismes de $\mathbb{R}^4$. La définition (C.2) de la multiplication donne directement leur matrice relativement à la base canonique de $\mathbb{R}^4$. On trouve

$$\mathrm{Mat}(T_q^g) = \begin{bmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{bmatrix} \tag{C.7}$$

et

$$\mathrm{Mat}(T_q^d) = \begin{bmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & q_3 & -q_2 \\ q_2 & -q_3 & q_0 & q_1 \\ q_3 & q_2 & -q_1 & q_0 \end{bmatrix} . \tag{C.8}$$

Lorsque q est un quaternion unité, les matrices $\mathrm{Mat}(T_q^g)$ et $\mathrm{Mat}(T_q^g)$ sont orthogonales (ce que l'on peut vérifer à la main) ; en effet, les applications T_q^g et T_q^d sont alors des isométries d'après (C.5).

C.2 Réflexions, rotation et conjugaison

Les quaternions purs étant identifiés à des vecteurs de $\mathbb{R}^3$, toute application de l'ensemble des quaternions purs dans lui-même peut être vu comme une application de $\mathbb{R}^3$ dans $\mathbb{R}^3$. Nous allons construire de cette manière les réflexions par rapport à un plan, puis les rotations dans l'espace. On utilisera la formule du double produit vectoriel :

$$a \wedge (b \wedge c) = (a \cdot c)\ b - (a \cdot b)\ c. \tag{C.9}$$

C.2.1 Réflexions

Soit $n = (0, \underline{n}) \in \mathbb{Q}_0$. On définit l'application

$$S_n : \begin{cases} \mathbb{Q}_0 & \longrightarrow \mathbb{Q}_0 \\ (0, \underline{v}) & \longmapsto nvn \end{cases} . \tag{C.10}$$

Un calcul explicite montre que

$$\forall v \in \mathbb{Q}_0,\ S_n(v) = (0, \underline{v}') \text{ avec } \underline{v}' = \underline{v} - 2\,\underline{n}\,(\underline{n} \cdot \underline{v})$$

qui, lorsque $\underline{n}$ est un vecteur de norme 1, est exactement la formule d'une réflexion ; on a donc prouvé le théorème suivant.

Théorème C.3 (Représentation des réflexions vectorielles)**.** *Si n est un quaternion unité pur, l'application S_n définie par (C.10), vue comme endomorphisme de $\mathbb{R}^3$, est la réflexion vectorielle par rapport à l'hyperplan vectoriel orthogonal à $\underline{n}$.*

C.2.2 Rotations

Il est connu que toute rotation de l'espace peut être décomposée sous la forme d'un produit de deux réflexions, comme le rappelle le résultat suivant. Soit $\underline{n}_1$ et $\underline{n}_2$ deux vecteurs unitaires de $\mathbb{R}^3$, et $\theta \in [0, 2\pi]$ l'angle (unique dans $[0, 2\pi]$) tel que $\underline{n}_1 \cdot \underline{n}_2 = \cos(\frac{\theta}{2})$. Soit S_1 la réflexion vectorielle par rapport au plan vectoriel orthogonal à $\underline{n}_1$, de même pour S_2 et $\underline{n}_2$. On va s'intéresser au produit $R := S_2\, S_1$; si $\underline{n}_1$ et $\underline{n}_2$ sont colinéaires, $R = \mathrm{I}_{\mathbb{R}^3}$. Sinon, $\frac{\theta}{2} \neq 0[\pi]$ et on pose $\underline{p} := \frac{1}{\sin(\frac{\theta}{2})}\, \underline{n}_1 \wedge \underline{n}_2$. Ainsi $\underline{p}$ est un vecteur unitaire.

Théorème C.4. *La composée $S_2\, S_1$ est la rotation d'angle θ autour de l'axe (orienté) $\underline{p}$.*

Il faut faire attention à l'ordre des rotations : on applique d'abord S_1, puis S_2. Pour décomposer une rotation donnée de paramètres $(\theta, \underline{p})$ comme produit de réflexions, il suffit de trouver $\underline{n}_1$ et $\underline{n}_2$ (on voit facilement qu'ils existent toujours, et ne sont pas uniques).

On introduit les deux quaternions unité purs $n_1 := (0, \underline{n}_1)$ et $n_2 := (0, \underline{n}_2)$, et leur produit $q := n_2 n_1 = -(\cos(\frac{\theta}{2}), \sin(\frac{\theta}{2})\, \underline{p})$. Grâce au théorème C.3 de représentation des réflexions par les quaternions, on voit donc que la rotation d'angle θ quelconque autour du vecteur unitaire $\underline{p}$ est l'application

$$C_q : \begin{cases} \mathbb{Q}_0 & \longrightarrow \mathbb{Q}_0 \\ v & \longmapsto n_2(n_1 v n_1) n_2 = qvq^{-1} \end{cases}$$

(on a utilisé le fait que $n_i^{-1} = n_i^c = -n_i$ car n_i est un quaternion unité pur). Cette application est appelée la conjugaison par q. On voit facilement que C_q est une isométrie et qu'elle est invariante par homothétie de rapport non-nul sur q :

$$\forall \lambda \in \mathbb{R}^*,\ C_{\lambda q} = C_q, \tag{C.11}$$

en particulier C_q ne dépend pas du signe de q, et de plus on peut se restreindre aux quaternions unités, pour lesquels $q^{-1} = q^c$. Ceci présente l'avantage de remplacer l'inversion q^{-1} par une conjugaison q^c, un peu plus rapide et moins sujette aux erreurs numériques. On aboutit donc au théorème suivant.

Théorème C.5. *Soient $\theta \in \mathbb{R}$ et $\underline{p} \in \mathbb{R}^3$ un vecteur unitaire. La conjugaison par le quaternion unitaire $q = (\cos(\frac{\theta}{2}), \sin(\frac{\theta}{2})\,\underline{p})$ est un endormorphisme de $\mathbb{Q}_0$ identifié à la rotation de $\mathbb{R}^3$ d'angle θ autour du vecteur unitaire $\underline{p}$.*

C.2.3 Conversion entre les représentations

On dispose maintenant de trois manières de représenter une rotation donnée : par les angles d'Euler (annexe B), par une matrice du groupe spécial orthogonal (l'idée la plus naturelle) ou par un quaternion unité. On peut avoir besoin de passer d'une représentation à l'autre, ce qui représente six conversions possibles. Par exemple, lorsqu'on formule les équations du mouvement d'un solide en utilisant l'approche eulérienne et en paramétrant son orientation par un quaternion unité q (sous-section C.3.2 ci-dessous), on obtient des équations différentielles qui donnent l'évolution de q au cours du temps. Pour représenter à l'écran la position du solide, on a parfois besoin de la matrice de rotation Q correspondante (certains codes de visualisation existants l'exigent). On ne donne donc ici que la formule qui nous intéresse, pour passer de q à Q, mais les cinq autres conversions sont possibles aussi avec quelques calculs. On définit d'abord la matrice $S(\underline{p})$ de l'application $x \mapsto \underline{p} \wedge x$, c'est-à-dire

$$S(\underline{p}) = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0 \end{bmatrix}. \tag{C.12}$$

La formule de conversion de q à Q est donnée par le théorème suivant.

