Problèmes inverses : aspects numériques

Michel Kern

ÉCOLE SUPERIEURE D'INGÉNIEURS
LÉONARD DE VINCI

PROBLÈMES INVERSES ASPECTS NUMÉRIQUES

Michel KERN [1]

2002–2003 CS 305

1. INRIA, ROCQUENCOURT, BP 105, 78153 LE CHESNAY, Michel.Kern@inria.fr

Table des matières

Première partie

Introduction et exemples

Chapitre 1

Introduction

1.1 Introduction

D'après J.B. Keller [41], deux problèmes sont dits *inverses* l'un de l'autre si la formulation de l'un met l'autre en cause. Cette définition comporte une part d'arbitraire, et fait jouer un rôle symétrique aux deux problèmes considérés. Une définition plus opérationnelle est qu'un problème inverse consiste à déterminer des *causes* connaissant des *effets*. Ainsi, ce problème est l'inverse de celui appelé problème direct, consistant à déduire les effets, les causes étant connues.

Cette seconde définition montre que nous sommes plus habitués à étudier des problèmes ń directs ż. En effet, depuis Newton la notion de causalité est ancrée dans notre subconscient scientifique, et à un niveau plus prosaïque, nous avons appris à poser, puis résoudre des problèmes pour lesquels les causes sont données, et l'on cherche les effets. Cette définition montre aussi que les problèmes inverses risquent de poser des difficultés particulières. Nous verrons plus loin qu'il est possible de donner un contenu mathématique à la phrase ń les mêmes causes produisent les mêmes effets ż, autrement dit, qu'il est raisonnable d'exiger que le problème direct soit ń bien posé ż. Par contre, il est facile d'imaginer, et nous en verrons de nombreux exemples, que les mêmes effets puissent provenir de causes différentes. Cette idée contient en germe la principale difficulté de l'étude des problèmes inverses : ils peuvent avoir plusieurs solutions, et il est nécessaire de disposer d'informations supplémentaires pour discriminer entre elles.

La prédiction de l'état futur d'un système physique, connaissant son état actuel, est l'exemple type du problème direct. On peut envisager divers problèmes inverses : par exemple, reconstituer l'état passé du système connaissant son état actuel (si ce système est irréversible), ou la détermination de paramètres du système, connaissant (une partie de) son évolution. Ce dernier problème est celui de *l'identification* de *paramètres*, qui sera notre principale préoccupation dans la suite.

Une difficulté pratique de l'étude des problèmes inverses est qu'elle demande souvent une bonne connaissance du problème direct, ce qui se traduit par le recours à une grande variété de notions tant physiques que mathématiques. Le succès dans la résolution d'un problème inverse repose en général sur des éléments spécifiques à ce problème. Il existe toutefois quelques techniques qui possèdent un domaine d'applicabilité étendu, et ce cours est une introduction aux principales d'entre elles : la régularisation des problèmes mal posés, et la méthode des moindres carrés.

La plus importante est la reformulation d'un problème inverse sous la forme de la minimisation d'une fonctionnelle d'erreur entre les mesures réelles et les ń mesures synthétiques ż (c'est-à-dire la solution du problème direct). Il sera commode de distinguer entre les problèmes linéaires et non-

linéaires. Précisons ici que la non-linéarité dont il s'agit ici fait référence au problème inverse lui-même, et non pas au problème direct (en considérant les paramètres comme connus).

Dans le cas des problèmes linéaires, le recours à l'algèbre linéaire et à l'analyse fonctionnelle permet d'obtenir des résultats précis, et des algorithmes efficaces. L'outil fondamental est ici la décomposition en valeurs singulières de l'opérateur considéré. Nous étudierons en détail la méthode de régularisation, qui consiste à ń modifier ż légèrement le problème étudié en un autre qui possède de ń meilleures ż propriétés. Ceci sera précisé aux chapitres 4 et 6.

Les problèmes non-linéaires sont plus difficiles, et il existe moins de résultats généraux. Nous étudierons l'application des algorithmes d'optimisation aux problèmes obtenus par la reformulation évoquée plus haut. Un ingrédient technique essentiel (du point de vue numérique) est le calcul du gradient de la fonctionnelle à minimiser. Nous étudierons la méthode de l'état adjoint au chapitre 8. Elle permet ce calcul pour un coût qui est un (petit) multiple de celui de la résolution du problème direct.

Comme on le voit, le contenu de ce cours vise surtout à présenter des *méthodes numériques* pour aborder les problèmes inverses. Cela ne veut pas dire que les questions *théoriques* n'existent pas, ou soient sans intérêt. Le choix délibéré de ne pas les aborder est dicté par l'orientation pratique de ce cours, par le goût et les connaissances de l'auteur, mais aussi par le niveau mathématique élevé que ces questions nécessitent.

1.2 Problèmes bien et mal posés

Dans un livre célèbre, Hadamard [32] a introduit dès 1923 la notion de *problème bien posé*. Il s'agit d'un problème dont :

- la solution existe ;
- elle est unique ;
- elle dépend continûment des données.

Bien entendu, ces notions doivent être précisées par le choix des espace (et des topologies) dans lesquels les données et la solution évoluent.

Dans ce même livre Hadamard laissait entendre (et c'était une opinion répandue jusqu'à récemment) que seul un problème bien posé pouvait modéliser correctement un phénomène physique. Après tout, ces trois conditions semblent très naturelles. En fait, nous verrons que les problèmes inverses ne vérifient souvent pas l'une ou l'autre de ces conditions, voire les trois ensembles. Après réflexion, cela n'est pas si surprenant :

- Un modèle physique étant fixé, les données expérimentales dont on dispose sont en général bruitées, et rien ne garantit que de telles données proviennent de ce modèle, même pour un autre jeu de paramètres.
- Si une solution existe, il est parfaitement concevable (et nous le verrons sur des exemples) que des paramètres différents conduisent aux mêmes observations.
- Le fait que la solution d'un problème inverse puisse ne pas exister n'est pas une difficulté sérieuse. Il est habituellement possible de rétablir l'existence en relaxant la notion de solution (procédé classique en mathématique).
- La non-unicité est un problème plus sérieux. Si un problème a plusieurs solutions, il faut un moyen de choisir entre elles. Pour cela, il faut disposer d'informations supplémentaires (une information *a priori*).
- Le manque de continuité est sans doute le plus problématique, en particulier en vue d'une résolution approchée ou numérique. Cela veut dire qu'il ne sera pas possible (indépendamment

de la méthode numérique) d'approcher de façon satisfaisante la solution du problème inverse, puisque les données disponibles seront bruitées donc proches, mais différentes, des données ń réelles ż.

Un problème qui n'est pas bien posé au sens de la définition ci-dessus est dit ń mal posé ż (ill-posed en anglais). Nous allons en voir un exemple qui, bien que très simple, illustre les difficultés que l'on peut rencontrer dans des situations plus générales.

Exemple 1.1.
La différentiation est l'intégration sont deux problèmes inverses l'un de l'autre. Il est plus habituel de penser à la différentiation comme problème direct, et à l'intégration comme problème inverse. En fait, l'intégration possède de bonnes propriétés mathématiques qui conduisent à le considérer comme le problème direct. Et la différentiation est le ń prototype ż du problème mal posé, comme nous allons le voir.

Considérons l'espace de Hilbert $L^2(\Omega)$, et l'opérateur intégral A défini par

$$Af(x) = \int_0^x f(t)\,dt. \tag{1.1}$$

Il est facile de voir directement que $A \in \mathcal{L}(L^2(0,1)$, ou l'on peut appliquer le théorème 3.1 (voir l'exemple 3.1). Cet opérateur est injectif, par contre son image est le sous espace vectoriel

$$\mathrm{Im}A = \{f \in H^1(0,1), u(0) = 0\}$$

où $H^1(0,1)$ est l'espace de Sobolev. En effet, l'équation

$$Af = g$$

est équivalente à

$$f(x) = g'(x) \text{ et } g(0) = 0.$$

L'image de A n'est *pas* fermée dans $L^2(0,1)$ (bien entendu, elle l'est dans $H^1(0,1)$). En conséquence, l'inverse de A n'est pas continu sur $L^2(0,1)$, comme le montre l'exemple suivant.

Considérons une fonction $f \in C^1([0,1])$, et $n \in \mathbf{N}$. Soit

$$f_n(x) = f(x) + \frac{1}{n}\sin\left(n^2 x\right).$$

Alors

$$f_n'(x) = f'(x) + n\cos\left(n^2 x\right).$$

De simples calculs montrent que

$$\|f - f_n\|_2 = \frac{1}{n}\left(\frac{1}{2} - \frac{1}{4n}\sin\left(2n^2\right)\right)^{1/2} = 0\left(\frac{1}{n}\right)$$

alors que

$$\left\|f' - f_n'\right\|_2 = n\left(\frac{1}{2} + \frac{1}{4n}\sin\left(2n^2\right)\right)^{1/2} = O(n)$$

Ainsi, la différence entre f' et f_n peut-être arbitrairement grande, alors même que la différence entre f et f' est arbitrairement petite. L'opérateur de dérivation (l'inverse de A) n'est donc pas continu, au moins avec ce choix des normes. □

L'instabilité de l'inverse est typique des problèmes mal posés. Une petite perturbation sur les données (ici f) peut avoir une influence arbitrairement grande sur le résultat (ici f').

Une seconde classe de problèmes inverses est l'estimation de paramètres dans les équations différentielles. Nous allons en voir un exemple très simple.

Exemple 1.2.
On considère le problème elliptique en dimension 1 :

$$\begin{cases} -\big(a(x)u'(x)\big)' = f(x), & \text{pour } x \in]-1,1[\\ u(-1) = u(1) = 0. \end{cases} \tag{1.2}$$

Cette équation, ou d'autres analogues bien que plus complexes, dans de nombreux exemples au chapitre suivant. Dans cet exemple, nous prendrons $a(x) = x^2 + 1$, et la solution $u(x) = (1-x^2)/2$, ce qui donne $f(x) = 3x^2 + 1$.

Le problème direct consiste à calculer u, étant donné a et f. Pour le problème inverse, nous considérerons que f est connue, et nous chercherons à retrouver le coefficient a à partir d'une mesure de u. Pour cet exemple, volontairement simplifié, nous supposerons que l'on mesure u en tout point de l'intervalle $]-1,1[$, ce qui est bien évidemment irréaliste. Nous allons voir que même dans cette situation optimiste, nous sommes susceptibles de rencontrer des difficultés.

En intégrant l'équation (1.2), et en divisant par u', nous obtenons l'expression suivante pour a (en supposant que u' ne s'annule pas, ce qui est faux sur notre exemple) :

$$a(x) = \frac{C}{u'(x)} + \frac{1}{u'(x)} \int_0^x f(\xi)\, d\xi, \tag{1.3}$$

ce qui donne dans notre cas particulier :

$$a(x) = \frac{C}{x} + x^2 + 1 \quad \text{pour } x \neq 0, \tag{1.4}$$

où C est une constante d'intégration.

Nous voyons que, même dans ce cas particulier, a n'est pas déterminée par les données, c'est-à-dire u. Bien entendu dans ce cas, il est clair que la ń bonne ż solution correspond à $C = 0$, puisque c'est la seule valeur pour laquelle a est bornée. Pour pouvoir discriminer parmi les différentes solutions possibles, nous avons du faire appel à une information supplémentaire (on parle généralement d'ń information a priori ż).

Il y a dans ce problème deux sources d'instabilité : tout d'abord l'équation (1.3) fait intervenir u', et nous venons de voir que le passage de u à u' est source d'instabilité. Il s'agit là d'un phénomène commun aux problèmes linéaires et non-linéaires. Par contre, la division par u' montre une instabilité spécifique des problèmes non-linéaires. Si u' s'annule, la division est impossible. Si u' est simplement ń petit ż, la division sera cause d'instabilité.

□

Le reste de ce cours est consacré à l'étude de méthodes permettant de rétablir une certaine stabilité dans les problèmes mal posés. Il faut toutefois garder présent à l'esprit cette remarque de [25] ń No mathematical trick can make an inherently unstable problem stable ż (Aucun artifice mathématique ne peut rendre stable un problème intrinsèquement instable). Les méthodes que nous allons introduire dans la suite vont rendre le problème considéré stable, mais au prix d'une modification du problème résolu (et donc de sa solution !).

1.3 Plan du cours

Dans le reste de ce chapitre, nous donnerons plusieurs exemples de problèmes inverses, provenant de plusieurs domaines de la physique. Nous introduirons la notion fondamentale de *problème mal posé*, qui est caractéristique des problèmes inverses.

Au chapitre 3, nous introduirons une source importante de problèmes inverses linéaires : les équations intégrales de première espèce. Après avoir exposé les principales propriétés de opérateurs intégraux, nous expliquerons en quoi ils sont mal posés. Enfin nous introduirons des méthodes de discrétisation, conduisant à des problèmes de moindres carrés.

L'étude de ces problèmes fait l'objet des deux chapitres suivants. Au chapitre 4, nous étudierons leurs propriétés mathématiques, dans un cadre hilbertien : l'aspect géometrique, et le lien avec les équations normales, ainsi que les questions d'existence et d'unicité des solutions. Nous introduirons également l'outil fondamental, tant pour l'analyse théorique que pour l'approximation numérique, qu'est la *décomposition en valeurs singulières*, tout d'abord pour les matrices, puis pour les opérateurs entre espaces de Hilbert. L'aspect numérique des problèmes inverses sera étudié au chapitre 5. Après un exposé des avantages et inconvénients de la méthode des équations normales, nous introduirons la méthode basée sur une factorisation *QR* de la matrice, puis montrerons comment la décomposition en valeurs singulières conduit à une méthode numérique. Nous conclurons par quelques indications sur le calcul effectif de cette décomposition. Au chapitre 6, nous aborderons l'étude des techniques pour les problèmes mal posés, tout particulièrement la méthode de régularisation de Tikhonov, et la troncature spectrale.

Dans une deuxième partie, nous aborderons les problèmes non-linéaires, essentiellement les problèmes d'estimation de paramètres dans les équations différentielles, et aux dérivées partielles. Au chapitre 7, nous verrons comment poser les problèmes d'identification en terme de minimisation, quelles sont les principales difficultés auxquelles on peut s'attendre, ainsi que des rappels sur les méthodes numériques de base en optimisation. Le chapitre 8 abordera la technique importante de l'état adjoint pour calculer le gradient des fonctionnelles qui interviennent dans les problèmes de moindres carrés. Nous verrons sur plusieurs exemples comment mener à bien ce calcul de façon efficace.

Nous avons rassemblé quelques résultats d'analyse fonctionnelle qui nous serons utile, ainsi que des compléments sur les opérateurs linéaires dans un appendice.

Chapitre 2

Exemples de problèmes inverses

Nous présentons dans ce chapitre quelques exemples ń concrets ż de problèmes inverses, tels qu'ils interviennent dans les sciences de l'ingénieur. Cette liste est loin d'être exhaustive (voir les références à la fin de ce chapitre pour d'autres applications).

Parmi les domaines dans lesquels les problèmes inverses jouent un rôle important nous pouvons citer :

- l'imagerie médicale (échographie, scanners, rayons X, ...) ;
- l'ingénierie pétrolière (prospection par des méthodes sismiques, magnétiques, identification des perméabilités dans un réservoir ...) ;
- l'hydrogéologie (identification des perméabilités hydrauliques) ;
- la chimie (determination des constantes de réaction) ;
- le radar (determination de la forme d'un ń obstacle ż) ;
- l'acoustique sous-marine (même objectif !) ;
- la mécanique quantique (détermination du potentiel) ;
- le traitement d'image (restauration d'images floues).

Du point de vue mathématique, ces problèmes se répartissent en deux grands groupes :

- les problèmes linéaires (échographie, traitement d'image, ...), qui se ramènent à la résolution d'une équation intégrale de première espèce ;
- les problèmes non-linéaires, qui sont le plus souvent des questions d'estimation de paramètres dans des équations différentielles ou aux dérivées partielles.

La seconde catégorie peut elle-même se subdiviser en deux sous-catégories selon que le paramètre que l'on cherche à estimer est un vecteur (donc de dimension finie), où une fonction. Le second cas est évidemment plus difficile que le premier, puisqu'il faut en particulier décider de la paramétrisation de cette fonction, avant de résoudre numériquement le problème en dimension finie.

2.1 Problèmes inverses en thermique

Pour déterminer la répartition de la température dans un matériau inhomogène occupant un domaine (ouvert connexe) Ω de $\mathbf{R}^3$ on écrit tout d'abord la conservation de l'énergie :

$$\rho c \frac{\partial T}{\partial t} + \mathrm{div}\,(\vec{q}) = f(x,y,z) \quad \text{dans } \Omega \tag{2.1}$$

où T est la température, ρ la densité du fluide, c la chaleur spécifique, $\vec{q}$ représente un flux de chaleur et f une source volumique.

La loi de Fourier relie ensuite le flux de chaleur au gradient de température :

$$\vec{q} = -K \operatorname{grad} T, \tag{2.2}$$

où K est la conductivité thermique (qui peut être un tenseur, et dépend de la position).

En éliminant $\vec{q}$, on obtient l'équation de la chaleur en milieu hétérogène :

$$\rho c \frac{\partial T}{\partial t} - \operatorname{div}(K \operatorname{grad} T) = f \quad \text{dans } \Omega. \tag{2.3}$$

Cette équation soit être complétée par des conditions aux limites sur le bord de l'ouvert Ω, et une condition initiale.

Le problème direct est de déterminer T connaissant les coefficients physiques ρ, c et K, ainsi que la source de chaleur f. Ce problème est bien connu, tant du point de vue théorique (existence et unicité de la solution) que du point de vue numérique. Plusieurs problèmes inverses peuvent être posés :

- étant donné une mesure de la température à un instant $t_f >$, déterminer la température initiale. Nous l'aborderons à l'exemple 2.1 ;
- étant donné une mesure (partielle) de la température, déterminer certains des coefficients de l'équation.

Notons que le premier de ces problèmes est linéaire, alors que le second est non-linéaire : en effet l'application $(\rho, c, K) \mapsto T$ est non-linéaire.

Exemple 2.1 (Équation de la chaleur rétrograde)**.**
Nous prenons le cas idéal d'un matériau infini et homogène (en une dimension d'espace pour simplifier). La température est solution de *l'équation de la chaleur* :

$$\frac{\partial T}{\partial t} - \frac{\partial^2 T}{\partial x^2} = 0 \tag{2.4}$$

(il n'y a pas de source). On suppose connue la température à un certain instant t_f, soit $T_f(x) = T(x, t_f)$, et l'on cherche à retrouver la température initiale $T_0(x) = T(x, 0)$.

Le problème de déterminer T_f connaissant T_0 est le problème de Cauchy pour l'équation de la chaleur. Il a une solution unique, qui dépend continûment de la donnée initiale. Comme nous allons le voir, il n'en est rien pour le problème inverse que nous considérons ici. Physiquement, cela est du au caractère irréversible de la diffusion thermique. Il est bien connu que la température a tendance à s'homogénéiser au cours du temps, et cela entraîne qu'il n'est pas possible de revenir en arrière, c'est-à-dire de retrouver l'état antérieur qui peut être plus hétérogène que l'état actuel.

Grâce à la situation très simplifiée que nous avons choisie, nous pouvons calculer à la main la solution de l'équation de la chaleur (2.4). En prenant la transformée de Fourier en espace de l'équation (2.4) (nous notons $\hat{T}(k,t)$ la trnsformée de Fourier de $T(x,t)$ en gardant t fixé), nous obtenons une équation différentielle ordinaire (où cette fois c'est k qui joue le rôle d'un paramètre) dont la solution est

$$\hat{T}_f(k) = e^{-|k|^2 t_f} \hat{T}_0(k). \tag{2.5}$$

En prenant la tranformée de Fourier inverse, nous voyons que la solution à l'instant T_f est reliée à la condition initiale par une convolution avec la solution élémentaire de la chaleur :

$$T_f(x) = \frac{1}{2\sqrt{\pi t_f}} \int_{-\infty}^{+\infty} e^{-(x-y)^2/4t_f} T_0(y)\, dy. \tag{2.6}$$

Il est bien connu [13] que, pour toute fonction T_0 ń raisonnable ż (continue, bornée), la fonction T_f est indéfiniment dérivable, ce qui traduit mathématiquement l'irréversibilité mentionnée ci-dessus.