Théorème C.6. *Soit q un quaternion unitaire. Si $q = \pm 1$, la conjugaison par q est l'identité. Sinon, on pose $\theta := 2\arccos(q_0) \neq 0[2\pi]$ et $\underline{p} := \frac{1}{\sin(\frac{\theta}{2})}\underline{q}$. La matrice Q de la rotation associée à q par le théorème de représentation C.5 est*

$$\cos(\theta) I_3 + (1 - \cos(\theta))\underline{p}\,\underline{p}^\top + \sin(\theta) S(\underline{p}). \tag{C.13}$$

C.3 Equations d'Euler

Les équations d'Euler déterminent l'évolution du vecteur rotation ω d'un solide S. On voudrait obtenir également l'orientation de S au cours du temps, représentée soit par une matrice de rotation Q (sous-section C.3.2), soit par un quaternion unité (sous-section C.3.1). Pour cela, on établit une équation qui relie ω à $\dot{Q}$ ou $\dot{q}$, ce qui permet de retrouver l'orientation du solide Q ou q par intégration en temps. Dans cette section, les deux approches sont exposées dans le but de montrer que l'approche par les quaternions

n'est pas plus compliquée que celle utilisant des matrices, et qu'elle est plus avantageuse du point de vue numérique.

On considère un solide rigide S de masse m en mouvement dans l'espace, et un point de référence O. On se donne deux bases orthonormales directes $B := (u_1, u_2, u_3)$ et $B' := (U_1, U_2, U_3)$. On suppose que le référentiel (O, u_1, u_2, u_3) est inertiel (galiléen), et que la base B' est liée au solide. On notera $\mathbf{x}$ un vecteur de l'espace, $\underline{x}$ ses coordonnées dans la base B et $\underline{X}$ ses coordonnées dans la base B'.

C.3.1 Formulation matricielle

Les coordonnées de $\mathbf{x}$ dans les bases B et B' sont reliées par l'équation

$$\underline{x} = Q(t)\underline{X} \tag{C.14}$$

où $Q(t)$ est la matrice (orthogonale) de passage de la base "fixe" B à la base "mobile" B'. En dérivant (C.14) par rapport au temps, on obtient

$$\underline{\dot{x}} = \dot{Q}\underline{X} + Q\underline{\dot{X}} = \dot{Q}Q^\top \underline{x} + Q\underline{\dot{X}}. \tag{C.15}$$

En dérivant par rapport au temps la relation $QQ^\top = \mathrm{I}_3$, on voit que $\dot{Q}Q^\top$ est antisymétrique. Il existe donc $\underline{\omega} \in \mathbb{R}^3$ tel que

$$\forall \underline{x} \in \mathbb{R}^3,\ \dot{Q}Q^\top \underline{x} = \underline{\omega} \wedge \underline{x}. \tag{C.16}$$

Le vecteur $\underline{\omega}$ est appelé le vecteur rotation de B' par rapport à B, exprimé dans la base B. Le vecteur de l'espace correspondant ω est parfois noté $\omega_{B'/B}$ pour expliciter les bases B et B'. L'équation (C.15) devient

$$\underline{\dot{x}} = \underline{\omega} \wedge \underline{x} + Q\underline{\dot{X}} \tag{C.17}$$

ce qui s'écrit, dans la base B' (en utilisant l'invariance du produit vectoriel par une transformation orthogonale)

$$\underline{\dot{x}} = Q(\underline{\Omega} \wedge \underline{X} + \underline{\dot{X}}). \tag{C.18}$$

Ces deux formules sont l'expression de la formule de Varignon, exprimée dans B pour (C.17) et dans B' pour (C.18). Elle permet de relier la dérivée temporelle d'un vecteur $\mathbf{x}$ quelconque dans B à sa dérivée temporelle dans B'. En termes de vecteurs de l'espace et non plus de coordonnées, cette formule s'écrit

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{x})_{/B} = \omega_{B'/B} \wedge \mathbf{x} + \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{x})_{/B'}. \tag{C.19}$$

En partant de la formule de Varignon, on déduit la formule de composition des vitesses, puis celle des accélérations, et enfin les équations d'Euler (ces calculs sont faits au chapitre 2). On note $\mathbf{v_G}$ la vitesse du centre de gravité G[1], I la matrice d'inertie du solide S (dans sa configuration actuelle) exprimée dans la base B' et qui est donc constante,

[1] En adaptant les équations, on peut prendre un autre point de référence que G.

$\mathbf{f}_{\text{ext}}$ les forces extérieures et $\mathbf{m}_{\text{ext}}$ les moments des forces extérieures. On aboutit aux équations d'Euler (la première étant écrite dans la base B et la seconde dans la base B')

$$\left\{ \begin{array}{l} m\,\underline{\dot{v}}_G = \sum \underline{f}_{\text{ext}} \\ I\,\underline{\dot{\Omega}} \;\; = (I\underline{\Omega}) \wedge \underline{\Omega} + \sum \underline{M}_{\text{ext}} \end{array} \right. \tag{C.20}$$

qui déterminent l'évolution de $\underline{v}_G$ et $\underline{\Omega}$. Pour connaître en outre la position et l'orientation du solide au cours du temps, il faut également calculer la position $\underline{g}$ du centre de gravité G et la matrice de rotation Q qui permet de passer de la base fixe B à la base mobile B'. Soit $S(\underline{\Omega})$ la matrice de l'application $X \mapsto \Omega \wedge X$ explicitée par l'équation (C.12) ; l'équation (C.16) s'écrit alors $\dot{Q}Q^\top x = QS(\underline{\Omega})Q^\top x$ ou encore $\dot{Q} = QS(\underline{\Omega})$. On peut donc remonter de l'histoire de ($\underline{v}_G$, $\underline{\Omega}$) à la position et l'orientation ($\underline{g}$, Q) du solide grâce aux équations différentielles

$$\left\{ \begin{array}{l} \underline{\dot{g}} \;\; = \underline{v}_G, \\ \dot{Q} = QS(\underline{\Omega}). \end{array} \right. \tag{C.21}$$

Cependant, cette technique consistant à résoudre numériquement l'équation $\dot{Q} = QS(\underline{\Omega})$ est assez maladroite, pour deux raisons.

- D'une part, on intègre une équation différentielle de dimension 9, alors que le paramétrage de l'orientation de B' par rapport à B ne nécessite que 3 degrés de liberté. On fait donc un calcul "trois fois trop coûteux".
- D'autre part, l'intégration numérique étant inexacte, la matrice Q (qui devrait théoriquement rester orthogonale) va dériver et s'éloigner de l'orthogonalité, de sorte qu'elle représentera de moins en moins une matrice de passage ; ceci nécessite d'utiliser des techniques de stabilisation qui compliquent l'intégration en temps.

On pourrait lutter contre le premier problème en écrivant seulement trois équations différentielles sur les trois angles d'Euler, mais nous les avons exclus en raison des problèmes de singularité artificielle qu'ils introduisent. L'utilisation des quaternions permet de trouver un compromis intéressant entre les angles d'Euler et la matrice Q :

- la paramétrage de l'orientation par un quaternion unité q ne nécessite qu'une équation différentielle de dimension 4 (ce qui est mieux que la dimension 9 de Q, mais moins bien que la dimension 3 des angles d'Euler) ;
- le seul phénomène de dérive possible est que la norme de q s'éloigne de 1. La technique de stabilisation est immédiate (il suffit de diviser q par sa norme de temps en temps).

Le paramétrage par un quaternion unité permet donc d'éviter à la fois les phénomènes de dérive qui pénalisent le paramétrage par Q et la singularité introduite par les angles d'Euler, au seul prix de l'introduction d'une variable supplémentaire (dimension 4 au lieu de 3 avec les angles d'Euler). Cette approche est exposée dans la sous-section suivante.

C.3.2 Formulation en terme de quaternions

Soient x et X les quaternions purs associés aux triplets $\underline{x}$ et $\underline{X}$, et q le quaternion unité associé à la matrice de rotation Q, de sorte que

$$x = qXq^c. \tag{C.22}$$

En dérivant cette équation par rapport au temps (de la même manière que nous avions dérivé l'équation $\underline{x} = Q\underline{X}$ dans la section précédente), on obtient

$$\dot{x} = \dot{q}Xq^c + qX\dot{q}^c + q\dot{X}q^c = (\dot{q}q^c)x + x(\dot{q}q^c)^c + q\dot{X}q^c. \tag{C.23}$$

En dérivant par rapport au temps l'équation $\|q\|^2 = 1$, on obtient le lemme suivant (qui est à comparer au fait que la matrice $\dot{Q}Q^\top$ est anti-symétrique, obtenu en dérivant l'équation $QQ^\top = \mathrm{I}_3$).

Lemme C.7. *Le quaternion $\dot{q}q^c$ est un quaternion pur.*

On introduit donc le quaternion pur $v = \dot{q}q^c = (0, \underline{v})$ et l'équation (C.23) se récrit

$$\dot{x} - q\dot{X}q^c = vx - xv = (0, 2\underline{v} \wedge \underline{x}). \tag{C.24}$$

On reconnaît la formule de Varignon (C.17) exprimée dans la base B, ce qui permet d'identifier le quaternion pur $2v$ au vecteur rotation $\underline{\omega}$. On peut donc remonter de l'histoire de $\underline{\omega}$ à celle du quaternion q grâce à l'équation différentielle de dimension 4

$$\dot{q} = \frac{1}{2}\omega q = \frac{1}{2}q\Omega = \frac{1}{2}\mathrm{Mat}(T^d_\Omega)q \tag{C.25}$$

où la matrice $\mathrm{Mat}(T^d_\Omega)$ de l'application "multiplication à droite par Ω" est définie à l'équation (C.8).

Remarque C.8. En pratique, même si on fait tous les calculs en représentant l'orientation par un quaternion, il faut parfois repasser en représentation matricielle pour dessiner à l'écran le mouvement du solide (par exemple, le langage OpenGL représente les rotations par des matrices). On peut alors utiliser la formule (C.13) du théorème C.6 pour reconstruire la matrice Q.

Conclusion

Lorsque les configurations singulières introduites par les angles d'Euler ne peuvent pas être évitées, comme c'est le cas en simulation de matériaux granulaires, l'introduction d'un quaternion unité permet d'éviter les singularités au prix d'une variable et d'une contrainte supplémentaire. L'utilisation des quaternions n'est pas plus compliquée que celle des angles d'Euler et le coût supplémentaire dû à l'augmentation de la dimension du problème est compensé par la sérénité qu'apporte la certitude de ne pas faire empirer artificiellement le conditionnement de la matrice de masse.

Annexe D

Différentiation des fonctions auxiliaires

Cette annexe rassemble les calculs, un peu pénibles, des différentielles de la fonction d'Alart et Curnier et de celle de De Saxcé.

D.1 Formulation d'Alart et Curnier

D.1.1 Partie normale

On rappelle que la fonction f_{N} (partie normale de la fonction d'Alart et Curnier, voir l'équation (1.23)) est définie par

$$f_{\mathrm{N}} : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ (u, r) \longmapsto \mathrm{P}_{\mathbb{R}^+}(P_{\mathrm{N}}\, r - \rho_{\mathrm{N}}\, P_{\mathrm{N}}\, u) - P_{\mathrm{N}}\, r \end{cases}$$

où ρ_{N} et ρ_{T} sont des constantes strictement positives (disons, égales à 1) et on rappelle que $P_{\mathrm{N}} \in \mathbb{R}^{1\times d}$ et $P_{\mathrm{T}} \in \mathbb{R}^{d-1\times d}$ sont les matrices de projection sur la direction normale et le plan tangent, définies par (1.19).

D'après sa définition, f_{N} est continue et affine par morceaux. Plus précisément, si $P_{\mathrm{N}}\, r - \rho_{\mathrm{N}}\, P_{\mathrm{N}}\, u < 0$, alors

$$f_{\mathrm{N}}(u, r) = -P_{\mathrm{N}}\, r,\ \frac{\partial f_{\mathrm{N}}}{\partial u} = 0_{1\times d},\ \frac{\partial f_{\mathrm{N}}}{\partial r} = -P_{\mathrm{N}}$$

et sinon,

$$f_{\mathrm{N}}(u, r) = -\rho_{\mathrm{N}}\, P_{\mathrm{N}}\, u,\ \frac{\partial f_{\mathrm{N}}}{\partial u} = -\rho_{\mathrm{N}} P_{\mathrm{N}},\ \frac{\partial f_{\mathrm{N}}}{\partial r} = 0_{1\times d}.$$

D.1.2 Partie tangentielle

Introduisons maintenant la fonction auxiliaire g par

$$g : \begin{cases} \mathbb{R} \times \mathbb{R}^{d-1} \longrightarrow \mathbb{R}^{d-1} \\ (\lambda, y) \longmapsto \mathrm{P}_{B(0,\lambda)}(y) \end{cases}$$

avec par convention,

$$\forall y \in \mathbb{R}^{d-1},\ \mathrm{P}_{\emptyset}(y) = 0_{d-1}.$$

On distingue trois cas, et après quelques calculs on trouve les formules suivantes pour la jacobienne de g.