En restant dans le domaine de Fourier, nous pouvons inverser ponctuellement l'équation (2.5), mais la fonction

$$k \mapsto e^{|k|^2 t}\, \hat{T}_f(k)$$

ne sera dans $L^2(\mathbf{R})$ que pour des fonctions T_f décroissant très rapidement à l'infini, ce qui est une restriction très sévère. Une température mesurée expérimentalement a peu de chances de la satisfaire, et c'est ce qui entraîne l'instabilité du problème inverse.

Nous retrouverons l'analogue de cette condition au chapitre 4, 4.3.1 quand nous étudierons la condition de Picard. □

Passons maintenant à des exemples d'estimamtion de paramètres.

Exemple 2.2 (Identification d'un coefficient dans un modèle stationnaire)**.**
Pour simplifier, on ne considère que le régime permanent, et on suppose que le bord du domaine est maintenue à une température de 0. L'équation de la chaleur (2.3), et la condition au bord, donnent alors :

$$\text{(2.7)} \qquad \begin{cases} -\operatorname{div}(K \operatorname{grad} T) = f(x,y,z) & \text{dans } \Omega \\ T = 0 & \text{sur } \partial\Omega \end{cases}$$

Le problème direct suppose connues la conductivité thermique K et la source de chaleur f, et cherche à déterminer la répartition de la température T en tout point du matériau. Il s'agit là du prototype des équations elliptiques du second ordre, et sa résolution est des plus classiques : sous des hypothèses raisonnables sur K ($K \in L^\infty(\Omega)$ avec $0 < K_* \leq K \leq K^* < \infty$) et sur f ($f \in L^2(\Omega)$, il admet une solution unique, d'après le théorème de Lax-Milgram. De plus, un calcul numérique avec une méthode d'éléments finis est tout-à-fait standard (voir [19, chap. V]).

Il en va différemment du problème inverse. Pour pouvoir le spécifier, il est tout d'abord nécessaire de préciser de quelle *observation* l'on dispose. Cela dépend bien évidemment du dispositif expérimental utilisé, mais en tout état de cause, il ne sera généralement pas réaliste de supposer que l'on connait la température en tout point. Dans notre cas, ces observations pourraient être, par exemple, des mesures de la température à l'intérieur du matériau, ou bien des mesures du flux de chaleur $-K\dfrac{\partial T}{\partial n}$ sur le bord du domaine (on parle dans ce cas d'observatio frontière). Le problème inverse est alors de chercher la (ou une) fonction de conductivité, telle qu'il existe une fonction T solution de (2.7) qui coïncide avec les observations.

On voit immédiatement plusieurs difficultés possibles :

- Tout d'abord celle d'obtenir les observations ! Une expérience n'est jamais facile à réaliser. Dans notre exemple, il n'est pas réaliste de supposer que l'on puisse mesurer la température en tout point du domaine (penser à une pièce dans un appartement).
- Mais alors on risque de ne pas disposer de suffisamment d'observations par rapport au nombre de paramètres que l'on cherche. Ici, si l'on dispose d'une observation frontière, c'est-à-dire d'une fonction de deux variables, il sera difficile de retrouver une fonction de trois variables, indépendamment de la méthode utilisée.
- En particulier, on voit tout de suite que si la température est constante dans un sous-domaine de Ω, la conductivité n'y est pas déterminée. Il faudrait donc disposer d'informations supplémentaires permettant de combler ce manque de mesure.

– Enfin, toute mesure est entachée d'erreur, et d'ailleurs le modèle mathématique (2.7) ne reflète pas exactement la réalité. Ainsi, il n'y a en fait aucune raison pour que le problème inverse possède une solution.

Ce problème difficile a fait l'objet de nombreuses études, tant théoriques que numériques (voir [39] pour une introduction). Signalons qu'il intervient dans d'autres domaines d'application (médical, prospection géophysique par des méthodes électriques ou magnétiques, ...) □

Exemple 2.3 (Identification en une dimension d'espace)**.**
Nous pouvons mieux comprendre cet exemple en le réduisant à une seule dimension. Il s'agit de déterminer la fonction $K(x)$ à partir de l'équation

$$-\big(K(x)T'(x)\big)' = f(x), \quad \text{pour } x \in]0,1[\tag{2.8}$$

(avec des conditions aux limites convenables), et de la connaissance de T. Pour simplifier, nous supposerons que l'on connaît T en tout point de l'intervalle $]0,1[$. Sous l'hypothèse que T' ne s'annule en aucun point de l'intervalle (qui n'est pas nécessairement vérifiée), on peut intégrer l'équation (2.8), pour obtenir :

$$K(x) = \frac{1}{T'(x)}\left(K(0)T'(0) - \int_0^x f(t)\,dt\right). \tag{2.9}$$

Cette équation montre, sous l'hypothèse que nous avons faite ci-dessus, que K est uniquement déterminée dès que sa valeur en un point est connue. Par contre, la formule donnant K fait intervenir la *dérivée* de l'observation T. Comme nous l'avons vu au paragraphe 1.1, cette opération n'est pas continue. Ainsi l'application (non-linéaire) $T \mapsto K$ ne le sera pas non plus. Une petite erreur sur la mesure de T pourra se traduire par une erreur arbitrairement grande sur la conductivité K.

De plus, une seconde instabilité provient de la division par $T'(x)$, ce qui est un effet essentiellement non-linéaire. Si $T'(x)$ est petit, la division par $T'(x)$ risque d'amplifier les erreurs déjà présentes dans la mesure de T. Ceci est bien sur lié au fait que si T' s'annule, K n'est pas déterminé du tout.

Pour plus de renseignements sur cet exemple, on pourra consulter l'article [24] ou le livre [25]. □

2.2 Problèmes inverses en hydrogéologie

L'hydrogéologie, ou l'étude des nappes phréatiques, est une autre source abondantes de problèmes inverses. Il en effet difficile d'accéder aux couches du sous-sol pour mesurer les propriétés aqueuses des roches. Un problème actuellement d'actualité est le contrôle des polluants dans les nappes d'eau souterraines. Pour mentionner un seul exemple pratique, la thèse d'habilitation de R. Mosé [52] consistait en l'étude de l'influence d'un accident de camion transportant du gaz CCL4 dans l'est de la France en 1970 sur l'eau qui est consommée actuellement. Un paramètre fondamental de cette étude est la conductivité hydraulique du sous-sol, qui dépend évidemment de la position.

Il existe une grande variété de modèles physiques, incluant diverses approximations. Nous en présentons un ci-dessous, en nous inspirant de la thèse de P. Siegel [57] (voir aussi [59, 8, 20]) :

Exemple 2.4 (Transport d'un polluant par un aquifère)**.**
Un milieu poreux est constitué d'une matrice rocheuse, comprenant des pores qui peuvent laisser passer l'eau. Il est essentiellement impssible de décrire l'écoulement d'un fluide dans un tel milieu hétérogène, dans la mesure où l'on doit prendre en compte des échelles spatiales allant du centimètre (le pore) au kilomètre (le modèle régional), et que la disposition précise des pores n'est de toutes

fçons pas connue. On utilise alors des modèles physiques simplifiés, le plus courant étant la loi de Darcy, qui relie la hauteur de l'eau dans le milieu, appelée *charge piézométrique* et notée $h(x,y,z,t)$, à la vitesse de filtration $\vec{q}(x,y,z,t)$. Cette loi exprime que la vitesse est proportionnelle à l'opposé du gradient hydraulique :

$$\vec{q} = -K \operatorname{grad} h \tag{2.10}$$

où K est le coefficient de conductivité hydraulique. Ce peut en principe être un tenseur, mais nous nous restreindrons au cas où c'est un scalaire.

On exprime également la conservation de la masse (on fait l'hypothèse que le milieu est incompressible) :

$$S\frac{\partial h}{\partial t} + \operatorname{div} \vec{q} = f \tag{2.11}$$

où S est le coefficient d'emmagasinement spécifique, et f est une source (supposée connue).

L'élimination de $\vec{q}$ donne pour h l'équation parabolique :

$$S\frac{\partial h}{\partial t} - \operatorname{div}(K \operatorname{grad} h) = f \tag{2.12}$$

à laquelle on ajoute des conditions initiales (h donné à $t = 0$) et aux limites (Dirichlet, correspondant à une charge imposée, ou Neumann, correspondant à un flux imposé).

Les problèmes de transport de contaminant font intervenir, en plus de l'écoulement, la façon dont évolue la concentration d'une espèce (composé chimique, hydrocarbure, radionucléide) ń portée ż par l'écoulement. Ce phénomène met en jeu trois mécanismes : la convection (imposée par la vitesse de filtration $\vec{q}$), la diffusion moléculaire et la dispersion cinématique. Nous ne décrirons pas ces deux derniers mécanismes en détail (pour cela voir les références citées plus haut). La quantité étudiée est la concentration $C(x,y,z,t)$ du polluant, qui obéit à une équation de type convection–diffusion :

$$\frac{\partial C}{\partial t} + \frac{1}{\varepsilon} \operatorname{div}(\vec{q}C) - \operatorname{div}(D \operatorname{grad} C) = f_c \tag{2.13}$$

ou ε est la porosité cinématique (fraction des pores occupés par l'eau en mouvement), D est le tenseur de diffusion (en agrégeant diffusion moléculaire et la dispersion cinématique), et f_c est une éventuelle source de polluant. On ajoute une condition initiale (concentration connue à l'instant initial), et des conditions aux limites.

Le problème direct est constitué par les équations (2.12) et (2.13). Ce problème couplé est théoriquement non-linéaire, à cause du terme $\operatorname{div}(\vec{q}C)$. En pratique, cependant, on peut souvetnt résoudre d'abord l'équation (2.12), puis (2.13), $\vec{q}$ étant connu.

On mesure, par exemple, la concentration en un certain nombres de points de mesures, et à des instants discrets (il n'est pas réaliste ici de supposer que la mesure est continue en temps). On connaît donc $C(x_o,y_o,z_o,t_o)$, $o = 1,\ldots,N_o$. Le problème inverse est alors de chercher la conductivité hydraulique (et dans une moindre mesure les autres paramètres du modèle), connaissant ces mesures. Ce problème est sous-determiné, car il est rare que l'on ait accès à suffisamment de mesures. □

Exemple 2.5 (Le cas stationnaire)**.**
On considère le régime stationnaire dans le modèle ci-dessus. Dans ce cas, on ne prend en compte que l'équation (2.12), dans laquelle on suppose de plus que $S = 0$, et que la source f est indépendante du temps. On a alors simplement une équation elliptique du second ordre : $-\operatorname{div}(K \operatorname{grad} h) =$

f dans un ouvert Ω. Prenons par exemple le cas ou la charge piézoélectrique est imposée sur le bord $\partial\Omega$. On mesure le flux $K\dfrac{\partial h}{\partial n}$ sur le bord, et on cherche toujours à identifier le coefficient K. On retrouve ici un modèle elliptique du type de celui étudié au paragraphe 2.1. □

Exemple 2.6 (Hydrogéologie 1D)**.**
Ici encore, nous considérerons le problème simplifié ou l'écoulement est essentiellement mono-dimensionnel, dans une direction horizontale que nous prendrons comme axe Ox. Un tel modèle s'obtient par intégration sur des couches verticales du modèle précédent. Les équations s'écrivent :

$$\text{(2.14)} \qquad \begin{cases} S\dfrac{\partial h}{\partial t} - \dfrac{\partial}{\partial z}\left(K\dfrac{\partial h}{\partial z}\right) = f & \text{dans } [0,L]\times]0,T[\\ q = -K\dfrac{\partial h}{\partial z} & \text{dans } [0,L]\times]0,T[\\ \dfrac{\partial C}{\partial t} + \dfrac{1}{\varepsilon}\dfrac{\partial(\vec{q}C)}{\partial z} - \dfrac{\partial}{\partial z}\left(D\dfrac{\partial C}{\partial z}\right) = f_c & \text{dans } [0,L]\times]0,T[\end{cases}$$

avec (par exemples) des conditions initiales données, h fixé aux deux extrémités, et C donné en $x = 0$. On mesure $C(L,t)$, et l'on cherche à identifier $K(x)$. □

2.3 Problèmes inverses en sismique

La prospection pétrolière par des méthodes sismiques (et la sismologie) donne lieu à un problème inverse qui a été largement étudié en raison de l'intérêt économique qui s'attache à sa solution. Il s'agit en réalité d'une famille de problèmes inverses, dont le but commun est de déterminer les propriétés élastiques du sous-sol (densité, vitesses de propagation des ondes élastiques) à partir de mesures des champs de déplacement, ou de pression, en surface.

Lors d'une campagne sismique, une source (en général une explosion) provoque un ébranlement des roches formant le sous-sol. L'écho est enregistré par une série de capteurs placés en surface. Cette expérience est répétée pour plusieurs positions de la source (de plusieurs centaines à plusieurs milliers). De cette façon une très grande quantité de données est mesurée (pouvant atteindre des centaines de giga-octets). Le but, est encore une fois, d'estimer les propriétés du milieu étant donné un modèle de propagation. La communauté géophysique a mis au point une grande quantité de méthodes spécifiques pour traiter ce problème. Le livre récent [10] présente ces méthodes de façon synthétique.

Exemple 2.7 (Le modèle acoustique)**.**
Il existe plusieurs modèles physiques pouvant rendre compte (avec des degrés divers d'approximation) de l'expérience décrite ci-dessus. Nous nous bornerons à étudier l'un des plus simples : nous ferons l'hypothèse que la région étudiée se compose d'un fluide (ce qui correspond à une expérience de sismique sous-marine). Dans ce cas, on peut démontrer (cf. [19, vol. 1]) que la propagation des ondes est régie par *l'équation des ondes acoustiques*, et la quantité mesurée est un champ (scalaire) de pression. Il est commode de faire l'hypothèse que le domaine d'étude est le demi-espace $\{z > 0\}$ (la terre n'est évidemment ni plane ni infinie, mais ces approximations sont justifiées par les échelles considérées, qui sont ici de l'ordre de quelques kilomètres), l'axe Oz étant orienté vers le bas. Nous noterons :

$p = p(x,y,z,t)$	la pression,
$\rho = \rho(x,y,z)$	la densité,

$c = c(x,y,z)$ la vitesse de propagation,

$f = f(x,y,z,t)$ la source,

$\Omega = \mathbf{R}^2 \times \mathbf{R}_+$ le domaine spatial d'étude, et T la durée de l'expérience.

Le problème direct est alors, connaissant f, c et ρ de trouver p solution de

$$\text{(2.15)} \qquad \begin{cases} \dfrac{1}{\rho c^2}\dfrac{\partial^2 p}{\partial t^2} - \operatorname{div}\left(\dfrac{1}{\rho}\operatorname{grad} p\right) = f & \text{dans } \Omega \times]0,T[\\ p(x,y,z,0) = \dfrac{\partial p}{\partial t}(x,y,z,0) = 0 & \text{dans } \Omega \\ \dfrac{\partial p}{\partial z}(x,y,0,t) = 0 & \text{sur } \{z=0\} \end{cases}$$

La première équation de (2.15) représente la loi de Newton. La seconde est l'état initial du système (ici supposé au repos), la troisième est une condition aux limites, ici de surface libre sur la surface du sol. Il est bien connu (voir par exemple Dautray–Lions [19, vol. 1]) que sous des hypothèses ń raisonnables ż sur les coefficients ρ et c et la source f, (2.15) admet une unique solution p, qui dépend continûment de f. Il est également vrai, mais plus difficile à démontrer car la dépendance est non-linéaire, que p dépend de façon continue de c et ρ.

Notons que l'équation (2.15) décrit également le mouvement des ondes SH dans un solide élastique, en faisant l'hypothèse que la terre est bidimensionnelle, (les paramètres c, ρ et p ont alors des significations physiques différentes).

Une fois de plus, ce problème direct a été abondammen étudié, ses propriétés numériques sont boen connues, ainsi que des méthodes efficaces pour sa résolution numérique.

Le problème inverse consiste à déterminer c, ρ et f (qu'il n'est pas réaliste de supposer connu) à partir des mesures, c'est-à-dire de la connaissance de $\{p(x_g,y_g,z_g,t),\ g = 1,\dots,N_g,\ t \in [0,T]\}$. Ce problème est non-linéaire, puisque la solution p dépend de façon non-linéaire de c et ρ, même si l'équation aux dérivées partielles (2.15) est linéaire.

Il n'est pas réaliste de supposer que l'on connaisse la pression à chaque instant, en chaque point du domaine $\Omega \times]0,T[$. On dispose de plus de capteurs (*géophones* dans le cas de la sismique terrestre, *hydrophones* dans celui de la sismique marine), et on supposera pour simplifier que les enregistrements sont effectués en des points discrets, mais sont continus en temps (ceci est encore une fois justifié par la considération des échelles des phénomènes). On notera N_g le nombre de capteurs, et (x_g,y_g,z_g), $g = 1,\dots,N_g$ leurs positions. On extrait alors les mesures simulées : $p(x_g,y_g,z_g,t)$, $g = 1,\dots,N_g$, $t \in [0,T]$ du champ de pression solution de (2.15).

Dans la réalité, il y a un paramètre supplémentaire : l'expérience décrite ci-dessus est répétée en déplaçant le ń dispositif ż source–récepteurs. L'ensemble de ces ń tirs ż fournit une immense quantité de données. Ce problème inverse est surdéterminé. Une information importante à exploiter est que tous ces enregistrement proviennent du même sous-sol. Plusieurs méthodes d'inversion récentes partent de cette idée [61, 60, 17].

On fait couramment l'hypothèse que la densité du milieu est constante. Dans ce cas, l'équation (2.15) devient l'équation des ondes classiques :

$$\text{(2.16)} \qquad \begin{cases} \dfrac{1}{c^2}\dfrac{\partial^2 p}{\partial t^2} - \Delta p = f & \text{dans } \Omega \times]0,T[\\ p(x,y,z,0) = \dfrac{\partial p}{\partial t}(x,y,z,0) = 0 & \text{dans } \Omega \\ \dfrac{\partial p}{\partial z}(x,y,0,t) = 0 & \text{sur } \{z=0\}. \end{cases}$$

□

Remarque 2.1. Avant de considérer des modèles encore plus simples que celui-ci, signalons en quoi ce modèle lui-même est simplifié : en réalité, la terre n'est pas acoustique. Il faudrait utiliser un modèle plus réaliste, qui prenne mieux en compte la physique. Un grand nombre de modèles sont possibles, selon les phénomènes que l'on veut considérer : 2D ou 3D, acoustique, élastique (la terre est un solide), visco-élastique (tenant compte des mécanismes d'amortissement dans le sol), isotrope ou anisotrope,... Mais en fait, il n'est pas évident qu'un modèle ń raffiné ż soit supérieur à celui que nous venons d'évoquer. En effet, il ne sert à rien de rajouter des paramètres à un modèle si l'on n'est pas capable de mesurer des données supplémentaires qui permettraient de les déterminer. Ainsi, notre modèle représente (peut-être) un compromis raisonnable. Une discussion voisine (concernant le choix entre modèles 2D et 3D) se trouve dans le livre récent de Bleistein et al. bcs :00.