– Si $\|y\| < \lambda$, alors

$$\begin{cases} g(\lambda, y) & = & y, \\ \dfrac{\partial g}{\partial \lambda}(\lambda, y) & = & 0_{d-1\times 1}, \\ \dfrac{\partial g}{\partial y}(\lambda, y) & = & \mathrm{I}_{d-1}. \end{cases}$$

– Si $\|y\| > \lambda > 0$, alors

$$\begin{cases} g(\lambda, y) & = & \lambda \frac{y}{\|y\|}, \\ \dfrac{\partial g}{\partial \lambda}(\lambda, y) & = & \frac{y}{\|y\|}, \\ \dfrac{\partial g}{\partial y}(\lambda, y) & = & \frac{\lambda}{\|y\|}(\mathrm{I}_{d-1} - \frac{yy^{\top}}{\|y\|^2}). \end{cases}$$

– Si $\|y\| > 0 > \lambda$, alors

$$\begin{cases} g(\lambda, y) & = & 0_{d-1\times 1}, \\ \dfrac{\partial g}{\partial \lambda}(\lambda, y) & = & 0_{d-1\times 1}, \\ \dfrac{\partial g}{\partial y}(\lambda, y) & = & 0_{d-1\times d-1}. \end{cases}$$

Dans les cas limites ($\|y\| = \lambda$, $\|y\| = 0$ et $\lambda = 0$), la fonction g n'est a priori pas différentiable.

On rappelle maintenant que la partie tangentielle de la fonction d'Alart-Curnier (1.24) est définie par

$$f_{\mathrm{T}} : \begin{cases} \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^{d-1} \\ (u, r) \longmapsto \mathrm{P}_{B(0,\mu P_{\mathrm{N}} r)}(P_{\mathrm{T}} r - \rho_{\mathrm{T}} P_{\mathrm{T}} u) - P_{\mathrm{T}} r \end{cases}$$

c'est-à-dire

$$f_{\mathrm{T}}(u, r) = g(\underbrace{\mu P_{\mathrm{N}} r}_{=:\lambda}, \underbrace{P_{\mathrm{T}} r - \rho_{\mathrm{T}} P_{\mathrm{T}} u}_{=:y}) - P_{\mathrm{T}} r.$$

En dérivant f_{T} avec la formule de dérivation composée (on néglige l'indice « T » pour alléger les notations)

$$\begin{cases} \dfrac{\partial f}{\partial u} = \dfrac{\partial g}{\partial \lambda}\dfrac{\partial \lambda}{\partial u} + \dfrac{\partial g}{\partial y}\dfrac{\partial y}{\partial u} \\ \dfrac{\partial f}{\partial r} = \dfrac{\partial g}{\partial \lambda}\dfrac{\partial \lambda}{\partial r} + \dfrac{\partial g}{\partial y}\dfrac{\partial y}{\partial r} - P_{\mathrm{T}} \end{cases}$$

on obtient, dans les trois cas, les formules suivantes.

– Si $\|y\| < \lambda$, alors

$$\left\{\begin{array}{l} f(u,r) \quad = -\rho_{\mathrm{T}} P_{\mathrm{T}}\, u, \\ \dfrac{\partial f}{\partial u}(u,r) = -\rho_{\mathrm{T}} P_{\mathrm{T}}, \\ \dfrac{\partial f}{\partial r}(u,r) = 0_{d-1\times d}. \end{array}\right.$$

– Si $\|y\| > \lambda > 0$, alors

$$\left\{\begin{array}{l} f(u,r) \quad = \lambda \frac{y}{\|y\|} - P_{\mathrm{T}}\, r, \\ \dfrac{\partial f}{\partial u}(u,r) = -\rho_{\mathrm{T}} \frac{\lambda}{\|y\|}(\mathrm{I}_{d-1} - \frac{yy^{\top}}{\|y\|^2}) P_{\mathrm{T}}, \\ \dfrac{\partial f}{\partial r}(u,r) = \mu \frac{y}{\|y\|} P_{\mathrm{N}} + \frac{\lambda}{\|y\|}(\mathrm{I}_{d-1} - \frac{yy^{\top}}{\|y\|^2}) P_{\mathrm{T}} - P_{\mathrm{T}}. \end{array}\right.$$

– Si $\|y\| \geq 0 \geq \lambda$, alors

$$\left\{\begin{array}{l} f(u,r) \quad = -P_{\mathrm{T}}\, r, \\ \dfrac{\partial f}{\partial u}(u,r) = 0_{d-1\times d}, \\ \dfrac{\partial f}{\partial r}(u,r) = -P_{\mathrm{T}}. \end{array}\right.$$

D.1.3 Particularisation à la dimension 2

Les formules des sous-sections précédentes se simplifient en dimension $d = 2$, dans le cas où $\|y\| > \lambda > 0$. En effet, on a alors

$$\frac{y}{\|y\|} = \mathrm{sign}(y),\ \mathrm{I}_{d-1} = \mathrm{I}_1 = 1,\ \frac{yy^{\top}}{\|y\|^2} = 1.$$

Donc : si $\|y\| > \lambda > 0$, alors

$$\left\{\begin{array}{l} f(u,r) \quad = \mu\, P_{\mathrm{N}}\, r\, \mathrm{sign}(y) - P_{\mathrm{T}}\, r, \\ \dfrac{\partial f}{\partial u}(u,r) = 0, \\ \dfrac{\partial f}{\partial r}(u,r) = \mu\, P_{\mathrm{N}}\, \mathrm{sign}(y) - P_{\mathrm{T}}. \end{array}\right.$$

D.1.4 Un résultat négatif

Cette sous-section donne un exemple de situation très générale dans laquelle la méthode de Newton appliquée à la fonction d'Alart et Curnier n'a pas de sens, faute de pouvoir inverser la matrice jacobienne. En admettant que la fonction d'Alart et Curnier f_{AC} est différentiable au point $(Wr + q, r)$, la matrice jacobienne de l'application $r \to f_{AC}(Wr + q, r)$ est

$$\frac{\partial f_{AC}}{\partial u}(Wr + q, r)\, W + \frac{\partial f_{AC}}{\partial r}(Wr + q, r).$$

La méthode de Newton n'a de sens que si cette matrice est inversible à chaque itération. Or, ceci n'est pas garanti, comme le montre l'exemple suivant. D'une part, la matrice

$$\frac{\partial f_{AC}}{\partial r}(Wr + q, r)$$

est la matrice nulle lorsque r vérifie à la fois

$$r_{\mathrm{N}} - \rho_{\mathrm{N}} u_{\mathrm{N}} \geq 0 \quad \text{et} \quad \|y\| > \lambda > 0.$$

D'autre part, W n'est pas inversible en général, comme le montre l'exemple suivant.