Exemple 2.8 (Le modèle stratifié)**.**
L'exemple que nous venons d'étudier, bien que déjà simplifié par rapport à la situation ń réelle ż, est encore compliqué : le problème direct requiert la solution d'une équation des ondes pour chaque position de la source ; de plus, le nombre d'inconnues nécessaires pour représenter la vitesse peut devenir gigantesque dans une situation 3D de géologie complexe. En fait, si la modélisation 3D est à la portée des super-ordinateurs modernes, l'inversion doit pour l'instant se contenter de modèles 2D, et ceux-ci sont extrêmement coûteux. On utilise donc encore des approximations du modèle acoustique. L'une d'elles, que l'on justifiera intuitivement en regardant un paysage rocheux en montagne, est de supposer que la terre est *stratifiée*, c'est à dire que les paramètres ρ et c ne dépendent que de la profondeur (la variable z). Si l'on suppose de plus que la source f est une onde plane (et l'on peut par un traitement approprié des données se ramener approximativement à cette situation), la pression p ne dépend plus que de z (et du temps), et l'équation (2.15) devient :

$$\text{(2.17)} \qquad \begin{cases} \dfrac{1}{\rho c^2}\dfrac{\partial^2 p}{\partial t^2} - \dfrac{\partial}{\partial z}\left(\dfrac{1}{\rho}\dfrac{\partial p}{\partial z}\right) = f & \text{dans } [0,Z]\times]0,T[\\ p(z,0) = \dfrac{\partial p}{\partial t}(z,0) = 0 & \text{dans } [0,Z] \\ p(0,t) = 0 & \text{sur } \{z=0\} \end{cases}$$

et l'on mesure $p(z_G,t), t \in [0,T]$ (la même simplification a lieu pour l'équation (2.16)).

Ce problème est plus simple, puisque l'équation (2.17) est à une seule dimension d'espace. Pour ce problème, un assez grand nombre de résultats sont connus, voir par exemple [3, 4, 5]. Cet exemple présente déjà les difficultés essentielles du cas général, mais permet une résolution économique du problème direct. □

Nous venons de voir que l'on peut poser des problèmes inverses pour les trois types d'équations aux dérivées partielles usuels : hyperbolique, parabolique et elliptique. Passons maintenant à des exemples d'un type différent, conduisant à des équations intégrales.

2.4 Imagerie médicale

Les sciences médicales fournissent un grand nombre de problèmes inverses, dont l'importance pratique n'échappera à personne. Nous allons évoquer rapidement quelques-uns d'entre eux. La de-

scription que nous présentons ici est empruntée à l'article de Louis [48], où l'on trouvera plus de détails.

Exemple 2.9 (Tomographie par rayons X)**.**
C'est la technique utilisée par les scanners. Un tube à rayons X est monté sur un portique qui entoure le patient. Les rayons émis sont mesurés par des détecteurs placés en face de l'émetteur. Nous considérons la situation bidimensionnelle, où le domaine représente une section transverse du patient. On suppose que les rayons suivent une ligne droite, et sont atténués à la traversée des tissus, proportionnellement à l'intensité elle-même, et à la distance parcourue (loi de Bouger). Les rayons X suivent des lignes droites, et nous paramétriserons ces lignes par leur vecteur normal $u \in \mathbf{R}^2$, et leur distance s à l'origine (voir la figure 2.1).

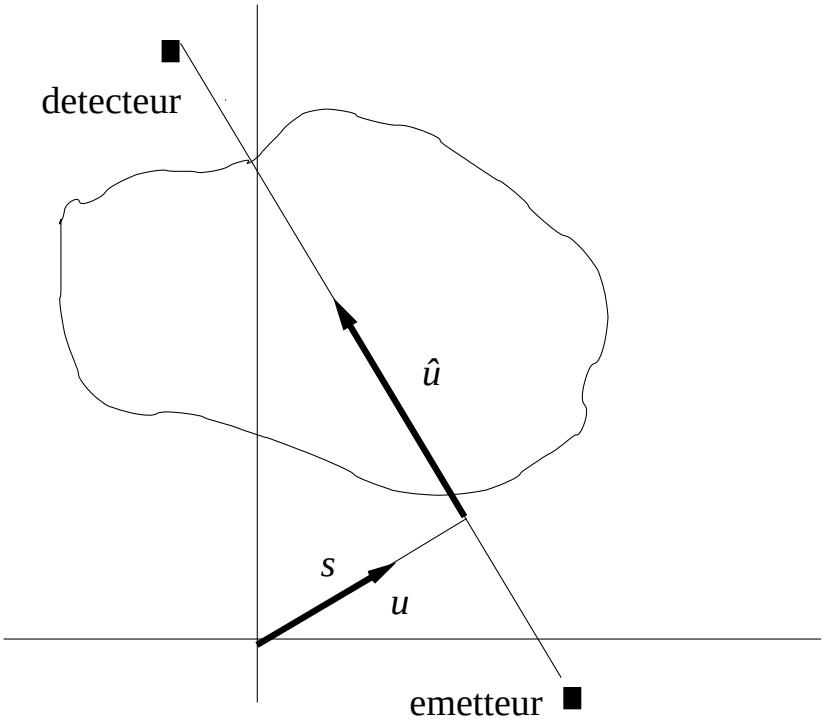

FIGURE 2.1 – Géométrie de l'expérience

En notant f le coefficient d'atténuation (qui peut dépendre du point), on obtient l'équation suivante :

$$\Delta I(su+t\hat{u}) = -I(su+t\hat{u})f(su+t\hat{u})\Delta t,$$

où $\hat{u}$ est un vecteur unitaire orthogonal à u. En faisant tendre Δt vers 0, on obtient l'équation différentielle suivante :

$$\frac{d}{dt}I(su+t\hat{u}) = -I(su+t\hat{u})f(su+t\hat{u}).$$

Notons I_0 et I_L les intensités à l'émetteur et au récepteur respectivement (que nous supposerons en dehors de l'objet, ce qui revient à dire qu'il sont à l'infini), l'équation différentielle précédente s'intègre en

$$-\ln\frac{I_l(s,u)}{I_0(s,u)} = \int_{\mathbf{R}} f(su+t\hat{u})\,dt. \tag{2.18}$$

Le problème direct consiste à déterminer l'intensité mesurée au détecteur connaissant celle à l'émetteur ainsi que la fonction d'atténuation f. Le problème inverse est donc de déterminer la fonction f connaissant les deux intensités.

L'opérateur intégral intervenant à droite de l'équation précédente s'appelle la *transformée de Radon* de f, d'après le mathématicien autrichien J. Radon, qui a d'ailleurs donné (en 1917 !) la formule d'inversion permettant en principe de reconstruire la fonction d'atténuation f à partir de la connaissance des transformées sur toutes les lignes du plan. Malheureusement le qualificatif *en principe* de la phrase précédente est important. En effet, la formule d'inversion suppose que Rf est connu pour toutes les directions u. Cela veut dire, en pratique, qu'il faut que les données soient mesurées de façon à peu près uniforme sur un cercle autour du patient (ce qui peut ou non être réalisable). Si cela n'est pas le cas, le problème est beaucoup plus délicat, et il est difficile de retrouver f de façon stable. Par ailleurs, comme nous le verrons dans u cas particulier ci-dessous, la formule de reconstruction fait intervenir la dérivée des mesures, ce qui montre également son caractère instable.

Ce problème est étudié en détails dans [53, 38]. □

Nous allons une fois de plus nous placer dans une situation simplifiée, où les calculs sont accessibles.

Exemple 2.10 (Tomographie à symétrie circulaire)**.**
Nous supposons maintenant que le milieu est un cercle de rayon ρ, et que la fonction f ne dépend que d'une variable ($f(s,u) = F(s)$). De plus, toute l'information est contenue dans l'intégrale prise selon une seule ligne, dont nous notons u_0 la direction. en notant

$$g(s) = -\ln \frac{I_l(s,u_0)}{I_0(s,u_0)}$$

L'équation (2.18) devient dans ce cas :

$$\int_0^\rho \frac{rF(r)}{\sqrt{r^2-s^2}}\,dr = \frac{g(s)}{2}. \tag{2.19}$$

Il s'agit d'une équation d'Abel, qui intervient dans plusieurs applications (nous y reviendrons avec l'exemple 2.13). Une référence très complète sur cette équation est l'ouvrage[29]. On démontre que l'équation (2.19) a une solution (unique) donnée par

$$F(r) = -\frac{1}{\pi}\int_r^\rho \frac{g'(s)}{\sqrt{s^2-r^2}}\,ds. \tag{2.20}$$

Une fois de plus, cette formule fait intervenir la dérivée de la donnée g ! □

Exemple 2.11 (Échographie)**.**
Cette méthode d'investigation présente le grand avantage d'être sans risques pour la patient. Les sources sont ici de brèves impulsions d'une onde acoustique à très haute fréquence, les mesures sont des échos acoustiques, et l'on recherche les discontinuités de la vitesse de propagation du milieu. Le problème direct est de calculer u_s connaissant q (et u_i), et le problème inverse est de retrouver q à partir de mesures de u_s effectuées loin de l'obstacle. Par rapport aux exemples du paragraphe 2.3, le problème est posé ici dans le domaine fréquentiel. Lorsque l'onde traverse le patient, elle est réfléchie par les changements dans la densité et les paramètres élastiques du ń milieu ż.

Notons ω la pulsation de la source, ρ la densité et p la pression. En posant $u = \rho^{-1/2}p$, on obtient l'équation d'Helmholtz :

$$\Delta u + (k^2 + q)u = 0 \tag{2.21}$$

où $k = \omega/c$ et et le potentiel q est relié à la densité par

$$q = \frac{1}{2\rho}\Delta\rho - \frac{3}{4}\left(\frac{1}{\rho}\nabla\rho\right)^2.$$

On considère que la source est une onde plane, notée u_i, qui est une solution de l'équation sans l'obstacle

$$(\Delta + k^2)u_i = 0,$$

et l'on définit l'onde diffractée u_s (s pour *scattered*) par $u_s = u - u_i$. L'onde diffractée est alors solution de

$$(\Delta u + k^2)u = -q(u_i + u_s) \tag{2.22}$$

On introduit alors la fonction de Green pour l'opérateur $\Delta + k^2$

$$G(x,y) = \begin{cases} \dfrac{i}{4}H_0^{(1)}(k|x-y|) & \text{en dimension 2} \\ \dfrac{1}{4\pi}e^{ik|x-y|}/|x-y| & \text{en dimension 3,} \end{cases}$$

où $H_0^(1)$ est la fonction de Hankel de première espèce et d'ordre 0. On peut alors montrer que l'équation (2.22) se met sous la forme d'une équation intégrale (dite équation de Lippmann-Schwinger) :

$$u_s(x) = k^2 \int_\Omega G(x,y)q(y)(u_i + u_s)(y)\, dy. \tag{2.23}$$

Du fait de la présence du terme qu_s au second membre, le problème inverse est non-linéaire. Une approximation raisonnable (l'approximation de *Born* dans certaines applications est de supposer que l'onde diffractée est négligeable devant l'onde incidente. L'équation de Lippmann-Schwinger devient

$$u_s(x) = k^2 \int_\Omega G(x,y)q(y)u_i(y)\, dy.$$

Il s'agit cette fois d'une équation intégrale *linéaire* pour q, qui est encore une équation de première espèce. Ce problème inverse est donc *mal posé*, comme celui de l'exemple 2.12.

La version non-linéaire du problème est plus difficile. Elle a fait l'objet de nombreux travaux, tant théoriques que numériques. Le livre de Colton et Kress [18] contient un état de l'art très complet sur ce problème. □

2.5 Autres exemples

Exemple 2.12 (Prospection gravimétrique)**.**
Il s'agit ici de déterminer l'emplacement, ou la forme, d'anomalies magnétiques dans une structure connue, à partir de mesures de force en surface. Soit Ω une partie de la terre, ρ la densité. La force due à la gravité en un point $x \notin \Omega$ est donné par la loi de Newton (G est la constante de gravitation) :

$$\phi(x) = \frac{G}{4\pi}\int_\Omega \frac{\rho(y)}{\|x-y\|}^2\, dy \tag{2.24}$$

Nous nous bornerons cette fois à un modèle unidimensionnel, inspiré de Kirch [43].

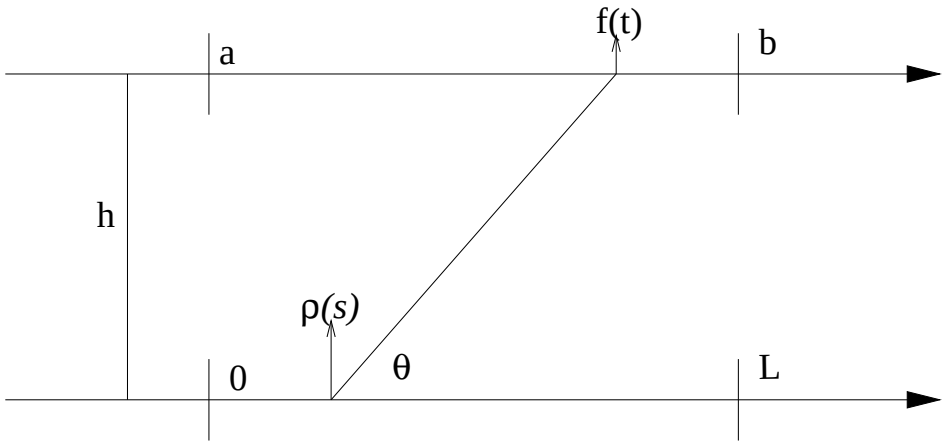

FIGURE 2.2 – Géométrie de l'expérience

On veut déterminer la répartition $\rho(x)$, $0 \leq x \leq 1$, de la densité de masse d'une anomalie localisée à une profondeur h, à partir de mesures de la force verticale $f(x)$.

La contribution à $f(t)$ due au segment ds de l'axe des s est

$$G\frac{\sin\theta}{r^2}\rho(s)\,ds$$

ou $r = \sqrt{h^2+(s-t)^2}$. Avec $\sin\theta = h/r$, il vient

$$f(t) = G\int_0^L \frac{h}{(h^2+(t-s)^2)^{3/2}}\rho(s)\,ds \quad a \leq t \leq b \tag{2.25}$$

Le problème direct, qui consiste à calculer la force connaissant la répartition de densité est cette fois simplement l'évaluation d'une intégrale. Le problème inverse est la résolution d'une équation intégrale de première espèce. Il s'agit d'un problème analogue à la différentiation vue au paragraphe 1.1, mais avec un noyau intégral (la fonction $(s,t) \mapsto \dfrac{h}{(h^2+(t-s)^2)^{3/2}}$) général. Nous étudierons ce type de problème plus en détail au chapitre 3.1. □

Exemple 2.13 (Lancer de rayon)**.**
Il s'agit d'une variante du modèle sismique, dans laquelle on considère un modèle de propagation simplifié. Nous suivrons la présentation du livre [29].

Nous supposerons encore que le modèle géologique est stratifié, et nous ferons de plus l'hypothèse que la fonction $z \mapsto c(z)$ est croissante sur l'intervalle $[0,Z]$. On considère (c'est le principe de Fermat) que les ondes voyagent selon des *rayons*, et l'on suit le rayon reliant une source (en surface) à un récepteur (également en surface) fixé, comme sur la figure 2.3. ce rayon, après s'être (éventuellement) retourné à la profondeur $Z(p)$, émerge au point (Xp). On mesure l'instant d'émergence de ce rayon, noté $T(X)$.

On démontre (c'est toujours le principe de Fermat, voir [1] et [29]) que la distance entre la source et le récepteur est donnée en fonction du *paramètre du rayon* $p = \sin i/c(z)$ (qui est une constante) par :

$$X(p) = 2\int_0^{Z(p)} \frac{c(z)\,p}{\sqrt{1-c(z)^2\,p^2}}\,dz, \tag{2.26}$$

de même, le temps de parcours est donné par l'intégrale

$$T(p) = 2\int_0^{Z(p)} \frac{1}{c(z)} \frac{1}{\sqrt{1-c(z)^2\,p^2}}\,dz, \tag{2.27}$$

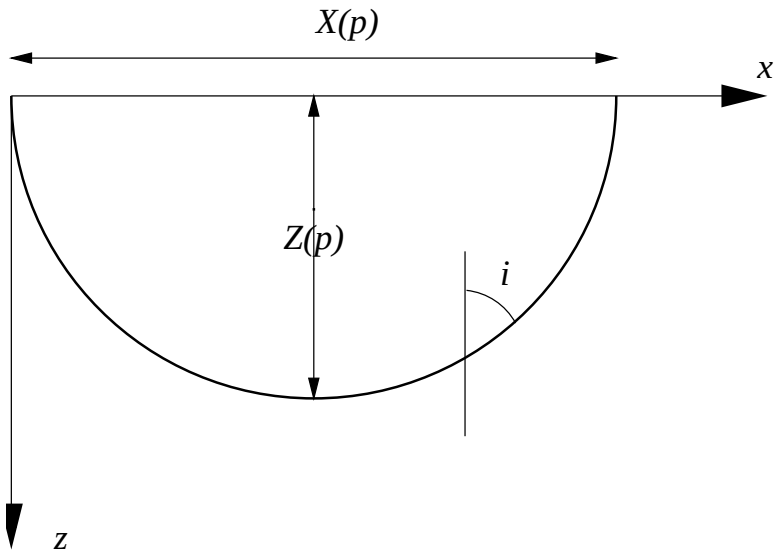

FIGURE 2.3 – Rayon dans une terre stratifiée

Nous avons donc obtenu X et T en fonction de p. Rien ne prouve que ces fonctions soient monotones (il existe d'ailleurs des exemples). On introduit alors la fonction de temps de retard, définie par :

$$\tau(p) = T(p) - pX(p) = 2\int_0^{Z(p)} \sqrt{c(z)^{-2} - p^2}\,dz,$$

qui est monotone. On calcule alors $\tau'(p)$ de deux façons différentes :

$$\tau'(p) = T'(p) - pX'(p) - X(p) = -2\int_0^{Z(p)} \frac{p}{\sqrt{c(z)^{-2} - p^2}}\,dz = -X(p)$$

de sorte que nous obtenons $p = dT/dX$. Ceci prouve que p est mesurable (c'est la pente de la courbe $X \mapsto T$).

Le changement de variable $\xi = c(z)^{-1}$ (fonction inverse qui existe sous l'hypothèse que $z \mapsto c(z)$ est une fonction monotone) donne :

$$\tau(p) = 2\int_{c_0^{-1}}^{p} \frac{\xi f(\xi)}{\sqrt{\xi^2 - p^2}}\,d\xi \tag{2.28}$$

où f est la fonction inverse de la fonction $z \mapsto c^{-1}(z)$.

Cette équation est encore une équation intégrale d'Abel, pour laquelle nous avons déjà vu qu'une formule d'inversion simple existe. On obtient donc

$$f(\xi) = -\frac{1}{\pi}\int_{c_0^{-2}}^{\xi} \frac{\tau'(p)}{\sqrt{\xi^2 - p^2}}\,dp \tag{2.29}$$

La formule 2.29 montre que la résolution du problème inverse passe le calcul de la dérivée de la fonction f. Nous retrouvons encore une fois la différentiation au coeur d'un problème inverse. Si les données $T(X)$ ne sont connues qu'approximativement, le calcul de la dérivée de la fonction inverse de la dérivée de $T(X)$ sera très imprécis. Le problème inverse est donc susceptible d'instabilité. □

Exemple 2.14 (Scattering inverse)**.**
Il s'agit d'un problème analogue à celui du paragraphe 2.11, mais le but est cette fois de retrouver la forme de l'obstacle diffractant. Ce problème intervient, par exemple, dans la détection par radar ou par sonar, ce qui motive de nombreuses recherches. On envoie une onde acoustique de fréquence donnée sur un obstacle (représenté par un ouvert Ω), on mesure le signal réfléchi, et l'on veut retrouver la frontière de l'obstacle. Les équations sont les mêmes qu'au paragraphe 2.11, à ceci près qu'elles sont posées sur l'extérieur de l'obstacle, et qu'il faut donc ajouter une condition aux limites sur le bord de celui-ci (qui peut être une condition de type Dirichlet, Neumann ou mixte). On peut écrire des équation intégrales analogues à (2.23), et on montre que l'onde diffractée a la représentation asymptotique

$$u_s(x) = \frac{e^{ikx}}{\sqrt{|x|}}\left[u_\infty\left(\frac{x}{|x|}\right) + O\left(\frac{1}{x}\right)\right] \quad \text{quand } |x| \to \infty \tag{2.30}$$

La fonction u_∞ est appelée le *champ lointain*, et c'est ce qui est mesuré. Le problème inverse est donc de retrouver le bord de l'obstacle connaissant une mesure du champ lointain. Ce problème est non-linéaire. On se rapportera ici encore à [18] pour en savoir plus. □

Exemple 2.15 (Problèmes spectraux inverses)**.**
La motivation de ces problèmes vient d'un article célèbre de Marc Kac [40] ń Can One Hear the Shape of a Drum ? ż La question mathématique que pose cet article est de retrouver la forme (la frontière) d'un ouvert connaissant ses fréquences de vibration (c'est-à-dire les fréquences propres de l'opérateur Laplacien sur cet ouvert). On sait que la réponse est en général négative, c'est-à-dire que l'on sait construire des ouverts isospectraux (possédant le même spectre), mais non isométriques. Des exemples d'ouverts polygonaux ont été construits par T. Driscoll [23]. Une introduction abordable se trouve dans le livre de Kirch [43]. □

On trouvera d'autres exemples dans les livres de Kirch [43], d'Engl et al. [25], ou d'Isakov [39].

Deuxième partie

Problèmes linéaires

Chapitre 3

Opérateurs intégraux et équations intégrales

Ce chapitre présente une brève introduction aux opérateurs intégraux, ainsi qu'aux équations intégrales de première espèce. Ces dernières fournissent le principal exemple de problèmes inverses linéaires.

Nous présenterons tout d'abord les principales propriétés des opérateurs intégraux, dans le cadre L^2, en particulier le fait que si le noyau est de carré intégrable, l'opérateur associé est compact. Nous étudierons ensuite l'approximation numérique des équations de première espèce par la méthode de Galerkin et la méthode de quadrature–collocation.

Nous ne traîterons pas de la convergence de ces approximations dans ce chapitre : le cadre naturel pour cela est l'étude des propriétés régularisantes des méthodes de projection, que nous verrons au paragraphe 6.2. Pour plus de détails, on pourra se reporter aux livres [43] et surtout [44].

Enfin, cette étude sera complétée au paragraphe 4.3 par l'étude des propriétés spectrales des opérateurs.

3.1 Définition et premières propriétés

Théorème 3.1. *Soit K une fonction de l'espace $L^2(]c,d[\times]a,b[)$. L'opérateur*

$$Au(t) = \int_a^b K(t,s)u(s)\,ds, \quad t \in]a,b[\tag{3.1}$$

est bien défini en tant qu'opérateur de $L^2(a,b)$ dans $L^2(c,d)$.

Preuve. La linéarité est évidente, seule la continuité (et le fait que Ku est élément de $L^2(c,d)$ si $u \in L^2(a,b)$, qui en sera une conséquence immédiate) sont à démontrer.

Bien entendu, nous voulons majorer

$$\int_c^d |Au(t)|^2\,dt = \int_c^d \left(\int_a^b K(t,s)u(s)\,ds \right)^2 dt. \tag{3.2}$$

Par l'ínégalité de Cauchy-Schwarz, il vient :

$$\int_c^d |Au(t)|^2\,dt \leq \int_c^d \left(\int_a^b |K(t,s)|^2\,ds \right) \left(\int_a^b |u(s)|^2\,ds \right) dt \leq M^2 \int_a^b |u(s)|^2\,ds. \tag{3.3}$$

(avec $M^2 = \iint_{]c,d[\times]a,b[} |K(t,s)|^2\,ds\,dt < \infty$, puisque $K \in L^2(]c,d[\times]a,b[)$), ce qui prouve que (3.1) définit bien un opérateur continu de $L^2(a,b)$ dans $L^2(c,d)$, et montre au passage que sa norme est majorée par M. □

Définition 3.1. L'opérateur A défini au théorème 3.1 s'apelle l'opérateur intégral de noyau K.

Exemple 3.1 (Opérateurs de Volterra)**.**
Il s'agit d'opérateurs de la forme

$$Au(t) = \int_0^t k(t,s)u(s)\,ds, \quad \text{pour } t \in [0,1]$$

avec $k \in L^2([0,1]\times[0.1])$.

Pour nous placer dans le cadre du théorème, nous devons tout d'abord nous ramener à un intervalle fixe. Pour cela, nous utilisons une astuce classique en théorie de l'intégration : nous introduisons la fonction caractéristique de l'intervalle $[0,t]$, notée $\chi_{[0,t]}$. La définition de A devient :

$$Au(t) = \int_0^1 \chi_{[0,t]}(s)k(t,s)u(s)\,ds.$$

Pour obtenir une définition plus symétrique, nous remarquons que $0 \le s \le t \le 1$ est équivalent à $(s,t) \in T$, où T est le triangle inférieur du carré unité $[0,1]\times[0,1]$. On a donc :

$$Au(t) = \int_0^1 \chi_T(s,t)k(s,t)u(s)\,ds.$$

Nous devons donc vérifier que la fonction $(s,t) \to \chi_T(s,t)k(s,t) \in L^2([0,1]\times[0,1])$. Or il est clair que

$$\int_T |k(s,t)|^2\,ds\,dt \le \int_{[0,1]\times[0,1]} |k(s,t)|\,ds\,dt < \infty.$$

Ce résultat s'applique ne particulier à l'exemple 1.1 : il suffit de prendre $k(s,t) = 1, \forall (s,t)$.

□

Une classe particulièrement simple d'opérateurs intégraux est constituée des opérateurs à noyau dits *dégénérés*, c'est-à-dire de la forme :

$$K(t,s) = \sum_{j=1}^{p} a_j(t)b_j(s).$$

Les opérateurs correspondants sont de *rang fini* :

Proposition 3.1. *Soit A un opérateur intégral à noyau dégénéré. L'image de A est de dimension finie.*

Preuve. Nous allons montrer que l'image de A est engndrée par les fonctions $a_1,\ldots,a_p$. Sa dimension est donc majorée par p.

Soit $u \in L^2(a,b)$.

$$Au(t) = \int_a^b \sum_{j=1}^{p} a_j(t)b_j(s)u(s)\,ds = \sum_{j=1}^{p} \left(\int_a^b b_j(s)u(s)\,ds\right) a_j(t),$$

qui est bien un élément de l'espace vectoriel vect$\{a_1,\ldots,a_p\}$. □

Ce résultat sera utilisé plus loin pour démontrer la compacité de l'opérateur A.

Proposition 3.2. *Soit A l'opérateur intégrable de noyau K. A^* est l'opérateur intégral de noyau K^*, avec*

$$K^*(t,s) = K(s,t) \tag{3.4}$$

Preuve. Il suffit de partir de la définition. Soit $u \in L^2(a,b), v \in L^2(c,d)$.

$$(Au,v) = \int_c^d \left(\int_a^b K(t,s)u(s)\,ds \right) v(t)\,dt = \int_{]a,b[\times]c,d[} K(t,s)u(s)v(s)\,ds\,dt$$

par le théorème de Fubini. En échangeant encore l'ordre d'intégration, il vient :

$$(Au,v) = \int_a^b \left(\int_c^d K(t,s)v(t)\,dt \right) u(s)\,ds = (u,A^*v)$$

d'après la définition de l'adjoint. En permuttant le nom des varaiables, on obtient la définition (3.4). □

Corollaire 3.1. *L'opérateur intégral A de noyau K est auto-adjoint si, et seulement si, le noyau est symétrique :*

$$K(s,t) = K(t,s), \quad \forall (s,t) \in [a,b] \times [c,d]. \tag{3.5}$$

Proposition 3.3. *Soient A_1 et A_2 les opérateurs intégraux de noyaux respectifs $K_1 \in L^2(]a,b[\times]c,d[)$ et $K_2 \in L^2(]c,d[\times]e,f[)$. Le composé $A_2A_1 \in \mathcal{L}(L^2(a,b),L^2(e,f))$ est un opérateur intégral de noyau*

$$K(t,s) = \int_c^d K_2(t,r)K_1(r,s)\,dr. \tag{3.6}$$

Preuve. Ici encore, il suffit de suivre les définitions :

$$\begin{aligned}(A_2A_1)(u)(t) &= \int_c^d K_2(t,r) \left(\int_a^b K_1(r,s)u(s)\,ds \right) dr \\ &= \int_a^b \left(\int_c^d K_2(t,r)K_1(r,s)\,dr \right) u(s)\,ds,\end{aligned}$$

toujours par le théorème de Fubini. Il reste à vérifier que ce noyau est bien de carré intégrable. Tout d'abord, par l'inégalité de Cauchy-Schwarz,

$$\left(\int_c^d K_2(t,r)K_1(r,s)\,dr \right)^2 \leq \int_c^d |K_2(t,r)|^2\,dr \int_c^d |K_1(r,s)|^2\,dr,$$

puis par une double application du théorème de Fubini,

$$\int_{[a,]\times[e,f]} \left(\int_c^d K_2(t,r)K_1(r,s)\,dr \right)^2 \leq \int_{[a,]\times[c,d]} |K_1(r,s)|^2\,dr\,ds \int_{[c,d]\times[e,f]} |K_2(t,r)|^2\,dr\,dt.$$

□

Il est habituel de classer les équations intégrales que l'on peut associer à l'opérateur intégral A en deux catégories :

Équations de première espèce Il s'agit de l'équation

$$Au = f, \quad \text{où} f \in L^2(c,d) \text{ est donnée;} \tag{3.7}$$

Équations de seconde espèce Il s'agit de l'équation

$$u - Au = f, \quad \text{où} f \in L^2(c,d) \text{ est donnée;} \tag{3.8}$$

Cette distinction est justifiée par les propriétés très différentes de ces deux types d'équation pour des noyaux de carré intégrable. Les équations de première espèce, auxquelles nous concentrerons notre attention, conduisent à des problèmes mal posés. En revanche, celles de seconde espèce ont, en général, une solution unique (cela relève de l'alternative de Fredholm, énoncée au théorème A.6). Cette distinction est liée à la *compacité* de l'opérateur. On a effet le résultat suivant :

Théorème 3.2. *Soit $K \in L^2(]a,b[\times]c,d[)$. L'opérateur intégral A de noyau K est compact de $L^2(a,b)$ dans $L^2(c,d)$.*

***Preuve** (partielle, peut être omise).* Nous admettrons qu'il est possible d'approcher le noyau K dans $L^2(]a,b[\times]c,d[)$ par une suite de noyaux $(K_n)_{n\in\mathbf{N}}$ dégénérés [7]. Notons A_n l'opérateur intégral de noyau K_n. D'après la proposition 3.1, A_n est un opérateur de rang fini. Montrons que la suite A_n converge vers A.

On a

$$(A_n - A)u = \int_a^b (K_n(t,s) - K(t,s))u(s)\,ds$$

et

$$\begin{aligned}
\|(A_n - A)u\|_F^2 &= \int_c^d \left(\int_a^b (K_n(t,s) - K(t,s))u(s)\,ds \right)^2 dt \\
&\leq \left(\iint_{]a,b\times]c,d[} |K_n(t,s) - K(t,s)|^2 \, ds\,dt \right) \|u\|_E^2 \\
&= \|K_n - K\|_{L^2(]a,b[\times]c,d[)} \|u\|_E^2 .
\end{aligned}$$

Le premier terme tend vers 0, d'après le choix de K_n, ce qui achève la démonstration. □

En rapprochant ce résultat du corollaire A.3, nous voyons que les équations intégrales de première espèce donenront toujours lieu à des problèmes mal posés. Nous reviendrons sur ce point quand nous aurons introduit la décomposition en valeurs singulières au chapitre 4.

Une autre façon de comprendre le caractère mal posé de ces équations passe par le lemme de Riemann–Lebesgue, qui affirme que

$$\int_0^p iK(t,s)\sin(ns)\,ds \xrightarrow[n\to\infty]{}, \text{ dans } L^2(0,1)$$

pour tout noyau $K \in L^2([0,1]times[0,1])$ (il s'agit d'une résultat sur les séries de Fourier). Ainsi, des perturbations *haute fréquence* (c'est le cas de $\sin(ns)$ pour n grand), sont annihilées par l'opération d'intégration avec un noyau régulier. Une telle perturabtion est donc ń indétectable ż du point de vue de la résolution de l'équation intégrale (3.7).

3.2 Discrétisation des équations intégrales

Nous allons nous borner à étudier brièvement deux méthodes pour discrétiser une équation intégrale : la méthode de quadrature – collocation et la méthode de Galerkin.

3.2.1 Discrétisation par quadrature

Commençons par quelques rappels sur les formules de quadrature, qui consistent à approcher une intégrale

$$I = \int_a^b \varphi(s)\,ds$$

par une somme pondérée des valeurs de φ, en des points appelés *noeuds*.

Dans les exemples qui suivent, n est un entier. Nous nous contenterons de présenter les formules les plus simples : formule des rectangles, des trapèzes et de Simpson.

Exemple 3.2 (Formule des rectangles)**.**
Les noeuds de la formule des rectangles sont les points $s_{j-1/2} = s_j - h/2,\ j = 1,\ldots,n$, et les poids sont tous égaux à h. On approche l'intégrale (l'aire sous la courbe) par la somme des aires des rectangles

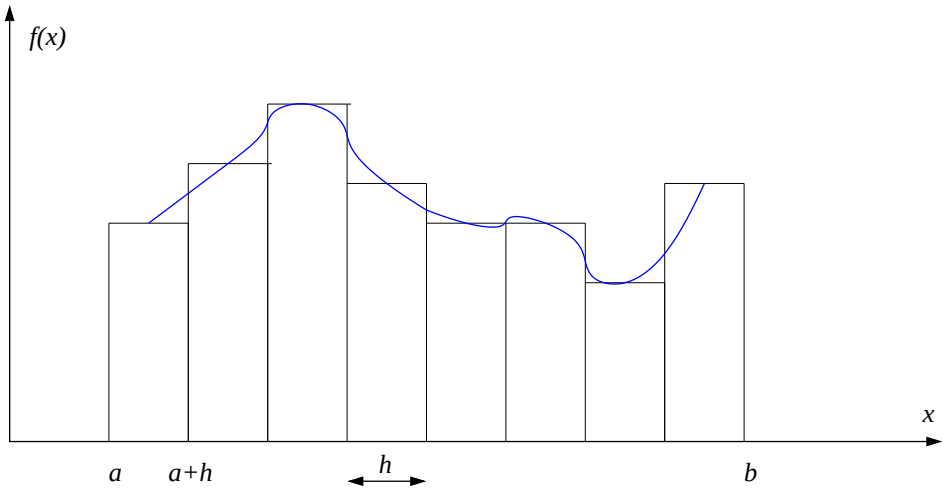

FIGURE 3.1 – Formule des rectangles

de largeur h, et de hauteur est determinée par $\varphi(s_{j-1/2})$:

$$I^R = h\big(\varphi(s_{1/2}) + \cdots + \varphi(s_{j-1/2}) + \cdots + \varphi(s_{n-1/2})\big).$$

Exemple 3.3 (Formule des trapèzes)**.**
Les noeuds de la formule des trapèzes sont les points $s_j,\ j = 1,\ldots,n$, où $h = (b-a)/n$. Les poids sont $w_j = h, j = 2,\ldots,n-1$ ainsi que $w_1 = w_n = h/2$. La formule de quadrature s'obtient en remplaçant

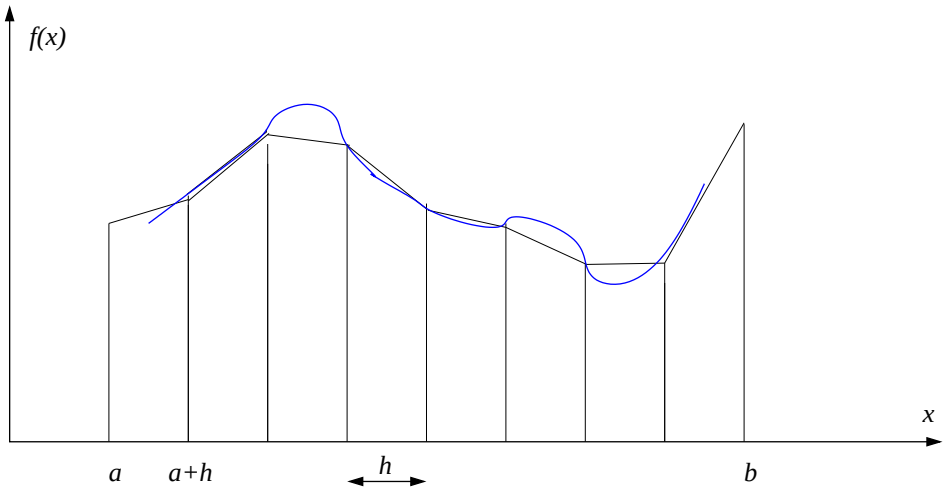

FIGURE 3.2 – Formule des rectangles

l'intégrale pa la somme des aires des trapèzes déterminés par les valeurs de la fonction aux points s_j et s_{j+1}, ce qui donne :

$$I^T = h\left(\frac{1}{2}\varphi(a) + \varphi(s_1) + \cdots + \varphi(s_{n-1}) + \frac{1}{2}\varphi(b)\right).$$

Exemple 3.4 (Formule de Simpson)**.**
On suppose cette fois que n est pair. Les noeuds de la formule de Simpson sont les mêmes que pour la formule des trapèzes. Les poids sont $w_{2j} = 2h/3$, $j = 1,\ldots,n/2-1$, $w_{2j+1} = 4h/3$, $j = 0,\ldots,n/2-1$, et $w_0 = w_n = h/3$. La formule de quadrature de Simpson est :

$$I^S = \frac{h}{3}\Big(\varphi(a) + 4\varphi(s_1) + 2\varphi(s_2) + \cdots + 2\varphi(s_{n-2}) + 4\varphi(s_{n-1}) + \varphi(b)\Big)$$

Sous des hypothèses convenables, on peut majorer l'erreur commise en remplaçant l'intégrale par les différentes formules de quadrature.

Proposition 3.4. *On a les majorations suivantes :*
Formule des rectangles *Sous l'hypothèse* $\varphi \in C^2[a,b]$,

$$\left|I - I^R\right| \leq \frac{(b-a)h^2}{24}\left\|\varphi''\right\|_\infty. \tag{3.9}$$

Formule des trapèzes *Sous l'hypothèse* $\varphi \in C^2[a,b]$,

$$\left|I - I^T\right| \leq \frac{(b-a)h^2}{12}\left\|\varphi''\right\|_\infty. \tag{3.10}$$

Formule de Simpson *Sous l'hypothèse* $\varphi \in C^4[a,b]$,

$$\left|I - I^S\right| \leq \frac{(b-a)h^4}{180}\left\|\varphi^{(4)}\right\|_\infty. \tag{3.11}$$

Preuve. Nous la ferons pour la formule des rectangles, les autres cas sont similaires (voir [44]).