Exemple D.1. Si le système mécanique comporte deux contacts qui impliquent la même paire de solides rigides, alors la matrice W est automatiquement singulière. En effet, il existe des vitesses relatives cinématiquement interdites (celles qui ne satisfont pas l'hypothèse de rigidité) et l'image de W (qui donne les vitesses relatives u par la formule $u = Wr + q$) ne peut pas être égale à $\mathbb{R}^{nd}$ tout entier : W n'est donc pas inversible.

En utilisant l'exemple D.1, il est donc facile de construire des systèmes mécaniques très simples pour lesquels la méthode d'Alart et Curnier n'a pas de sens.

Exemple D.2. Considérons un carré rigide, en dimension $d = 2$, posé sur un plan immobile comme sur la figure D.1. On prend comme paramètres la position (x, y) de

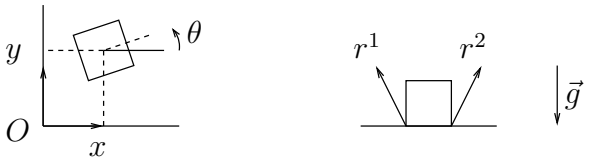


FIG. D.1 – Un exemple très simple où W est singulière

son centre et son orientation θ. On déclare deux points de contact au niveau des deux sommets du carré qui touchent le sol, avec les coefficients de frottement $\mu^1 = \mu^2 = 1$. On suppose que la matrice de masse est $M = \mathrm{I}_3$, que la gravité (dirigée vers le bas) est $g = 1$, et on discrétise avec un pas de temps $h = 1$. Les données du problème incrémental sont alors

$$d = 2, m = 3, n = 2, \mu = (1, 1), M = \mathrm{I}_3, w = 0_{3\times 1}, f = [0, 1, 0]^\top$$

et

$$E = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

On remarque que le noyau de $H^\top$ (qui est aussi celui de W) est la droite générée par $(1, 0, -1, 0)$, conformément à l'intuition : si l'on applique deux forces opposées et horizontales $r^1 = (\alpha, 0)$ et $r^2 = (-\alpha, 0)$ aux deux coins du carré rigide, il n'en résulte

aucun mouvement. Initialisons maintenant l'algorithme avec la valeur $r_0 := (1, 6, 9, 8)$. On constate que cette initialisation produit bien

$$\frac{\partial f_{AC}}{\partial r}(Wr + q, r) = 0$$

et donc que la matrice

$$\frac{\partial f_{AC}}{\partial u} W + \frac{\partial f_{AC}}{\partial r} = \frac{\partial f_{AC}}{\partial u} W$$

n'est pas inversible. L'algorithme de Newton échoue à la première itération. Pourtant, ce problème admet une solution évidente (il suffit de prendre deux forces de contact égales qui compensent la gravité). Cette solution est d'ailleurs trouvée en quelques itérations à partir d'une autre initialisation (l'origine, par exemple).

D.2 Formulation de De Saxcé

On calcule la matrice jacobienne de f_{DS} définie par (4.9) là où ceci a un sens, c'est-à-dire lorsque $\widetilde{u}$ est différentiable en u et P_L est différentiable en $r - \rho\,\widetilde{u}(r)$. La règle de dérivation composée donne pour le i-ème bloc de cette matrice

$$\frac{\partial f^i_{DS}}{\partial r} = \frac{\partial \mathrm{P}_{K^i}}{\partial r^i}\left(\frac{\partial r^i}{\partial r} - \rho^i \frac{\partial \widetilde{u}^i}{\partial u^i} W^i\right) \tag{D.1}$$

où la jacobienne

$$\frac{\partial \mathrm{P}_{K^i}}{\partial x}$$

(de taille $d \times d$) de la projection sur le cône du second ordre est donnée dans l'annexe E,

$$\frac{\partial r^i}{\partial r} = [0_{d,(i-1)d}, \mathrm{I}_d, 0_{d,(n-i)d}],$$

et pour $u^i_{\mathrm{T}} \neq 0$

$$\frac{\partial \widetilde{u}^i}{\partial u^i} = \mathrm{I}_d + \frac{\mu^i}{\|u^i_{\mathrm{T}}\|} e^i_{\mathrm{N}} (P_{\mathrm{T}}\, u^i)^\top P_{\mathrm{T}}. \tag{D.2}$$

On a aussi noté W^i (de taille $d \times nd$) le i-ème bloc diagonal de W. Ces formules permettent d'implémenter le calcul de f_{DS} de l'équation (4.9) et sa matrice jacobienne, quand elle existe.

Annexe E

Projection sur le cône du second ordre

L'opérateur de projection sur le cône du second ordre apparaît naturellement dans l'étude du frottement de Coulomb ; elle apparaît par exemple dans la formulation de De Saxcé avec projection (sous-section 4.2.2). Nous allons calculer une formule explicite par morceaux pour la projection, et calculer sa matrice jacobienne sur chacun des morceaux.

E.1 Formule de projection

Soit d un entier naturel supérieur ou égal à 2, et K le cône du second ordre d'ouverture $\mu \in [0, \infty]$ dans $\mathbb{R}^d$ défini par

$$K := \{x_1, \dots, x_d \,:\, \|(x_1, \dots, x_{d-1})\| \leq \mu\, x_d\}. \tag{E.1}$$

Dans la suite, on notera $x_{\mathrm{T}} := (x_1, \dots, x_{d-1})$ et $x_{\mathrm{N}} := x_d$ par analogie avec les composantes tangentielles et normales usuelles en mécanique. Soit $x \in \mathbb{R}^d$ fixé. On s'intéresse au problème de minimisation

$$\left\{ \begin{array}{l} \min f(r) := \frac{1}{2}\|x - r\|^2 \\ r \in K \iff g(r) := \frac{1}{2}\sum_{i=1}^{d-1}(r_i^2) - \mu^2 r_d^2 \leq 0. \end{array} \right. \tag{E.2}$$

Il est connu que ce problème possède une unique solution r^*, appelée projection de x sur le convexe K. D'une part, on a

$$x \in K \iff r^* = x$$

et d'autre part

$$x \in K^\circ = -K_{\frac{1}{\mu}} \iff r^* = 0$$

où K° est le cône polaire de K. Supposons maintenant que $x \notin K \cup K^\circ$, de sorte que r^* appartient à la frontière de K mais n'est pas nul. Alors

$$\nabla g(r) = \begin{bmatrix} r_{\mathrm{T}} \\ -\mu^2\, r_{\mathrm{N}} \end{bmatrix}$$

donc le vecteur $n := (\frac{r_{\mathrm{T}}}{\mu}, -\mu r_{\mathrm{N}})$ est un vecteur normal sortant à K en r. Le problème d'optimisation (E.2) est équivalent aux conditions d'optimalité suivantes

$$\left\{ \begin{array}{ll} r_{\mathrm{T}} - x_{\mathrm{T}} = -\frac{\lambda}{\mu} r_{\mathrm{T}} \\ r_{\mathrm{N}} - x_{\mathrm{N}} = \lambda\mu r_{\mathrm{N}} \\ \|r_{\mathrm{T}}\| \quad\quad = \mu r_{\mathrm{N}}. \end{array} \right. \tag{E.3}$$

D'après la seconde équation, $\lambda\mu = 1 - \frac{x_{\mathrm{N}}}{r_{\mathrm{N}}}$ que l'on injecte dans la première ($r_{\mathrm{T}}(\mu^2 + \lambda\mu) = \mu^2 x_{\mathrm{T}}$) pour obtenir $r_{\mathrm{T}}(1 + \mu^2 - \frac{x_{\mathrm{N}}}{r_{\mathrm{N}}}) = \mu^2 x_{\mathrm{T}}$. On remplace dans cette équation r_{N} par sa valeur $\frac{\|r_{\mathrm{T}}\|}{\mu}$ donnée par la troisième équation de (E.3), pour obtenir $r_{\mathrm{T}}(1 + \mu^2 - \frac{\mu x_{\mathrm{N}}}{\|r_{\mathrm{T}}\|}) = \mu^2 x_{\mathrm{T}}$. On prend ensuite la norme de cette équation, ce qui donne $\|r_{\mathrm{T}}\|(1 + \mu^2 - \frac{\mu x_{\mathrm{N}}}{\|r_{\mathrm{T}}\|}) = \mu^2 \|x_{\mathrm{T}}\|$ ou encore après quelques calculs $r_{\mathrm{T}} = \frac{\mu}{1+\mu^2}(x_{\mathrm{N}} + \mu\|x_{\mathrm{T}}\|)$. De cette dernière équation, et des formules $r_{\mathrm{N}} = \frac{\|r_{\mathrm{T}}\|}{\mu}$ et $r_{\mathrm{T}} = \|r_{\mathrm{T}}\| \frac{x_{\mathrm{T}}}{\|x_{\mathrm{T}}\|}$, on tire

$$r_{\mathrm{T}} = \frac{\mu}{1+\mu^2}(\mu + \frac{x_{\mathrm{N}}}{\|x_{\mathrm{T}}\|}) x_{\mathrm{T}}. \tag{E.4}$$

$$r_{\mathrm{N}} = \frac{1}{1+\mu^2}(x_{\mathrm{N}} + \mu\|x_{\mathrm{T}}\|) \tag{E.5}$$

Ces formules permettent de calculer explicitement la projection d'un point x sur le cône K. Il suffit de vérifier d'abord que l'on n'est pas dans l'un des deux cas faciles ($x \in K$ et $x \in K^\circ$), puis de calculer r_{N} et r_{T} avec (E.5) et (E.4).

E.2 Différentielle de la projection

On s'intéresse maintenant à la différentielle de l'application

$$P_K : \left\{ \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ x & \longmapsto & r^* \end{array} \right. \tag{E.6}$$

où r^* est l'unique solution de (E.2). Si $x \in K$, $r^* = x$, donc sur l'ouvert $\mathrm{int}(K)$, l'opérateur P_K est l'identité et par conséquent sa différentielle est également l'identité. De même, sur $\mathrm{int}(K^\circ)$, $r^* = 0$ et la différentielle de P_K est l'application nulle. Enfin, sur $\mathrm{int}(\mathbb{R}^d \setminus (K \cup K^\circ))$, on dispose des formules (E.5) et (E.5) que l'on peut différentier. Posons

$$f(x) := \begin{bmatrix} x_{\mathrm{T}} \\ \frac{\|x_{\mathrm{T}}\|}{\mu} \end{bmatrix} \text{ et } \alpha(x) = \frac{\mu}{1+\mu^2}(\mu + \frac{x_{\mathrm{N}}}{\|x_{\mathrm{T}}\|}) \tag{E.7}$$

de sorte que $P_K(x) = \alpha(x)\, f(x)$. On a donc

$$\mathrm{Jac}[P_K](x) = \alpha(x)\, \mathrm{Jac}[f](x) + f(x)\, \nabla\alpha(x)^\top. \tag{E.8}$$

On trouve facilement

$$\mathrm{Jac}[f](x) = \begin{bmatrix} I_{d-1} & 0_{d-1,1} \\ \frac{x_{\mathrm{T}}^\top}{\mu\|x_{\mathrm{T}}\|} & 0 \end{bmatrix} \tag{E.9}$$

et

$$\nabla\alpha(x) = \frac{\mu}{1+\mu^2}\begin{bmatrix} -\frac{x_{\mathrm{N}}}{\|x_{\mathrm{T}}\|^3}x_{\mathrm{T}} \\ \frac{1}{\|x_{\mathrm{T}}\|} \end{bmatrix} \tag{E.10}$$

ce qui donne

$$\mathrm{Jac}[P_K](x) = \begin{bmatrix} \alpha(x)I_{d-1} - \frac{\mu}{1+\mu^2}\frac{x_{\mathrm{N}}}{\|x_{\mathrm{T}}\|^3}(x_{\mathrm{T}}\, x_{\mathrm{T}}^\top) & \frac{\mu}{1+\mu^2}\frac{x_{\mathrm{T}}}{\|x_{\mathrm{T}}\|} \\ (\frac{\alpha(x)}{\mu\|x_{\mathrm{T}}\|} - \frac{x_{\mathrm{N}}}{(1+\mu^2)\|x_{\mathrm{T}}\|^2})x_{\mathrm{T}}^\top & \frac{1}{1+\mu^2} \end{bmatrix}. \tag{E.11}$$

Table des figures

Liste des tableaux

Index

Bibliographie

[AB08] V. Acary et B. Brogliato : *Numerical methods for nonsmooth dynamical systems*. Springer, 2008.