Plaçons nous d'abord sur l'intervalle $[s_j, s_{j+1}]$, et écrivons la formule de Taylor à l'ordre 2, autour du point $s_j + h/2$:

$$\varphi(s) = \varphi(s_j + h/2) + (s - s_j - h/2)\varphi'(s_j + h/2) + (s - s_j - h/2)^2/2\varphi''(\xi_h), \quad \xi_h \in [s_j, s_{j+1}].$$

Intégrons cette égalité sur $[s_j, s_{j+1}]$, en remarquant que l'intégrale du terme linéaire est nulle par symétrie :

$$\left|\int_{s_j}^{s_{j+1}} \varphi(s)\,ds - h\varphi(s_{j+1/2})\right| \leq \sup_{[s_j,s_{j+1}]}\left|\varphi''\right| \int_{sj}^{s_{j+1}} (s - h/2)^2/2\,ds = \sup_{[s_j,s_{j+1}]}\left|\varphi''\right| h^3/24.$$

On obtient l'inégalité (3.9) en ajoutant les majorations ci-dessus sur tous les intervalles $[s_j, s_{j+1}]$. □

De façon générale, en notant s_j et w_j pour $j = 1,\ldots,n$ respectivement les noeuds et les poids, une formule de quadrature s'écrit :

$$I^a = \sum_{j=1}^{n} w_j\,\varphi(s_j). \tag{3.12}$$

L'application d'une formule de quadrature à une équation intégrale (de première espèce) passe par une méthode de collocation. On exprime que l'équation intégrale est vérifiée en un nombre finis de points $t_i, i = 1, \ldots, m$,

$$\int_a^b K(t_i, s)\, u(s)\, ds = f(t_i), \quad i = 1, \ldots, m, \tag{3.13}$$

et on remplace l'intégrale ci-dessus par la formule de quadrature choisie dans l'équation (3.12). On obtient le système d'équation (rectangulaire) :

$$\sum_{j=1}^{n} w_j K(t_i, s_j) u_j = f(t_i), \quad i = 1, \ldots, m. \tag{3.14}$$

Il s'agit là d'un système linéaire $Ax = b$, comme on le voit en posant

$$A_{hij} = w_j K(t_i, s_j),\ b_i = f(t_i,),\ x_j = u_j, \quad \text{pour } j = 1, \ldots, n,\ i = 1, \ldots, m.$$

Dans le cas où $m > n$, on obtient un système sur-determiné. Il n'y a en tout cas aucune raison de prendre $m = n$, et le système (3.14) devra en général être résolu au sens des moindre carrés. Nous reviendrons sur ces problèmes au chapitre 5.

Exemple 3.5.
Dans le cas de l'opérateur intégral

$$Au(t) = \int_0^1 u(s)\, ds, \quad t \in [0, 1],$$

la matrice A_h est triangulaire inférieure, avec $A_{ij} = h$, pour $j \leq 1$, c'est-à-dire que le système d'équation (3.14) est :

$$h \sum_{j=1} u_j = f_i, \quad i = 1, \ldots, n.$$

On voit facilement par récurrence que la solution s'écrit :

$$u_j = \frac{f_j - f_{j-1}}{h}, \quad j = 2, \cdots, n$$

avec $u_1 = f_1/h$, ce qui n'est pas surprenant si l'on reconnait une discrétisation de la dérivée première. L'inverse de A est bien une approximation de l'inverse de A !

3.2.2 Discrétisation par la méthode de Galerkin

Il s'agit cette fois d'une méthode de projection. On approche l'espace $L^2(a,b)$ (resp. $L^2(c,d)$) par une suite de sous-espaces de dimension finie E_n (resp. F_m) (on supposera $\dim E_n = n$, $\dim F_m = m$). On projette alors l'équation (3.7) sur F_n, c'est-à-dire que l'on cherche $u_n \in E_n$ solution de l'équation

$$(Au_n, v_n) = (f, v_m) \quad \forall v_m \in F_m \tag{3.15}$$

Cette équation est l'équation de Galerkin pour u_n. Afin d'expliciter cette équation, introduisons une base $\{e_1, \ldots, e_n\}$ dans l'espace E_n (resp. une base $\{f_1, \ldots, f_m\}$ dans l'espace F_m). Développons u_n dans cette base, sous la forme

$$u_n = \sum_{j=1}^{n} x_j e_j$$

et prenons $v_m = f_i$ dans l'équation de Galerkin (3.15). Il vient :

$$\sum_{j=1}^{n} (Ae_j, f_i) x_j = (f, f_i), \quad i = 1, \ldots, m. \tag{3.16}$$

Il s'agit encore d'un système (rectangulaire) d'équations, susceptible d'un traitement numérique.

Notons une différence pratique entre la méthode de Galerkin et la méthode de quadrature–collocation du paragraphe 3.2.1. Dans la méthode de quadrature–collocation, les éléments de la matrice et du second membre s'obtiennent simplement comme

$$A_{ij} = w_j K(t_i, s_j), \quad b_i = f(t_i) \tag{3.17}$$

alors que les éléments de la matrice de la méthode de Galerkin sont des intégrales doubles (simples pour le vecteur) :

$$A_{ij} = \iint_{]a,b[\times]c,d[} K(t,s)\, f_i(t)\, e_j(s)\, ds\, dt, \quad b_i = \int_c^d f(t) f_j(t)\, dt. \tag{3.18}$$

La méthode de Galerkin entrainera donc un sur-coût important par rapport à la méthode de quadrature–collocation. En contrepartie, la méthode de galerkin possède de meilleures propriétés de convergence (ordre plus élevé dans d'autres normes, voir [43]).

Exemple 3.6.
Un exemple simple, et utile en pratique, est fourni par le choix de fonctions constantes par morceaux pour les deux sous-espaces d'approximation.

Plus précisemment, posons $h_s = (b-a)/n$, $h_t = (d-c)/m$, et subdivisons les intervalles $]a,b[$ et $]c,d[$ en n et m intervalles de taille h_s et h_t respectivement. Notons $I_j^s =]a + (j-1)h_s, a + jh_s$ et $I_i^t =]c + (i-1)h_t, c + ih_t$ ces intervalles, et définissons les fonctions de base par :

$$e_j(s) = \begin{cases} h_s^{-1/2}, & s \in I_j^s \\ 0 & \text{sinon,} \end{cases} \quad j = 1, \ldots, n, \tag{3.19}$$

et

$$f_i(s) = \begin{cases} h_t^{-1/2}, & s \in I_i^t \\ 0 & \text{sinon,} \end{cases} \quad i = 1, \ldots, m. \tag{3.20}$$

Les éléments de la matrice et du second membre se calculent alors par

$$A_{ij} = \frac{1}{\sqrt{h_s h_t}} \int_{I_i^t} \int_{I_j^s} K(t,s)\, dt\, ds, \; b_i = \int_{I_i^t} f(t)\, dt \tag{3.21}$$

Terminons par un exemple numérique pour mettre en évidence les difficultés de résolution des équations de première espèce.

Exemple 3.7.
Reprenons l'exemple de la prospection magnétique, vu au chapitre 1 (exemple 2.12). Nous prenons l'équation (2.25), avec $[a,b] = [0,1]$,$L = 2$, $h = 0.25$ et différentes valeurs de n.

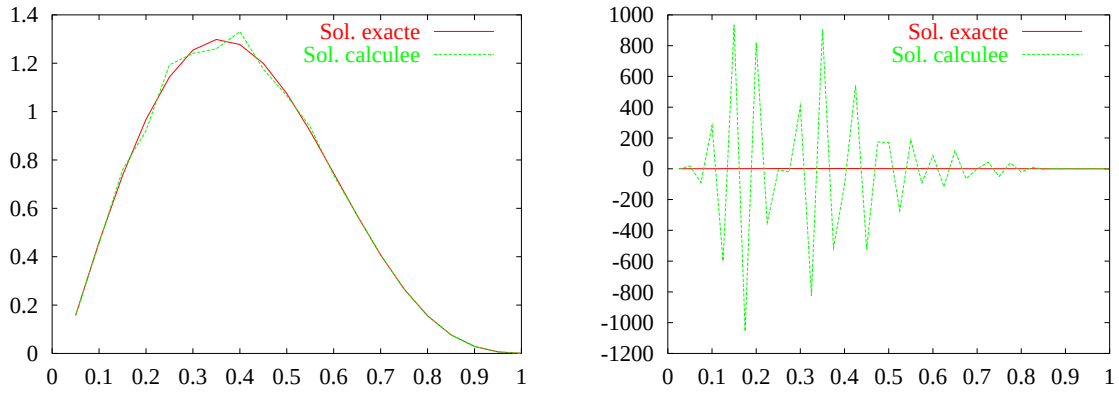

FIGURE 3.3 – Prospection géomagnétique. Comparaison de la solution exacte avec la solution calculée. À gauche $n = 20$, à droite $n = 40$

La solution exacte est $f(t) = \sin(\pi t) + 1/2\sin(2\pi t)$, et le second membre est calculé en conséquence. L'équation est discrétisée en utilisant la méthode de quadrature collocation vue au paragraphe 3.2.1, avec la formule des rectangles. La matrice obtenue est symétrique et définie positive.

Nous représentons sur la figure 3.3 les résultats correspondants aux valeurs de n égales à 20, puis à 40. Noter la différence d'échelle sur l'axe vertical entre les deux figures. Comme on peut le constater, si les résultats sont acceptables pour $n = 20$, ils sont catastrophiques pour $n = 40$. Cela est évidement du au caractère mal posé du problème continu. Plus n augmente, plus la discrétisation reproduit ce caractère mal posé, ce qui se traduit numériquement par un conditionnement énorme. Le conditionnement de la matrice de ce problème pour différentes valeurs de n est donné dans le tableau 3.1.

n	10	20	40	60	80	100
cond A	$6.3\,10^{7}$	$4.1\,10^{12}$	$7.5\,10^{19}$	$2.8\,10^{20}$	$1.5\,10^{20}$	$1.8\,10^{21}$

TABLE 3.1 – Prospection géomagnétique. Conditionnement de la matrice en fonction de n

Chapitre 4

Problèmes de moindres carrés linéaires – Décomposition en valeurs singulières

Nous étudions dans ce chapitre les principales propriétés des problèmes inverses linéaires. Nous nous placerons dans le cadre des espaces de Hilbert, pour que les résultats s'appliquent (par exemple) aux équations intégrales de première espèce, mais nous indiquerons les simplifications qui interviennent en dimension finie. Nous introduirons ensuite l'outil fondamental que constitue la *décomposition en valeurs singulières*. Cette fois, pour des raisons de simplicité d'exposition, nous traiterons d'abord le cas des matrices, avant de présenter les opérateurs, bien que le premier soit un cas particulier du second. Enfin, nous montrerons comment la décomposition en valeurs singulières permet d'analyser les problèmes de moindre de carrés.

Ce chapitre est uniquement concerné par les aspects mathématiques. Nous nous intéresserons aux méthodes numériques au chapitre suivant.

Dans tout ce chapitre nous désignerons par A un opérateur linéaire continu d'un espace de Hilbert E dans un espace de Hilbert F : $A \in \mathcal{L}(E,F)$.

4.1 Propriétés mathématiques des problèmes de moindres carrés

Étant donné $\hat{z} \in F$, nous cherchons $\hat{x} \in E$ solution de :

$$A\hat{x} = \hat{z}. \tag{4.1}$$

Revenons dans ce cas particulier sur la discussion du chapitre 1 concernant les problèmes bien et mal posés

- l'opérateur A peut ne pas être surjectif ;
- il peut ne pas être injectif ;
- si un inverse existe, il peut ne pas être continu.

Comme nous l'avons déjà dit, la première difficulté n'est pas sérieuse : il ń suffit ż de se restreindre à $\mathrm{Im}A$. La seconde est plus gênante : il faut pouvoir sélectionner, parmi plusieurs solutions, celle qui est appropriée au problème. La dernière, fondamentale pour les applications, est lié au caractère fermé ou non de $\mathrm{Im}A$:

Théorème 4.1. *Soit $A \in \mathcal{L}(E,F)$, E et F deux espaces de Hilbert. Supposons que A soit injectif, et notons $A^{-1} : \mathrm{Im}A \to E$ l'inverse de A. On a :*

$$\mathrm{Im}A \textit{ fermé } \iff A^{-1} \textit{ est continu}.$$

Preuve. $\Rightarrow$ Dans ce cas $W = \mathrm{Im}A$ est un espace de Hilbert (il est immédiat que W est un espace préhilbertien, et il est complet parce qu'il est fermé). L'opérateur $\tilde{A} : E \to W$, $\tilde{A}u = Au$ pour $u \in E$ est un opérateur linéaire continu et bijectif. Une conséquence classique du théorème de l'application ouverte (théorème A.2) est que $\tilde{A}^{-1}$ est continu. Il en est donc de même pour A^{-1}.

$\Leftarrow$ Puisque A^{-1} est continu, et que $E = \mathrm{Im}A^{-1}$ est fermé, $\mathrm{Im}A = (A^{-1})^{-1}(E)$ est fermé dans F.

□

Pour les situations que nous considérons dans ce cours, l'opérateur pourra ou non être injectif, mais la situation générale sera que ImA n'est pas *pas* fermée (le corollaire A.3 montre que si A est compact, A n'a pas d'inverse continu, et dans ce cas *ImaA* ne sera pas fermé).

Le problème (4.1) n'a de solution que pour les seconds membres dans l'image de A. Comme nous venons de le voir, cette condition peut-être trop restrictive. Nous reviendrons sur ce point après introduit la décomposition en valeurs singulières, avec la condition de Picard (théorème 4.6). Nous cherchons donc une autre formulation du problème original, qui permette d'étendre la notion de solution à un sous-espace plus grand.

Nous proposons une formulation comme un problème de *moindres carrés* : nous remplaçons donc (4.1) par :

$$\min_{x \in E} \frac{1}{2} \|Ax - \hat{z}\|_F^2 \tag{4.2}$$

Nous allons voir que ce problème est équivalent à une équation linéaire, mais pour un opérateur différent de A.

Théorème 4.2. *Soit $A \in \mathcal{L}(E,F)$, E, F deux espaces de Hilbert, et soit $\hat{z} \in F$. Un élément $\hat{x} \in E$ est une solution de (4.2) si et seulement si*

$$A^*A\hat{x} = A^*\hat{z} \tag{4.3}$$

Preuve. On peut obtenir facilement (4.3) en calculant le gradient de la fonctionnelle $x \to \frac{1}{2}\|Ax - z\|^2$. Nous laissons cette méthode en exercice au lecteur. Nous présenterons plutôt l'argument élémentaire suivant, emprunté à Björck [9].

Soit x vérifiant (4.3). On a pour tout $y \in E$:

$$\hat{z} - Ay = \hat{z} - Ax + A(x - y).$$

L'équation normale (4.3) implique que les deux termes de la somme sont orthogonaux (le résidu $\hat{z} - Ax$ est orthogonal a l'imae de A). Le théorème de Pythagore implique :

$$\|\hat{z} - Ay\|_F^2 = \|\hat{z} - Ax\|_F^2 + \|A(x - y)\|_F^2 \geq \|\hat{z} - Ax\|_F^2 .$$

x est donc bien solution de (4.2).

Réciproquement, soit x tel que $A^*(\hat{z} - Ax) = w \neq 0$. Choisissons $y = x + \varepsilon w$, avec $\varepsilon > 0$. On a alors :

$$\|\hat{z} - Ay\|_F^2 = (\hat{z} - Ay, \hat{z} - Ay) = \|\hat{z} - Ax\|_F^2 - 2\varepsilon(\hat{z} - Ax, w) + \|Aw\|_F^2 < \|\hat{z} - Ax\|_F^2$$

si ε est suffisamment petit. x n'est donc pas solution de (4.2). □

Remarque 4.1. L'équation normale (4.3) se réécrit :

$$A^* (A\hat{x} - \hat{z}) = 0,$$

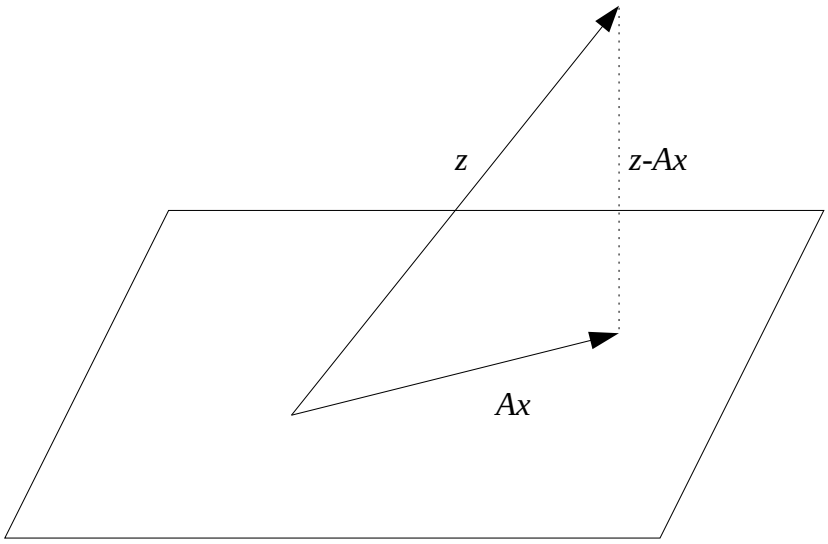

FIGURE 4.1 – Illustration géométrique des moindres carrés

ce qui exprime simplement que le résidu $\hat{z} - A\hat{x}$ est dans le noyau de A^*, c'est-à-dire orthogonal à (la fermeture de) l'image de A (voir proposition A.5). Ceci conduit à l'illustration géométrique bien connue :

La solution du problème de moindres carrés est telle que Ax est la projection de $\hat{z}$ sur l'image de A.

Notons que nous n'avons pour l'instant évoqué ni l'existence, ni l'unicité, pour les solutions de (4.2) (ou de (4.3)). Nous avons simplement montré l'équivalence des deux problèmes. L'unicité est évidemment liée à l'injectivité de A, comme le précise le résultat suivant.

Lemme 4.1. *La solution du problème* (4.2) *est unique si, et seulement si, l'opérateur A est injectif*

Preuve. Notons tout d'abord que $\mathrm{Ker}A^*A = \mathrm{Ker}A$. Un sens est évident. Pour l'autre, nous avons :

$$A^*Ax = 0 \Rightarrow (A^*Ax, x) = 0 \Rightarrow (Ax, Ax)_F = \|Ax\|_F = 0 \Rightarrow Ax = 0$$

Par conséquent, A et A^*A sont injectifs en même temps, ce qui donne le résultat. □

En ce qui concerne l'existence, on a le résultat suivant :

Proposition 4.1. **i)** *L'équation (4.3) admet une solution si et seulement si* $\hat{z} \in \mathrm{Im}A \oplus \mathrm{Im}A^{\perp}$.

ii) *Si* $\hat{z} \in \mathrm{Im}A \oplus \mathrm{Im}A^{\perp}$*, l'ensemble S des solutions de* (4.3) *est un convexe fermé non vide de E.*

Preuve. **i)** Soit $x \in E$ une solution de (4.3). On a donc $Ax - \hat{z} \in \mathrm{Ker}A^* = \overline{\mathrm{Im}A}^{\perp} = (\mathrm{Im}A)^{\perp}$. Donc $\hat{z} = Ax + (\hat{z} - Ax) \in \mathrm{Im}A \oplus (\mathrm{Im}A)^{\perp}$.
Inversement, soit $\hat{z} = z^1 + z^2$, avec $z^1 \in \mathrm{Im}A$, $z^2 \in \mathrm{Im}A^{\perp}$. Il existe donc $x \in E$, tel que $Ax = z^1$. Évidemment $A^*Ax = A^*z^1$. Mais, toujours parce que $(\mathrm{Im}A)^{\perp} = \mathrm{Ker}A^*$, $A^*z^2 = 0$, c'est-à-dire que $A^*\hat{z} = A^*z^1 = A^*Ax$, et $\hat{z}$ est une solution de (4.3).

ii) L'ensemble des solutions est non-vide d'après le point i. C'est un espace affine, c'est donc en particulier un convexe, et il est fermé puisque c'est l'image réciproque de $\{A^\star\hat{z}\}$ par l'opérateur continu $A^\star A$. Ceci prouve que $S = x_0 + \mathrm{Ker}A$, où x_0 est une solution quelconque de (4.3).

□

Lemme 4.2. *Le sous-espace* $\mathrm{Im}A + \mathrm{Im}A^\perp$ *est dense dans* F.

Preuve. Si $x \in (\mathrm{Im}A + \mathrm{Im}A^\perp)^\perp$, alors, pour tous $y \in \mathrm{Im}A$ et $z \in \mathrm{Im}A^\perp$, on a $(x, y+z) = 0$. En choisissant d abord $z = 0$, on obtient $x \in \mathrm{Im}A^\perp$, puis en choisissant $y = 0$, on obtient $x \in \mathrm{Im}A^{\perp\perp} = \overline{\mathrm{Im}A}$. Autrement dit, $(\mathrm{Im}A + \mathrm{Im}A^\perp)^\perp = \overline{\mathrm{Im}A} \cap \mathrm{Im}A^\perp = \{0\}$, ce qui est équivalent à la densité cherchée. □

On a donc bien réussi a étendre la notion de solution à un sous-espace dense dans F, ce qui est ń presqu'aussi bien ż que de l'étendre à F tout entier.

Remarque 4.2. Notons que, dans le cas général, la condition $\hat{z} \in \mathrm{Im}A \oplus \mathrm{Im}A^\perp$ est *non-triviale*. En effet, le théorème de projection dit seulement $F = \overline{\mathrm{Im}A} \oplus \mathrm{Im}A^\perp$, ce qui est différent si $\mathrm{Im}A$ n'est pas fermée.

On retrouve encore ici l'importance de cette condition. Dans ce cas, et seulement dans ce cas, le problème (4.3), et donc (4.2), admet *toujours* une solution.

En dimension finie, cette condition est bien entendu automatiquement vérifiée (tout sous-espace est fermé). Nous retrouverons ce résultat à la proposition 4.2.

Corollaire 4.1. *Si* $\hat{z} \in \mathrm{Im}A \oplus \mathrm{Im}A^\perp$*, le problème* (4.2) *admet une unique solution de norme minimale.*

Preuve. Notons S l'ensemble de solutions de (4.2). Chercher une solution de norme minimale de ce problème de moindres carrés revient à résoudre le problème suivant :

$$\min_{x \in S} \|x\|_E, \quad S = \{x \in E,\ \|Ax - b\|_F \text{ minimal}\}$$

c'est-à-dire à projeter l'origine sur l'ensemble S. D'après la proposition 4.1, S est un convexe fermé non-vide de E. Le théorème de projection (A.1) implique que S possède un élément de norme minimale, qui est la solution cherchée. □

Nous noterons $\hat{x}$ cette solution particulière.

Remarque 4.3. Il est facile de voir que $\hat{x}$ dépend linéairement de $\hat{z}$ (par exemple à cause de l'équation normale). L'opérateur linéaire $A^\dagger : \mathrm{Im}A \oplus \mathrm{Im}A^\perp \mapsto E$ défini par $A^\dagger \hat{z} = \hat{x}$ s'appelle le *pseudo-inverse* de A. On peut démontrer (voir par exemple [25]) qu'il n'est continu que si $\mathrm{Im}A$ est fermé. Nous donnerons quelques indications sur le pseudo-inverse au paragraphe 4.1.2.

Exemple 4.1 (Injection canonique)**.**
Notons A l'injection canonique de $H_0^1(0,1)$ dans $L^2(0,1)$ ($Au = u$, pour $u \in H_0^1(0,1)$). Inverser A revient en fait à dériver une fonction, et nous devons donc nous attendre à des difficultés. Nous pouvons déjà noter que l'image de A est (évidemment) $H_0^1(0,1)$, vu comme sous-espace de $L^2(0,10$, qui n'est pas fermé (il est dense, et strictement inclus dans $L^2(0,1)$). Bien évidemment, dans ce cas, A est injectif !

Commençons par chercher l'adjoint de A, et pour cela nous devons préciser les normes. Nous prenons comme norme sur $H_0^1(0,1)$ la quantité $\|u\|_1^2 = \int_0^1 |u'(x)|\ dx$ (c'est bien une norme d'après l'inégalité de Poincaré). Dans ces conditions, pour $v \in L^2(0,1)$, l'adjoint $A^* v$ est caractérisé par :

$$(u, A^* v) = (Au, v), \quad \forall u \in H_0^1(0,1),$$

soit, en explicitant les produits scalaires :

$$\int_0^1 u'(x)(A^* v)'(x)\ dx = \int_0^1 u(x)v(x)\ dx, \quad \forall u \in H_0^1(0,1).$$

On reconnaît dans cette dernière égalité la formulation variationnelle d'un problème elliptique du second ordre. A^*v est la solution (unique, d'après le lemme de Lax-Milgram) du problème de Dirichlet :

$$\begin{cases} -(A^*v)'' = v, & x \in]0,1[\\ (A^*v)(0) = (A^*v)(1) = 0 \end{cases} \tag{4.4}$$

Ainsi, l'adjoint de A se calcule en intégrant une équation différentielle. En particulier, l'image de A^* est composé des fonctions de $H_0^1(0,1) \cap H^2(0,1)$, donc de fonctions plus régulières.

Étant donné $\hat{z} \in L^2(0,1)$, le problème de moindres carrés (4.2) n'a de solution que si $z \in H_0^1(0,1)$, et la solution est alors $u = \hat{z}$. Si cette condition n'est pas vérifiée, le problème de minimisation n'a pas de solution (le minimum n'est pas atteint).

4.1.1 Cas de la dimension finie

Si E et F sont deux espace de dimension finie, on peut supposer que $E = \mathbf{R}^n$ et $F = \mathbf{R}^m$. Dans ce cas, A s'identifie à une matrice $A \in \mathbf{R}^{m\times n}$. Nous supposerons en général que $m > n$, c'est-à-dire que le problème est *sur-determiné*.

On peut préciser, dans ce cas, les résultats du paragraphe précédent.

Proposition 4.2. *Quand E et F sont deux espaces de dimension finie, le problème de moindres carrés admet toujours au moins une solution.*

Cette solution est unique si, et seulement si, A est de rang maximal (de rang n si le problème est sur-determiné).

Preuve. Il s'agit d'une simple application du lemme 4.1 et de la proposition 4.1, puisque l'image de A est toujours fermée. A est injective si et seulement si elle est de rang maximal (quand $m > n$).

On peut également donner une preuve directe, n'utilisant pas les résultats du paragraphe précédent. Nous montrons que les équations normales ont toujours au moins une solution, c'es-à-dire que $A^t\hat{z} \in \mathrm{Im}(A^t)$. Pour cela, commençons par remarquer que $\mathrm{Im}A^t = \mathrm{Im}(A^tA)$. En effet, $\mathrm{Im}(A^tA) = \mathrm{Ker}(A^tA)^\perp = \mathrm{Ker}(A)^\perp = \mathrm{Im}(A^t)$. Il suffit donc de vérifier que $A^t\hat{z} \in \mathrm{Ker}(A)^\perp$, or c'est immédiat, puisque, si $x \in \mathrm{Ker}(A)^\perp$, $(A^t\hat{z}, x) = (\hat{z}, Ax) = 0$. □

Proposition 4.3. *Sous l'hypothèse que A est de rang n, la matrice des équations normales A^tA est* définie positive.

Preuve. Il est facile de voir que A^tA est semi-définie positive (c'est en fait toujours vrai) :

$$(A^tAx, x) = (Ax, Ax) = \|x\|^2 \geq 0.$$

De plus, quand A est de rang n, nous avons vu (au lemme 4.1) que la matrice (carrée) A^tA est injective, donc inversible. Elle est donc définie positive. □

Ce résultat pourrait laisser penser qu'une méthode pour résoudre (4.2) est de résoudre (4.3) par la méthode de Choleski. Nous verrons au chapitre 5 que ce n'est pas forcément une bonne idée.

Exemple 4.2 (Régression linéaire)**.**
Nous cherchons à faire passer une droite $y(t) = \alpha + \beta t$ par un ensemble de points expérimentaux

$(t_i, y_i), i = 1, \ldots, m$. Cela conduit au système sur-déterminé

$$\begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 & y_1 \\ 1 & y_2 \\ \vdots & \vdots \\ 1 & y_m \end{pmatrix}.$$

L'équation normale

$$\begin{pmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m y_i t_i \end{pmatrix}$$

a pour solution

$$\beta = \frac{\sum_{i=1}^m y_i t_i - m\bar{y}\bar{t}}{\sum_{i=1}^m t_i^2 - m\bar{t}^2}, \qquad \alpha = \bar{y} - \beta\bar{t},$$

où nous avons noté $\bar{y} = \sum_{i=1}^m y_i/m$ et $\bar{t} = \sum_{i=1}^m t_i/m$.

Noter que la droite obtenue passe par les moyennes $(\bar{t}, \bar{y})$.

4.1.2 Compléments : projections et pseudo-inverse

4.2 Décomposition en valeurs singulières de matrices

La décomposition en valeurs singulières (singular value decomposition en anglais, souvent abrégée **SVD**) est devenue depuis quelques dizaines d'années un outil fondamental pour étudier un nombre croissant de problèmes linéaires. La décomposition a été découverte il y a plus de cent ans par Beltrami, mais n'est devenu un outil numérique que depuis la fin des années 1960, quand G. Golub a montré comment on pouvait la calculer de façon stable et (raisonnablement) efficace.

Cette décomposition est une sorte de diagonalisation qui donne une ń carte d'identité ż complète de l'opérateur. Nous verrons en particulier qu'elle donne une solution simple (en théorie) du problème de moindres carrés.

Nous énonçons le théorème principal dans le cas des matrices réelles.

Théorème 4.3. *Soit $A \in \mathbf{R}^{m\times n}$ une matrice de rang r. Il existe deux matrices orthogonales $U \in \mathbf{R}^{m\times m}$, $(U^tU = UU^t = I_m)$ et $V \in \mathbf{R}^{n\times n}$, $(V^tV = VV^t = I_n)$ telles que*

$$A = U\Sigma V^t, \qquad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.5}$$

où $\Sigma \in \mathbf{R}^{m\times n}$, $\Sigma_1 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, et

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

Composante par composante, l'identité matricielle (4.5) *devient :*

$$Av_j = \sigma_j u_j, \quad A^t u_j = \sigma_j v_j, \qquad \text{pour } j = 1, \ldots, n, \tag{4.6}$$

$$A^t u_j = 0, \qquad \text{pour } j = n+1, \ldots, m. \tag{4.7}$$

Si l'on note $U = (u_1, \cdots, u_m)$, $V = (v_1, \cdots, v_n)$ les colonnes des matrices U et V , les vecteurs u_j et v_j sont, respectivement, les *vecteurs singuliers* droits et gauches associés à la *valeur singulière* σ_j.

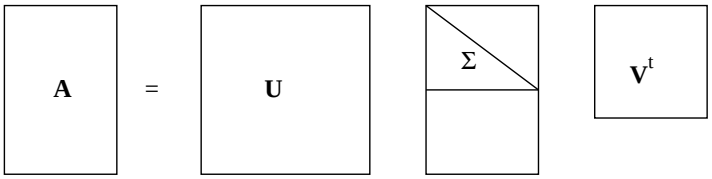

FIGURE 4.2 – Illustration de la SVD

Preuve. La preuve se fait par récurrence sur n.

Par définition de ce qu'est une norme matricielle subordonnée, il existe un vecteur $v_1 \in \mathbf{R}^n$ tel que

$$\|v_1\|_2 = 1, \quad \|Av_1\|_2 = \|A\|_2 \underset{\text{def}}{=} \sigma,$$

où σ est strictement positif (si $\sigma = 0$, alors $A = 0$, et il n'y a rien à démontrer). Posons $u_1 = 1/\sigma A v_1 \in \mathbf{R}^m$. Complétons le vecteur v_1 en une base orthogonale de $\mathbf{R}^n$, et notons $V = (v_1, V_1) \in \mathbf{R}^{n\times n}$ la matrice formée par les vecteurs de base. Faisons de même pour u_1 et $\mathbf{R}^m$, notant $U = (u_1, U_1) \in \mathbf{R}^{m\times m}$. Remarquons que les matrices U et V sont orthogonales par construction.

D'après notre choix de U_1, $U_1^t A v_1 = \sigma U_1^t u_1 = 0$, et donc le produit $U^t AV$ a la structure par blocs suivante :

$$A_1 \underset{\text{def}}{=} U^t AV = \begin{pmatrix} \sigma & w^t \\ 0 & B \end{pmatrix}$$

avec $w^t = u_1^t AV_1$ et $B = U_1^t AV_1 \in \mathbf{R}^{(m-1)\times(n-1)}$.

Comme U et V sont orthogonales, $\|A_1\|_2 = \|A\|_2 = \sigma$. Mais la double inégalité

$$\|A_1\|_2 \left(\sigma^2 + w^t w\right)^{1/2} \geq \left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma^2 + w^t w \\ Bw \end{pmatrix} \right\|_2 \geq \sigma^2 + w^t w,$$

montre que $\|A_1\|_2 \geq (\sigma^2 + w^t w)^{1/2}$. On doit donc avoir $w = 0$. On peut alors terminer la démonstration en appliquant l'hypothèse de récurrence à B. □

Du point de vue géométrique, ce théorème exprime que toute application linéaire peut être vue, après un changement de bases *orthogonales* dans chacun des espaces, comme agissant comme une dilatation dans chaque direction. Insistons sur l'importance du fait que les bases considérées sont orthonormées. Ceci constitue la principale différence avec la diagonalisation. En effet, si toute matrice possède une DVS, seules les matrices normales (ce sont celles qui commutent avec leur adjoint) sont diagonalisables dans une base orthonormée.

Exemple 4.3.

Illustrons cet aspect géometrique sur la figure 4.3, avec l'exemple de la matrice

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

dont la décomposition en valeurs singulières est (exercice !) :

$$A = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}^t$$

Nous calculons l'image du cercle unité. On commence par l'action de V^t, qui est une rotation (d'angle $-45°$), puis par Σ, dont l'action dilate l'axe des x dans un rapport 4, et l'axe des y dans un rapport 2, enfin par U qui est une rotation d'angle $45°$.

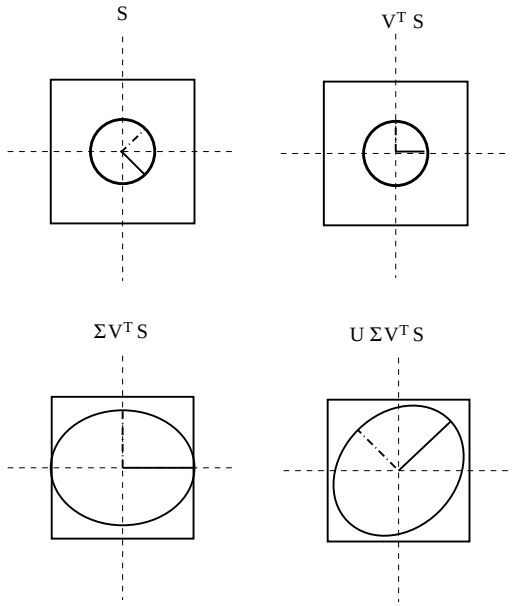

FIGURE 4.3 – Signification géometrique de la SVD

Les valeurs singulières sont reliées aux valeurs propres de matrices ń dérivées ż de A. La proposition suivante s'obtient facilement à partir de la SVD de A.

Proposition 4.4. *Soit $A \in \mathbf{R}^{m\times n}$ une matrice. Notons $A = U\Sigma V^t$ sa décomposition en valeurs singulières.*

i) *Les valeurs propres de la matrice A^tA sont les nombres σ_j^2, $j = 1,\ldots,n$, et ses vecteurs propres sont les vecteurs singuliers à gauche de A, v_j, $j = 1,\ldots,n$;*

ii) *Les valeurs propres de la matrice $\begin{bmatrix} 0 & A^t \\ A & 0 \end{bmatrix}$ sont les nombres $\pm\sigma_j$, $j = 1,\ldots,n$, et ses vecteurs propres sont $\frac{1}{\sqrt{2}} \begin{bmatrix} v_j \\ \pm u_j \end{bmatrix}$*

Cette proposition nous permet de préciser en quel sens la SVD d'une matrice A est unique. Puisque les valeurs singulières sont les valeurs propres de A^tA, elles sont determinées par A, donc uniques. Les

vecteurs singuliers à droite appartenant à une valeur singulière simple sont également uniques (à un facteur ± 1 près), par contre pour une valeur singulière multiple, seul le sous-espace est unique. Enfin, les vecteurs singuliers à gauche correspondant à une valeur singulière simple sont également uniques. Par contre, les $m-n$ derniers vecteurs singuliers à droite ne sont pas determinés uniquement par A, seul le sous-espace engendré par $u_{n+1},\dots,u_m$ est unique.

Il peut être intéressant d'écrire la SVD de A sous une forme légèrement différente de (4.5) :

– Tout d'abord, en posant $U_n = (u_1,\dots,u_n)$, on a

$$A = U_n \Sigma V^t \tag{4.8}$$

que l'on appelle parfois la ń factorisation en valeurs singulières ż de A.

– De même si A n'est pas de rang n, notons $U_r = (u_1,\dots,u_r)$, $V_r = (v_1,\dots,v_r)$ les matrices formées par les r premiers vecteurs singuliers, et $\Sigma_r = \mathrm{diag}(\sigma_1,\dots,\sigma_r)$. Alors :

$$A = U_r \Sigma_r V_r^t = \sum_{i=1}^{r} \sigma_i u_i^t v_i. \tag{4.9}$$

Notons que les matrices U_r et V_r sont orthogonales, puisque c'est le cas des vecteurs $u_1,\dots,u_r$ et $v_1,\dots,v_r$. La décomposition (4.9) correspond à une somme de matrices de rang 1.

Cette dernière forme est celle qui se prête le mieux à la généralisation aux opérateurs dans les espaces de Hilbert, que nous verrons au paragraphe 4.3. Elle exprime que les colonnes de V_r forment une base de $(\mathrm{Ker}A)^\perp$, que celles de U_r forment une base de $\mathrm{Im}A$, et qu'en développant les vecteurs dans cette base, on a

$$x = x^0 + \sum_{i=1}^{r} (x, v_i)\, v_i \Longrightarrow Ax = \sum_{i=1}^{r} \sigma_i (x, v_i)\, u_i, \quad \text{où } x^0 \in \mathrm{Ker}A. \tag{4.10}$$

Comme nous l'avons annoncé, on peut lire les principales propriétés de la matrice A sur sa DVS.

Proposition 4.5. *On a les relations suivantes :*

i) *Le rang de A est égal au nombre de valeurs singulières non-nulles ;*

ii) $\mathrm{Ker}A = \mathrm{vect}(v_{r+1},\cdots,v_n), \quad \mathrm{Im}A = \mathrm{vect}(u_1,\cdots,u_r)$ *;*

iii) $\mathrm{Ker}A^t = \mathrm{vect}(v_1,\cdots,v_r), \quad \mathrm{Im}A^t = \mathrm{vect}(u_{r+1},\cdots,u_m)$ *;*

iv) $\|A\|_2 = \sigma_1$.

Preuve. Les matrices U et V sont des isométries, donc des bijections. Les différentes propriétés sont donc des conséquences de ce qu'elles sont vraies pour Σ.

i) Il est clair que le rang d'une matrice diagonale est égal au nombre d'éléments diagonaux non-nuls. Le résultat général est une conséquence de la remarque ci-dessus.

ii) Soit $v \in \mathrm{Ker}A$, alors $V^t v \in \mathrm{Ker}U^t AV = \mathrm{Ker}\Sigma$, et réciproquement. Encore une fois, il est clair que le noyau de Σ est engendré par les colonnes $r+1$ à n de l'identité, et donc le noyau de A est engendré par le produit de V par ces colonnes, c'est-à-dire les vecteurs $v_{r+1},\dots,v_n$.

Un raisonnement analogue donne le résultat pour l'image de A.

iii) Le raisonnement est identique à celui du point précédent.

iv) Puisque U et V sont des isométries, on a pour $x \in R^n$:

$$\frac{\|Ax\|_2}{\|x\|_2} = \frac{\|U\Sigma V^t x\|_2}{\|x\|_2} = \frac{\|\Sigma y\|_2}{\|y\|_2}$$

avec $y = V^t x$ (et donc $\|y\|_2 = \|x\|_2$). Donc :

$$\frac{\|Ax\|_2}{\|x\|_2} = \frac{(\sum_{j=1}^r \sigma_j^2 y_j^2)^{1/2}}{(\sum_{j=1}^r y_j^2)^{1/2}} \leq \sigma_1,$$

avec égalité pour $y = (1,0,\ldots,0)^t$.