[AC91] P. Alart et A. Curnier : A mixed formulation for frictional contact problems prone to Newton like solution methods. *Comput. Methods Appl. Mech. Eng.*, 92(3):353–375, 1991.

[ACLM09] V. Acary, F. Cadoux, C. Lemaréchal et J. Malick : An optimisation-based approach to the discrete Coulomb friction problem. Article en préparation, 2009.

[AFMS08] R. Andreani, A. Friedlander, M. P. Mello et S. A. Santos : Box-constrained minimization reformulations of complementarity problems in second-order cones. *Jour. Nonlin. Conv. Anal.*, 2008.

[AG03] F. Alizadeh et D. Goldfarb : Second-order cone programming. *Math. Prog. Ser. B*, 95:3–51, 2003.

[Ala97] P. Alart : Méthode de Newton généralisée en mécanique du contact. *Journal de Mathématiques Pures and Appliquées*, 76(1):83, 1997.

[AP97] M. Anitescu et F.A. Potra : Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dyn.*, 14(3):231–247, 1997.

[BAC+06] F. Bertails, B. Audoly, M.P. Cani, B. Querleux, F. Leroy et J.L. Lévêque : Super-helices for predicting the dynamics of natural hair. *In International Conference on Computer Graphics and Interactive Techniques*, pages 1180–1187. ACM New York, NY, USA, 2006.

[Bal71] E. Balas : Intersection cuts – a new type of cutting planes for integer programming. *Operations Research*, 19(3):19–39, 1971.

[Bal79] E. Balas : Disjunctive programming. *In* P. L. Hammer, E. L. Johnson et B. H. Korte, éditeurs : *Discrete Optimization II, 5*, pages 3–51, Amsterdam, 1979. Annals of Discrete Mathematics, North-Holland.

[Bar89] D. Baraff : Analytical methods for dynamic simulation of non-penetrating rigid bodies. *In* Jeffrey Lane, éditeur : *Computer Graphics (SIGGRAPH '89 Proceedings)*, volume 23, pages 223–232, juillet 1989.

[Bar91] D. Baraff : Coping with friction for non-penetrating rigid body simulation. *In* Thomas W. Sederberg, éditeur : *Computer Graphics (SIGGRAPH '91 Proceedings)*, volume 25, pages 31–40, juillet 1991.

[Bar96] D. Baraff : Linear-time dynamics using Lagrange multipliers. *In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 137–146. ACM New York, NY, USA, 1996.

[BCC93] E. Balas, S. Ceria et G. Cornuéjols : A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math. Prog.*, 58(3):295–324, 1993.

[Ber06] F. Bertails : *Simulation de Chevelures Virtuelles*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 2006.

[BJ80] E. Balas et R.G. Jeroslow : Strengthening cuts for mixed integer programs. *European Journal of Operational Research*, 4(4):224–234, 1980.

[BL08] O. Bauchau et A. Laulusa : Review of contemporary approaches for constraint enforcement in multibody systems. *Journal of Computational and Nonlinear Dynamics*, 3(1):011005, 2008.

[Boy93] E. A. Boyd : Generating Fenchel cutting planes for knapsack polyhedra. *SIAM Journal of Optimization*, 3:734–750, 1993.

[Boy94] E. A. Boyd : Fenchel cutting planes for integer programs. *Operations Research*, 42:53–64, 1994.

[Boy95] E. A. Boyd : On the convergence of Fenchel cutting planes in mixed-integer programming. *SIAM Journal of Optimization*, 5:421–435, 1995.

[Bro99] B. Brogliato : *Nonsmooth mechanics. Models, dynamics and control. 2nd ed.* Communications and Control Engineering Series. London : Springer, 552 p., 1999.

[BTN01] A. Ben-Tal et A. Nemirovski : *Lectures on modern convex optimization. Analysis, algorithms, and engineering applications.* MPS/ SIAM Series on Optimization, 488 p., 2001.

[Cad08] F. Cadoux : Computing deep facet-defining disjunctive cuts for mixed-integer programming. *Math. Prog. A*, juin 2008. A paraître.

[CKPS98] P.W. Christensen, A. Klarbring, J.S. Pang et N. Stromberg : Formulation and comparison of algorithms for frictional contact problems. *Int. J. Numer. Methods Eng.*, 42(1):145–173, 1998.

[CL06] G. Cornuéjols et C. Lemaréchal : A convex-analysis perspective on disjunctive cuts. *Math. Prog.*, 106:567 – 586, 2006.

[Cor08] G. Cornuéjols : Valid inequalities for mixed integer linear programs. *Math. Prog.*, 112:3–44, 2008.

[CPS92] R. W. Cottle, J. Pang et R. E. Stone : *The linear complementarity problem*. Academic Press, Inc., Boston, MA, 1992.

[CR99] E. A. Coutsias et L. Romero : The quaternions with an application to rigid body dynamics. working paper, Univ. of New Mexico, 1999.

[CS97] S. Ceria et J. Soares : Disjunctive cut generation for mixed 0–1 programs : duality and lifting, 1997.

[CT05] J.-S. Chen et P. Tseng : An unconstrained smooth minimization reformulation of the second-order cone complementarity problem. *Math. Program*, 104(2-3):293–327, 2005.

[DF95] S. Dirkse et M. Ferris : The PATH solver : A non-monotone stabilization scheme for mixed complementarity problems. *Opt. meth. and Soft.*, 5:123–156, 1995.

[DSF98] G. De Saxcé et Z. Q. Feng : The bipotential method : a constructive approach to design the complete contact law with friction and improved numerical algorithms. *Math. Comput. Modelling*, 28(4-8):225–245, 1998.

[FJ00] A. Fischer et H. Jiang : Merit functions for complementarity and related problems : A survey. *Computational Optimization and Applications*, 17(2): 159–182, 2000.

[FK98] M. Ferris et C. Kanzow : Complementarity and related problems : A survey. Technical Report MP-TR-1998-17, University of Wisconsin, Madison, novembre 1998.

[Fle87] R. Fletcher : *Practical Methods of Optimization*. Wiley, 1987.

[FLT02] M. Fukushima, Z.-Q. Luo et P. Tseng : Smoothing functions for second-order-cone complementarity problems. *SIAM J. Opt.*, 12(2):436–460, 2002.

[FM97] M. Ferris et T. Munson : Interfaces to PATH 3.0 : Design, implementation and usage. Technical Report MP-TR-1997-12, University of Wisconsin, Madison, novembre 1997.

[FWA07] M.C. Ferris, A.J. Wathen et P. Armand : Limited memory solution of bound constrained convex quadratic problems arising in video games. *RAIRO Operations Research*, 41(1):19–34, 2007.