□

La décomposition en valeurs singulières permet d'établir un résultat d'approximation d'une matrice par des matrices de rang plus faible, qui nous sera utile quand nous étudierons la troncature spectrale (paragraphe 6.2.2).

Proposition 4.6. *Pour $k = 1,\ldots,n-1$, notons $A_k = U\Sigma_k V^t$.*
On a $\|A - A_k\|_2 = \sigma_{k+1}$, et la matrice A_k est la matrice de rang k qui minimise l'écart $\|A - B\|_2$.

Preuve. Il est clair que A_k est bien de rang k. Par ailleurs,

$$\|A - A_k\|_2 = \left\| U \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_n \end{pmatrix} V^t \right\|_2 = \left\| \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_n \end{pmatrix} \right\|_2 = \sigma_{k+1}.$$

Enfin, soit B une matrice de rang k. Le noyau de B est donc de dimension $n-k$, et l'espace engendré par $v_1,\ldots,v_{k+1}$ est de dimension $k+1$. La somme des dimensions vaut $(n-k)+(k+1) > n$, et donc ces deux sous-espaces ont une intersection non-triviale. Soit x un vecteur de norme 1 dans l'intersection. On a d'une part

$$\|A - B\|_2 \geq \|(A-B)x\|_2$$

et d'autre part ($(A-B)x = Ax$, puisque $x \in \mathrm{Ker}\, B$) :

$$\|(A-B)x\|_2 = \|Ax\|_2 = \left\|U\Sigma V^t x\right\|_2 = \left\|\Sigma(V^t x)\right\|_2 \geq \sigma_{k+1}\left\|V^t x\right\|_2 = \sigma_{k+1},$$

d'où le résultat en comparant les deux inégalités. □

4.2.1 Applications de la SVD aux problèmes de moindres carrés

La décomposition en valeurs singulières fournit la solution la plus claire du problème de moindres carrés. Nous allons l'utiliser dans ce paragraphe pour obtenir une meilleure compréhension théorique, en particulier des questions de perturbations, et nous y reviendrons au chapitre 5 en tant que méthode numérique.

Dans ce paragraphe, nous abandonnons l'hypothèse que A est de rang n, puisqu'un des points forts de la SVD est justement de permettre de traiter ce cas. Quand A est de rang $r < n$, nous chercherons la solution de norme minimale.

Introduisons la SVD de A (équation (4.5)) dans le problème de moindres carrés (4.2) :

$$\|Ax-z\|_2^2 = \left\|U\Sigma V^t x - z\right\|_2^2 = \left\|\Sigma(V^t x) - U^t z\right\|_2^2 \tag{4.11}$$

puisque U est orthogonale. Notons $w = U^t z$ (avec $\|w\|_2 = \|z\|_2$), et prenons comme nouvelle inconnue $y = V^t x$ (avec encore $\|y\|_2 = \|x\|_2$, puisque V est orthogonale, ce qui sera important pour calculer la solution de norme minimale).Comme Σ est diagonale, ce problème est découplé, et se résout composante par composante dans les bases $(u_1,\dots,u_m)$ et $(v_1,\dots,v_n)$.

Nous avons donc :

$$\|Ax-z\|_2^2 = \|\Sigma y - w\|_2^2 = \sum_{i=1}^{r} |\sigma_i y_i - w_i|^2 . \tag{4.12}$$

On obtient donc toutes les solutions du problème (4.2) en posant :

$$y_i = \begin{cases} w_i/\sigma_i, & \text{pour } i = 1,\dots,r \\ \text{quelconque} & \text{pour } i = r+1,\dots,n. \end{cases} \tag{4.13}$$

On retrouve bien les résultats du paragraphe 4.1.1 : dans le cas où A est de rang plein, il y a bien une solution unique, et dans le cas contraire, la solution est définie à l'addition d'un élément du noyau de A près. Dans ce dernier cas, la norme d'une solution est $\sum_{i=1}^{r} |w_i/\sigma_i|^2 + \|\text{élément du noyau}\|_2^2$. La solution de norme minimale est donc celle qui n'a pas de composante dans $\mathrm{Ker}A$. Énonçons le résultat sous une forme plus intrinsèque :

Théorème 4.4. *La solution de norme minimale du problème* (4.2) *est donnée par :*

$$x = \sum_{i=1}^{r} (z,u_i)/\sigma_i \, v_i \tag{4.14}$$

Preuve. Nous avons déjà fait la plus grande partie de la démonstration. Par définition, $y = V^t x$ s'inverse en $x = Vy$, dont les composantes sont $(x,v_i) = (Vy,v_i) = (y,V^t v_i) = y_i$ En utilisant (4.13), nous avons

$$x = \sum_{i=1}^{r} (x,v_i)\, v_i = \sum_{i=1}^{r} y_i \, v_i = \sum_{i=1}^{r} (z,u_i)/\sigma_i \, v_i$$

□

Ce résultat nous permet de donner des premières indications sur la sensibilité de la solution x par rapport à des perturbations sur la donnée z. Pour simplifier, plaçons nous dans le cas où A est de rang n. Remplaçons z par $z+\delta z$, et notons $x+\delta x$ la solution du problème perturbé. Par linéarité, on déduit immédiatement de (4.14) que

$$\delta x = \sum_{i=1}^{r} (\delta z,u_i)/\sigma_i \, v_i \tag{4.15}$$

et donc que $\|\delta x\|_2^2 = \sum_{i=1}^{r} |\delta z|^2 / \sigma_i^2$.

En l'absence d'information plus précise, le mieux que l'on puisse déduire de (4.15) est

$$\|\delta x\|_2 \leq \frac{\|\delta z\|_2}{\sigma_n},$$

puisque σ_n est la plus petite valeur singulière de A, ce qui veut dire que l'erreur sur la solution a été amplifiée par *l'inverse de la plus petite valeur singulière* par rapport à l'erreur sur la donnée. Dans le cas où cette plus petite valeur singulière est petite, cette amplification peut devenir dramatique. Nous verrons que cela sera le cas général pour la discrétisation des problèmes mal posés, et cela constitue l'explication fondamentale à l'instabilité rencontrée lors de leur solution.

4.3 Développement en valeurs singulières des opérateurs compacts

Dans ce paragraphe nous supposerons que A est un opérateur compact de E dans F (voir le paragraphe A.2.3). Nous allons généraliser la décomposition en valeurs singulières à cette situation. La principale différence sera l'existence d'une *infinité* (dénombrable) de valeurs singulières. Nous obtiendrons en conséquence un critère pour l'existence d'une solution au problème de moindres carrés (le critère de Picard, théorème 4.6).

Théorème 4.5. *Soit $A : E \to F$ un opérateur* compact. *Il existe une suite $(\sigma_j)_{j\in\mathbf{N}} \in \mathbf{R}_+$, et deux familles orthonormales $(e_j)_{j\in\mathbf{N}} \in E$, $(f_j)_{j\in\mathbf{N}} \in F$ telles que :*

i) $(\sigma_j)_{j\in\mathbf{N}}$ *est décroissante*, $\lim_{j\to+\infty}\sigma_j = 0$

ii) $Ae_j = \sigma_j f_j$; $\quad A^*f_j = \sigma_j e_j, \quad j \in \mathbf{N}$

iii) *Pour tout $x \in E$, on a le développement*

$$x = x_0 + \sum_{j=1}^{+\infty}(x,e_j)e_j, \quad \text{où } x_0 \in \mathrm{Ker}A \tag{4.16}$$

iv) *Pour tous $x \in E$ et $y \in F$ on a :*

$$Ax = \sum_{j=1}^{+\infty}\sigma_j(x,e_j)f_j, \qquad A^*y = \sum_{j=1}^{+\infty}\sigma_j(y,f_j)e_j \tag{4.17}$$

La suite $(e_j)_{j\in\mathbf{N}}$ est une base hilbertienne de $\mathrm{Ker}A^\perp$, la suite $(f_j)_{j\in\mathbf{N}}$ est une base hilbertienne de $\overline{\mathrm{Im}A}$

Preuve*(Peut être omise).* Considérons l'opérateur auto adjoint $T = A^*A$. T est compact comme composé de A compact et de A^* continu (et compact). La théorie spectrale des opérateurs auto-adjoints compacts (voir le théorème A.7 ou Brézis [12]) implique que T a une suite de valeurs propres non nulles $(\lambda_j)_{j\in\mathbf{N}} \in \mathbf{R}^*$, et de vecteurs propres $(e_j)_{j\in\mathbf{N}}$, tels que

$$Te_j = \lambda_j e_j$$

et si $x \in E$, il existe $x_0 \in \mathrm{Ker}T, x = x_0 + \sum_{j=1}^{+\infty}(x,e_j)e_j$.

Supposons les $(\lambda_j)_{j\in\mathbf{N}}$ rangés en une suite décroissante, chaque valeur étant comptée avec son ordre de multiplicité. Observons alors que, $\forall j, \lambda_j > 0$ puisque :

$$\lambda_j = (Te_j,e_j) = (A^*Ae_j,e_j) = \|Ae_j\|^2 \geq 0$$

et λ_j est non-nulle par hypothèse. Nous pouvons donc définir $\sigma_j = \lambda_j^{1/2}$, $\sigma_j > 0$ et poser $f_j = \dfrac{1}{\sigma_j}Ae_j \in \mathrm{Im}A \subset F$.

Nous avons bien $A^*f_j = \dfrac{1}{\sigma_j}A^*Ae_j = \dfrac{1}{\sigma_j}Te_j = \sigma_j e_j$ et $(f_j)_{j\in\mathbf{N}}$ est une suite orthonormale dans F, puisque :

$$(f_j,f_k) = \frac{1}{\sigma_j\sigma_k}(Ae_j,Ae_k) = \frac{\lambda_k}{\sigma_j\sigma_k}(e_j,e_k) = \sqrt{\frac{\sigma_k}{\sigma_j}}\delta_{jk} = \delta_{jk}.$$

Les identités (ii) sont immédiates. Pour la représentation (4.16), remarquons que $\mathrm{Ker}T = \mathrm{Ker}(A^*A) = \mathrm{Ker}A$ (une inclusion est triviale. Pour l'autre : $A^*Ax = 0 \Rightarrow (A^*Ax,x) = \|Ax\|^2 = 0 \Rightarrow x \in \mathrm{Ker}A$. Le raisonnement est le même qu'au lemme 4.3). Le développement (4.16) est alors celui obtenu plus haut.

Pour conclure, établissons le premier développement de (4.17). Pour $x \in E$, posons

$$X = \sum_{j=1}^{+\infty} \sigma_j (x, e_j) f_j.$$

La série est convergente dans E, avec $\|X\| \leq \sigma_1 \|x\|$, et $X \in \overline{\mathrm{Im} A}$. Par conséquent, $X - Ax \in \mathrm{Ker} A^* \cap \overline{\mathrm{Im} A} = \mathrm{Ker} A^* \cap (\mathrm{Ker} A^*)^\perp$ et donc $X = Ax$. Le second développement s'établit de la même façon. □

Définition 4.1. Les nombres σ_j sont appelés les valeurs singulières de A. Les vecteurs e_j et f_j sont les vecteurs singuliers. Le développement obtenu en (4.16) s'appelle le développement en valeurs singulières (ń singular value expansion ż, abrégé SVE, en anglais) de A.

Remarque 4.4. Nous avons supposé, pour simplifier, que l'opérateur A^*A avait une infinité de valeurs propres. Le cas d'un nombre fini de valeurs propres existe, mais ne correspond pas à des situations rencontrées en pratique. Nous avons choisi d'éliminer ce cas pour obtenir des énoncés plus simples.

Exemple 4.4 (Injection canonique)**.**
Cherchons le développement en valeurs singulières de l'opérateur de l'exemple 4.1. D'après (4.4), les valeurs sinulières et les fonctions associées sont solution de

$$\begin{cases} e_j = \sigma_j f_j, \\ -\sigma e_j'' = f_j, \quad \text{dans }]0,1[, e_j(0) = e_j(1) = 0. \end{cases} \tag{4.18}$$

En résolvant le système, on voit que les valeurs singulières sont les nombres $\sigma_k = 1/(k\pi)$, $k \geq 1$, et les fonctions propres sont données par :

$$e_j(t) = \sqrt{2} \sin(k\pi t), \quad f(j) = \sqrt{2} k\pi \sin(k\pi t).$$

L'interprétation des développements (4.16) et (4.17) est ici

$$u = \sum u_j \sqrt{2} \sin(k\pi t), \text{ avec } u_j = \sqrt{2} \int_0^1 u(t) \sin(k\pi t)\, dt.$$

On retrouve le développement de u en série de Fourier (au sens de $L^2(0,1)$).

4.3.1 Applications de la SVE aux problèmes de moindres carrés

Comme en dimension finie, Le développement en valeurs singulières de A permet une analyse complète des équations linéaires associées à l'opérateur A comme (4.1) ou (4.3).

Théorème 4.6. *Soit $z \in F$. L'équation* (4.1) *possède une solution dans E si et seulement si $z \in \overline{\mathrm{Im} A}$ et que de plus :*

$$\sum_{j=1}^{+\infty} \frac{|(z, f_j)|^2}{\sigma_j^2} < \infty. \tag{4.19}$$

Dans ce cas l'ensemble des solutions de (4.1) *est donné par*

$$x = \sum_{j=1}^{+\infty} \frac{(z, f_j)}{\sigma_j} e_j + \mathrm{Ker} A \tag{4.20}$$

Preuve. Développons z sur la base hilbertienne des (f_j) : $z = \sum_{j=1}^{+\infty}(z,f_j)f_j$, et de même pour x sur la base hilbertienne des (e_j) : $x = \sum_{j=1}^{+\infty} x_j e_j$. En appliquant (4.17), nous obtenons $Ax = \sum_{j=1}^{+\infty} \sigma_j x_j f_j$. Il suffit alors d'identifier les deux développements pour obtenir (4.20). La condition (4.19) s'obtient simplement en exprimant que la série des coefficients du développement de x doit être de carré intégrable. □

Remarque 4.5. La condition obtenue en (4.19) s'appelle la condition de Picard. Elle a été obtenue par ce mathématicien français au début du siècle. Elle exprime une restriction sur les coefficients d'un élément de l'image de A. Ceux-ci doivent tendre plus rapidement vers 0 que ne l'exige la simple appartenance à $\overline{\mathrm{Im}A}$. En effet, la suite σ_j tend vers 0, donc $|(z,f_j)|^2/\sigma_j^2 \leq |(z,f_j)|^2$ pour j assez grand. En particulier, ces coefficients doivent décroître plus rapidement que les valeurs singulières. Cette condition (ou son analogue en dimension finie) peut se vérifier numériquement. Voir sur ce sujet les travaux de P. C. Hansen [36, 37].

Remarque 4.6. La solution de norme minimale de (4.1) est évidemment obtenue quand la contribution de $\mathrm{Ker}A$ est nulle (en effet, la série est orthogonale à $\mathrm{Ker}A$ puisque $f_j \in \mathrm{Im}A$, pour tout j, et donc la somme est dans $\overline{\mathrm{Im}A}$).

Exemple 4.5.
Continuons l'exemple 4.4. D'après l'identité de Bessel-Parseval, on a $\|u\|_{L^2}^2 = \sum_{k=1}^{\infty} |u_k|^2$. Ici, la condition $u \in \mathrm{Im}A$ se traduit par $u = \sum_{k=1}^{\infty} u_k f_k$, avec $\sum_{k=1}^{\infty} k^2 |u_k|^2 < \infty$, ce qui est bien équivalent à $u \in H^1(0,1)$.

Remarque 4.7. En reprenant le raisonnement du paragraphe 4.2.1 sur la stabilité, on peut retrouver grâce à la formule d'inversion (4.20) les raisons qui rendent le problème inverse mal posé (au moins dans le cas d'un opérateur compact), et en particulier expliquer l'instabilité. Supposons que la composante z_i du second membre exact soit perturbée en $z_i' = z_i + \eta$. Par linéarité, on voit facilement que la seule composante de la solution x' qui est modifiée est $x_i' = x_i + \eta/\sigma_i$. En passant aux normes, on a donc :

$$\left\|x' - x\right\|_E = \frac{\eta}{\sigma_i}. \tag{4.21}$$

Comme $\sigma_i \underset{i\to\infty}{\longrightarrow} 0$, il n'est pas possible de borner la différence entre x et x' uniformément en fonction de $z - z'$. On retrouve donc la conclusion que la solution de (4.1) ne dépend pas continûment du second membre.

Chapitre 5

Méthodes numériques pour les problèmes de moindres carrés

Après avoir vu la structure mathématique des problèmes de moindres carrés, nous nous tournons maintenant vers les méthodes permettant de les traiter numériquement sur ordinateur. Nous ne considérerons donc que des problèmes en dimension finie (qui peuvent éventuellement provenir de la discrétisation d'un opérateur intégral, comme nous l'avons vu au paragraphe 3.2). Nous considérons donc une matrice $A \in \mathbf{R}^{m\times n}$, et nous supposerons que $m \geq n$ (le problème est sur-determiné). Enfin, nous supposerons, dans presque tout ce chapitre, que la matrice A est de rang maximum (i.e. n). Le cas général, où A est de rang $r < n$, fait appel à des méthodes spécifiques. Nous verrons, au paragraphe 5.4, que la SVD peut être employée dans ce cas, et nous y reviendrons au chapitre 6 après avoir étudié les méthodes régularisation.

Nous commencerons par étudier le conditionnement des problèmes de moindres carrés. Nous étudierons ensuite la méthode classique des équations normales, que l'on peut résoudre par une factorisation de Cholesky. Nous exposerons en particulier les limites de cette méthode. Nous insisterons plutôt sur la méthode basée sur une factorisation QR de la matrice du système qui, bien que plus coûteuse que la précédente, est plus robuste, et constitue la méthode de choix pour résoudre les problèmes de moindres carrés de taille ń raisonnable ż (quelques centaines). Enfin nous donnerons quelques indications (limitées) sur les méthodes de calcul de la SVD.

Pour des raisons de place, nous n'aborderons pas les méthodes spécialisées pour les problèmes de grande taille, c'est-à-dire les méthodes directes pour les matrices creuses, et les méthodes itératives.

Avant de rentrer dans le vif du sujet, donnons quelques références. L'algèbre linéaire est l'un des sujets sur lesquels il existe le plus de livre, et la plupart comportent un chapitre sur les problèmes de moindres carrés. En français, citons le livre [46]. En anglais, le classique est [28], une référence encyclopédique sur les problèmes de moindres carrés est [9]. D'autres livres moins complets, mais plus abordables, sont [63], [58] et [21].

Nous conservons dans ce chapitre les notations du chapitre précédent, à une exception près : puisque (cf A.1) l'adjoint d'une matrice est sa transposée, nous noterons A^t la matrice adjointe (transposée) de A.

5.1 Conditionnement des problèmes de moindres carrés

La définition du conditionnement pour une matrice rectangulaire est

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} \tag{5.1}$$

qui (d'après la proposition 4.5 iv) généralise bien la définition ń usuelle ż pour les matrices carrées.