[Has83] J. Haslinger : Approximation of the Signorini problem with friction, obeying the Coulomb law. *Math. Methods Appl. Sci*, 5:422–437, 1983.

[HKD04] J. Haslinger, R. Kucera et Z. Dostál : An algorithm for the numerical realization of 3d contact problems with Coulomb friction. *Journal of Computational and Applied Mathematics*, 164-165:387 – 408, 2004. Proceedings of the 10th International Congress on Computational and Applied Mathematics.

[HT83] J. Haslinger et M. Tvrdỳ : Approximation and numerical solution of contact problems with friction. *Applications of Mathematics*, 28(1):55–71, 1983.

[HUL93] J.-B. Hiriart-Urruty et C. Lemarechal : Convex Analysis and Minimization Algorithms, Vol. 1 and 2, 1993.

[HUL01] J.-B. Hiriart-Urruty et C. Lemaréchal : *Fundamentals of Convex Analysis*. Springer Verlag, Heidelberg, 2001.

[HYF05] S. Hayashi, N. Yamashita et M. Fukushima : A combined smoothing and regularization method for monotone second-order cone complementarity problems. *SIAM Journal on Optimization*, 15(2):593–615, 2005.

[JAJ98] F. Jourdan, P. Alart et M. Jean : A Gauss-Seidel like algorithm to solve frictional contact problems. *Comput. Methods Appl. Mech. Eng.*, 155(1-2):31–47, 1998.

[JF08] P. Joli et Z.Q. Feng : Uzawa and Newton algorithms to solve frictional contact problems within the bi-potential framework. *International Journal for Numerical Methods in Engineering*, 73(3), 2008.

[JFMR02] J. Júdice, A. Faustino et I. Martins-Ribeiro : On the solution of NP-hard linear complementarity problems. *TOP*, 10(1):125–145, June 2002.

[KC05] M. Karamanov et G. Cornuéjols : Branching on general disjunctions. Working paper, 2005.

[KEP05] D. Kaufman, T. Edmunds et D.K. Pai : Fast frictional dynamics for rigid bodies. *ACM Trans. Graph.*, 24(3):946–956, 2005.

[KMPdC06] Y. Kanno, J. Martins et A. Pinto da Costa : Three-dimensional quasi-static frictional contact by using second-order cone linear complementarity problem. *Int. Jour. for Num. Meth. Eng.*, 65(1), 2006.

[KP98] A. Klarbring et J.-S. Pang : Existence of solutions to discrete semicoercive frictional contact problems. *SIAM Journal on Optimization*, 8(2):414–442, 1998.

[Kuc07] R. Kucera : Minimizing quadratic functions with separable quadratic constraints. *Opt. Meth. Soft.*, 22, issue 3:453–467, 2007.

[LB08] A. Laulusa et O. Bauchau : Review of classical approaches for constraint enforcement in multibody systems. *Journal of Computational and Nonlinear Dynamics*, 3(1):011004, 2008.

[LH64] C.E. Lemke et J.T. Howson : Equilibrium points of bimatrix games. *SIAM Journal on Applied Mathematics*, 12:413–423, 1964.

[Löt84] P. Lötstedt : Numerical simulation of time-dependent contact and friction problems in rigid body mechanics. *SIAM Journal on Scientific and Statistical Computing*, 5:370, 1984.

[Mal09] J. Malick : Dual Newton methods to solve second-order cone least-squares. working paper, INRIA, 2009.

[MC95] B. Mirtich et J. Canny : Impulse-based simulation of rigid bodies. *In Proceedings of the 1995 symposium on Interactive 3D graphics*. ACM New York, NY, USA, 1995.

[Mor88] J.-J. Moreau : Unilateral contact and dry friction in finite freedom dynamics. Nonsmooth Mech. App., CISM Courses Lect. 302, 1-82, 1988.

[Mor94] J.-J. Moreau : Some numerical methods in multibody dynamics : application to granular materials. *Eur. J. Mech. A*, 13:93, 1994.

[Mor99] J.-J. Moreau : Numerical aspects of the sweeping process. *Comput. Methods Appl. Mech. Engrg*, 177(3-4):329–349, 1999.

[Mor03] J-J. Moreau : Modélisation and simulation de matériaux granulaires. *In Actes du 35ème Canum*, 2003.

[Mor06] J.-J. Moreau : Facing the plurality of solutions in nonsmooth mechanics. Nonsmooth/Nonconvex Mechanics with Applications in Engineering, II. NNMAE 2006, pp. 3–12, 2006.

[PB01] M. Perregaard et E. Balas : Generating cuts from multiple-term disjunctions. *Lecture Notes in Computer Science*, 2081:348–360, 2001.

[PB03] M. Perregaard et E. Balas : A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0-1 programming. *Mathematical Programming B*, 94:221–245, 2003.

[PC09] S. Pan et J.-S. Chen : A damped Gauss-Newton method for the second-order cone complementarity problem. *Applied Mathematics and Optimization*, 59(3):293–318, 2009.

[Roc70] R.T. Rockafellar : *Convex analysis.* Princeton, N. J. : Princeton University Press, XVIII, 451 p. , 1970.

[RS02] P. Rey et C. Sagastizábal : Convex normalizations in lift-and-project methods for 0-1 programming. *Journal Annals of Operations Research*, 116:91–121, 2002.

[RW98] R.T. Rockafellar et R.J.-B. Wets : *Variational Analysis.* Springer Verlag, Heidelberg, 1998.

[Sch86] A. Schrijver : *Theory of Linear and Integer Programming.* Wiley and Sons, 1986.

[Spe75] D. A. Spence : The Hertz contact problem with finite friction. *Jour. of Elasticity*, 5(3-4):297–319, 1975.

[ST00] D. Stewart et J.C. Trinkle : An implicit time-stepping scheme for rigid body dynamics with Coulomb friction. *In IEEE International Conference on Robotics and Automation*, volume 1, pages 162–169. IEEE ; 1999, 2000.

[Ste98] D. Stewart : Convergence of a Time-Stepping Scheme for Rigid-Body Dynamics and Resolution of Painleve's Problem. *Arch. Rational Mech. Anal.*, 154:215–260, 1998.

[Str91] W.J. Stronge : Unraveling paradoxical theories for rigid body collisions. *J. Appl. Mech.*, 58(4):1049–1055, 1991.

[Stu99] J. Sturm : Using SeDuMi 1. 02, a MATLAB toolbox for optimization over symmetric cones. *Opt. Meth. Soft.*, 1999.

[Wol76] P. Wolfe : Finding the nearest point in a polytope. *Math. Prog.*, 11:128 – 149, 1976.

[Wri06] P. Wriggers : *Computational contact mechanics. 2nd ed.* Springer, 2006.