Théorème 5.1. *c Soit $A \in \mathbf{R}^{m\times n}$ une matrice de rang n, $x \in \mathbf{R}^m$, et x la solution du problème* (4.2)*. Notons $r = Ax - z$ le résidu. Enfin, soit $\tilde{x}$ la solution du problème de moindres carrés*

$$\min_{x\in\mathbf{R}^n} \|(A+\delta A)\tilde{x} - (z+\delta z)\|_2 .$$

On a la majoration

$$\frac{\|\tilde{x}-x\|_2}{\|x\|_2} \lesssim \kappa(A)\left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{1}{\cos\theta}\frac{\|\delta z\|_2}{\|z\|_2}\right) + \kappa(A)^2 \operatorname{tg}\theta \frac{\|\delta A\|_2}{\|A\|_2} \tag{5.2}$$

où $\sin\theta = \dfrac{\|r\|_2}{\|z\|_2}$ *et où le symbole* $\lesssim$ *indique que cette majoration est vraie à des termes d'ordre* $\max\left(\dfrac{\|\delta A\|_2}{\|A\|_2}, \dfrac{\|\delta z\|_2}{\|z\|_2}\right)$ *près.*

Preuve. Écrivons $\delta x = \tilde{x} - x$. Développons l'équation normale pour $\tilde{x}$:

$$(A^t + \delta A^t)(A+\delta A)\tilde{x} = (A^t + \delta A^t)(z+\delta z)$$

au premier ordre, et soustrayons l'équation normale pour x. Il reste (a vec $r = z - Ax$) :

$$A^t A\delta x \approx -A^t \delta A x + A^t \delta z + \delta A^t r$$

En passant aux normes, nous obtenons

$$\|\delta x\|_2 \lesssim \left\|(A^tA)^{-1}A^t\right\|_2 \|\delta A\|_2 \|x\|_2 + \left\|(A^tA)^{-1}A^t\right\|_2 \|\delta z\|_2 + \left\|(A^tA)^{-1}\right\|_2 \|\delta A\|_2 \|r\|_2 \tag{5.3}$$

Notons $A = U\Sigma V^t$ la SVD de A, avec $\Sigma = \begin{pmatrix}\Sigma_1\\ 0\end{pmatrix}$. Alors

$$A^tA = V\begin{pmatrix}\Sigma_1 & 0\end{pmatrix} U^tU \begin{pmatrix}\Sigma_1\\ 0\end{pmatrix} V^t = V\Sigma_1^2 V^t$$

et

$$(A^tA)^{-1}A^t = V\Sigma_1^{-2}V^tV\begin{pmatrix}\Sigma_1 & 0\end{pmatrix}U^t = V\Sigma_1^{-1}U^t.$$

Toujours d'après la proposition 4.5 iv,

$$\left\|(A^tA)^{-1}A^t\right\|_2 = \frac{1}{\sigma_n} \text{ et } \left\|(A^tA)^{-1}\right\|_2 = \frac{1}{\sigma_n^2}. \tag{5.4}$$

On déduit donc de (5.3), (5.4) que

$$\frac{\|\delta x\|_2}{\|x\|_2} \lesssim \kappa(A)\left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta z\|_2}{\|z\|_2}\frac{\|z\|_2}{\|A\|_2\|x\|_2}\right) + \kappa(A)^2\frac{\|\delta A\|_2}{\|A\|_2}\frac{\|r\|_2}{\|A\|_2\|x\|_2}.$$

Pour conclure il nous reste à relier les deux termes $\frac{\|z\|_2}{\|A\|_2\,\|x\|_2}$ et $\frac{\|r\|_2}{\|A\|_2\,\|x\|_2}$ à l'angle θ de l'énoncé. Or le résidu r est orthogonal à Ax, donc

$$\cos^2\theta = 1 - \sin^2\theta = \frac{\|z\|_2^2 - \|r\|_2^2}{\|z\|_2^2} = \frac{\|Ax\|_2^2}{\|z\|_2^2}$$

d'après le théorème de Pythagore. Enfin,

$$\frac{\|z\|_2}{\|A\|_2\,\|x\|_2} \leq \frac{\|z\|_2}{\|Ax\|_2} = \frac{1}{\cos\theta}.$$

□

L'angle θ mesure l'écart entre les vecteurs z et Ax, autrement dit la taille du résidu. Il est petit si r est petit et proche de $\pi/2$ si r est proche de b. On voit donc que le conditionnement d'un problème de moindres carrés peut dépendre non pas de $\kappa(A)$ mais de $\kappa(A)^2$. Cet effet est toutefois mitigé par la présence du facteur $1/\operatorname{tg}\theta$ devant ce terme : l'effet de $\kappa(A)^2$ ne sera visible que si le résidu est important.

En détail, si θ est petit, le résidu est petit et le conditionnement ń effectif ż est proche de $\kappa(A)$. Si θ n'est ni petit ni proche de $\pi/2$, le résidu est modérément grand, et le conditionnement effectif est gouverné par le terme en $\kappa(A)^2$, qui peut être beaucoup plus grand que $\kappa(A)$. Enfin, si θ est proche de $\pi/2$, la solution est beaucoup plus petite que le résidu, et le conditionnement effectif devient non-borné même si $\kappa(A)$ faible.

5.2 Équations normales

Comme nous l'avons vu au chapitre 4 (équation (4.3)), la solution d'un problème de moindres carrés se ramène, en théorie du moins, à la résolution d'un système linéaire pour la matrice A^tA (dite matrice des équation normales) :

$$A^tAx = A^tz. \tag{5.5}$$

Nous avons vu également (corollaire 4.1) que, sous l'hypothèse que A est rang n, la matrice A^tA est définie positive. Cette matrice est de taille n par n, et les équations normales représentent une ń compression ż d'information, puisque $n \leq m$. Le système (5.5) peut donc (toujours en théorie) être résolu par la factorisation de Cholesky. Nous rappelons cette méthode bien connue :

Proposition 5.1. *Soit C une matrice symétrique et définie positive. Il existe une unique matrice R triangulaire supérieure, à éléments diagonaux strictement positifs, telle que*

$$C = R^tR \tag{5.6}$$

Preuve. Elle se trouve dans les références citées ci-dessus. □

Il suffit ensuite de résoudre les deux systèmes linéaires

$$\left\{\begin{array}{rcl} R^ty &=& A^tz \\ Rx &=& y \end{array}\right. \tag{5.7}$$

Algorithme 5.1 Factorisation de Cholesky

for $j = 1 : n$ **do**
  for $i = 1 : j-1$ **do**
    $$R_{ij} = \left(C_{ij} - \sum_{k=1}^{i-1} R_{ki} R_{kj} \right) / R_{ii}$$
  end for
  $$R_{jj} = \left(C_{jj} - \sum_{k=1}^{j-1} R_{kj}^2 \right)^{1/2}$$
end for

pour obtenir la solution. Le premier a une matrice triangulaire inférieure, le second une matrice triangulaire supérieure.

Nous rappelons également, dans l'algorithme 5.2, l'algorithme de calcul de la factorisarion, qui consiste essentiellement à identifier, colonne par colonne, les éléments de R à partir de l'équation (5.6).

Noter que nous avons écrit l'algorithme comme si les matrices C et R occupaient des places différentes en mémoire. En pratique, il est usuel de stocker R à la place de C (qui est perdue).

Le nombre d'opérations flottantes nécessaires à la formation de la matrice des équations normales est $n(n+1)m + mn$ (en tirant parti de la symétrie). Le nombre d'opérations de la factorisation de Choleski est $n^3/3$ opérations flottantes (additions et multiplications), plus n divisions et n extractions de racines carrées. La solution des équations triangulaires (5.7) prend n^2 opérations, et est donc négligeable.

Le coût de la méthode des équations normales est donc, en ne gardant que les termes principaux : $mn^2 + \frac{1}{3}n^3$.

En dépit de sa simplicité, et de son coût raisonnable (nous verrons que les autres méthodes de résolution des problèmes de moindres carrés sont plus chères), la méthodes des équations normales n'est pas recommandée, pour des raisons de stabilité. Elle cumule deux inconvénients :

i) Le simple fait de former la matrice A^tA peut faire perdre de l'information sur les petits coefficients de la matrice A, ce qui peut avoir des conséquences désastreuses, comme le montre l'exemple 5.1 ci-dessous.

ii) De plus, le conditionnement de la matrice A^tA est le carré de celui de A (puisque les valeurs propres de A^tA sont les carrés des valeurs singulières de A).

Exemple 5.1 (d'après Björck [9])**.**
Soit le problème $\min_{x \in \mathbf{R}^3} \|Ax - z\|$, avec

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}, \quad z = \begin{pmatrix} 1, & 0, & 0, & 0 \end{pmatrix}^t.$$

Cet exemple peut correspondre à un problème où la somme $x_1 + x_2 + x_3$ est connue avec beaucoup

plus de précision de que les composants de x individuellement. La calcul exact donne

$$A = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{pmatrix}, \quad A^t z = \begin{pmatrix} 1, & 1, & 1 \end{pmatrix}^t, \quad x = \frac{1}{3+\varepsilon^2} \begin{pmatrix} 1, & 1, & 1 \end{pmatrix}^t$$

Supposons que $\varepsilon = 10^{-4}$ et que le calcul soit effectué avec 8 chiffres significatifs. Alors $1+\varepsilon^2 = 1.000000001$ sera arrondi à 1, et la matrice A^tA calculée sera singulière. Tout se passe comme si les trois dernières lignes de A étaient nulles, et n'avaient contribué aucune information.

Les difficultés illustrées par cet exemple peuvent être évitées, ou tout du moins les limites repoussées, en passant en double précision. Dans ces conditions, les équations normales deviennent une méthode acceptable. Toutefois, la méthode présentée au paragraphe suivant lui est préférable comme méthode de résolution générale.

5.3 La factorisation *QR*

La méthode ń moderne ż pour résoudre les problèmes de moindres carrés est basée sur une factorisation dite QR de la matrice A, où Q est une matrice orthogonale, et R est triangulaire supérieure. La réduction de A à une forme triangulaire s'effectue par une suite de multiplications par des matrices de réflexion, dites matrices de Householder (en hommage à A. Householder, pionnier de l'analyse numérique matricielle). Nous commencerons par définir, et donner les principales propriétés, de ces matrices, avant de montrer comment les utiliser pour obtenir la factorisation QR. Nous montrerons ensuite comment cette factorisation permet de résoudre le problème de moindres carrés.

5.3.1 Matrices de Householder

Définition 5.1. Soit u de norme (euclidienne) 1. Une matrice de Householder est de la forme :

$$H(u) = I - 2uu^t \tag{5.8}$$

Rappelons qu'une matrice $Q \in \mathbf{R}^m m \times n$ est dite *orthogonale* si elle vérifie $Q^tQ = I_n$. Une telle matrice conserve la norme euclidienne, $\|Qx\|_2 = \|x\|_2$, et correspond donc à une application linéaire qui est une isométrie.

Lemme 5.1. *Une matrice de Householder est symétrique et orthogonale.*

Preuve. La symétrie est claire. Pour le caractère orthogonal, un simple calcul donne :

$$(I-2uu^t)^t(I-2uu^t) = I-4uu^t+4(uu^t)(uu^t) = I-4uu^t+4u(u^tu)u^t) = I,$$

puisque $u^tu = \|u\|_2^2 = 1$. □

Du point de vue géométrique (voir figure 5.1), une matrice de Householder est une symétrie orthogonale par rapport à l'hyperplan orthogonal à u.

Notons qu'il est simple de multiplier un vecteur ou une matrice par une matrice de Householder, en ne connaissant que le vecteur u. Pour le produit d'un vecteur, il suffit de revenir à la définition (5.8) :

$$H(u)x = x - 2(uu^t)x = x - 2(u^tx)u. \tag{5.9}$$

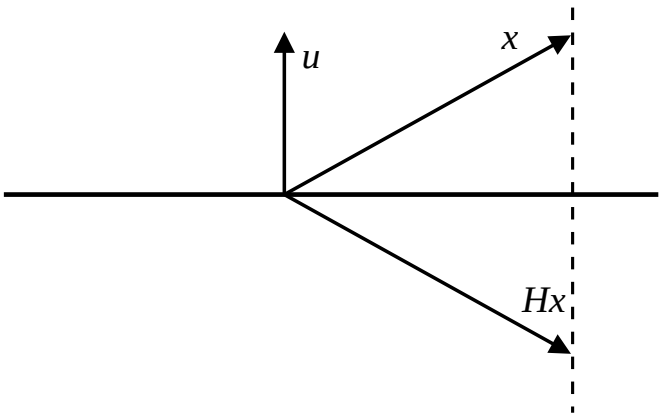

FIGURE 5.1 – Matrice de Householder – vue géometrique

qui ne demande que $2n$ opérations, au lieu de n^2 opérations si l'on forme et stocke la matrice $H(u)$. La produit par une matrice est aussi simple :

$$H(u)A = A - 2(uu^t)A = A - 2u(u^tA), \tag{5.10}$$

ne demande que $O(n^2)$ opérations, ce qui est plus économique que de former $H(u)$ et d'effectuer le produit.

Dans notre contexte, l'intérêt de ces matrices est qu'il en existe une qui transforme un vecteur donné en un multiple du premier vecteur de base.

Lemme 5.2. *Soit $x \in \mathbf{R}^n$, $x \neq 0$. Il existe un vecteur unitaire u tel que $H(u)$ soit parallèle à e_1.*

Preuve. Soit donc $x \neq 0$, nous cherchons donc u tel que $H(u)x = \alpha e_1$ (où α est inconnu). D'après (5.9), nous avons

$$H(u)x = x - 2(uu^t)x = x - 2(u^tx)u,$$

se trouve dans le plan engendré par x et u. On peut donc chercher u de norme 1, dans le plan engendré par x et e_1. u doit donc être parallèle à $\tilde{u} = x - \alpha e_1$, d'où $u = \dfrac{\tilde{u}}{\|\tilde{u}\|_2}$. Enfin, α est déterminé (au signe près) par $\|H(u)x\|_2 = \|x\|_2 = |\alpha|$, et il y a deux vecteurs solutions (qui sont d'ailleurs orthogonaux). Pour des raisons de stabilité numérique, il est préférable de prendre

$$u = \frac{x + \mathrm{sign}(x_1)e_1}{\|x + \mathrm{sign}(x_1)e_1\|_2}, \tag{5.11}$$

le choix du signe évite des annulations lors du calcul de la première composante de u. □

Maintenant que nous avons mis l'outil en place, voyons comment l'utiliser.

5.3.2 Factorisation *QR*

Théorème 5.2. *Soit $A \in \mathbf{R}^{m\times n}$ une matrice de rang n. Il existe une matrice orthogonale $Q \in \mathbf{R}^{m\times m}$ et une matrice triangulaire supérieure inversible telles que*

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix} \tag{5.12}$$

Preuve. Nous donnons une preuve constructive, en décrivant l'algorithme utilisé pour obtenir la factorisation. Dans cette méthode, la matrice Q n'apparaît qu'implicitement, comme produit de transformations de Householder.

Pour ce faire, nous procédons en n étapes, chacune d'entre elles agissant sur une colonne de A par une transformation de Householder.

La première étape consiste à déterminer une matrice de Householder qui envoie la première colonne de A sur une multiple du premier vecteur de base. C'est justement le problème que nous avons résolu au paragraphe précédent ! Nous prendrons donc $u_1 = \tilde{u_1}/\|u_1\|_2$ avec $\tilde{u_1} = a_1 + \text{sign}(a_{11})e_1$, et nous posons $P_1 = I - 2u_1u_1^t$. Nous obtenons

$$A^{(2)} = P_1A = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & \tilde{a}_{22} & \dots & \tilde{a}_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & \tilde{a}_{m2} & \dots & \tilde{a}_{mn} \end{pmatrix}.$$

On poursuit ensuite l'algorithme sur la sous-matrice de taille $(m-1)\times(n-1)$ située dans le coin en bas à gauche. Avant la $k^{\text{ième}}$ étape, la matrice réduite a la forme

$$A^{(k)} = \begin{pmatrix} R_{11}^{(k)} & R^{(k)} \\ 0 & \tilde{A}^{(k)} \end{pmatrix}, \tag{5.13}$$

où $R_{11}^{(k)}$ est triangulaire supérieure, et ne sera plus modifiée. On détermine alors une matrice de Householder $\tilde{P}^k$ qui envoie la première colonne de $\tilde{A}^{(k)}$ sur le premier vecteur de base, ce qui donne $u_k = \tilde{u_k}/\|u_k\|_2$ avec $\tilde{u_k} = \tilde{a}_k + \text{sign}(a_{kk})e_1$, et on pose

$$P_k = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{P}_k \end{pmatrix}, \quad A^{(k+1)} = P_kA^{(k)},$$

qui est encore de la même forme.

Après n étapes, nous avons l'égalité :

$$P_n\cdots P_2P_1A = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

avec R triangulaire supérieure, qui est bien le résultat attendu, en posant

$$Q^t = P_n\cdots P_2P_1.$$

Concluons en notant que la matrice R est inversible, puisque nous avons supposé A de rang n, et qu'une matrice orthogonale est inversible. □

Comme nous l'avions annoncé, la matrice orthogonale Q est obtenue implicitement, comme un produit de n matrices de Householder. Il n'est pas nécessaire (et même inutile) d'effectuer ce produit. Il suffit de conserver les différents vecteurs u_k, $k = 1,\dots,n$.

Nous pouvons illustrer l'algorithme dans le cas $m = 6$, $n = 4$. Sur la figure 5.2, un x représente un élément (a-priori non-nul) de la matrice A, un 0 un élément certainement nul, et un r un élément de R qui ne sera plus modifié. Une telle figure s'appelle un diagramme de Wilkinson.

Précisons quelques détails d'implémentation. À chaque étape, l'on opère que sur la sous-matrice notée $\tilde{A}^{(k)}$ plus haut. De plus, conformément à ce que nous avons noté au paragraphe précédent, le produit $\tilde{P}_k\tilde{A}^{(k)}$ se calcule par :

$$\tilde{P}_k\tilde{A}^{(k)} = \tilde{A}^{(k)} - 2u_k(u_k^t\tilde{A}^{(k)}).$$

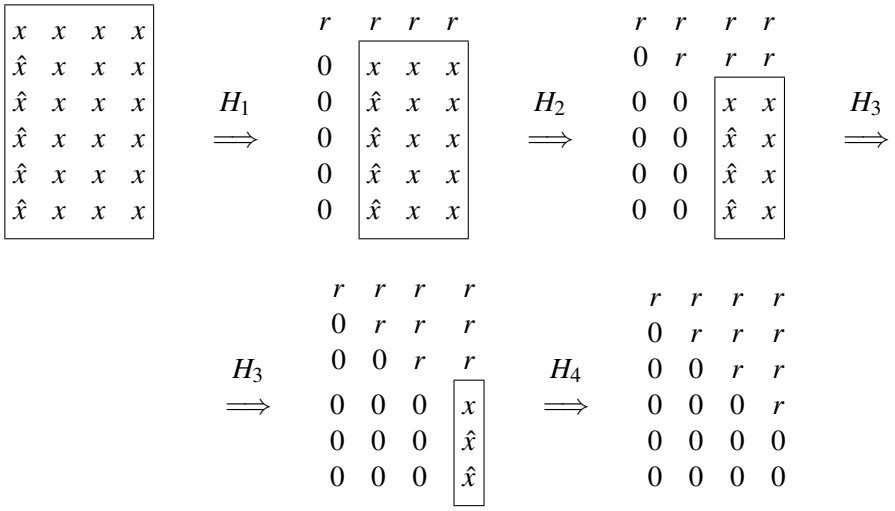

FIGURE 5.2 – Factorisation QR

Algorithme 5.2 Factorisation QR

for $k = 1, \ldots, n$ **do**
  Déterminer le vecteur u_k de la matrice de Householder $\tilde{P}_k$ (par l'équation (5.11)).
  for $j = k+1, \ldots, n$ **do**
    $r_{kj} = \tilde{a}_{kj}^{(k)} - u_k^t \tilde{a}_j^{(k)}$
    $\tilde{a}_j^{(k+1)} = \tilde{a}_j^{(k)} - (u_k^t \tilde{a}_j^{(k)}) u_k$
  end for
end for

Donnons, enfin, une description formelle de l'algorithme, voir l'algorithme 5.3.2.

En pratique, la matrice R écrase la partie supérieure de A, et les vecteurs u_k peuvent être stockés à la place de la partie inférieure (il faut stocker le premier élément de chaque vecteur dans un tableau supplémentaire).

On peut montrer (voir [63]) que le nombre d'opérations flottantes est (au premier ordre) $2mn^2 - \frac{2}{3}n^3$. Le calcul de la factorisation QR est donc plus cher que la formation, et la factorisation, de la matrice des équations normales. Le sur-coût n'est toutefois que d'un facteur 2 (l'ordre de complexité est le même). De plus, (voir les références citées au début de ce chapitre), les propriétés numériques de la méthode QR sont supérieures (dues au fait que l'on travaille avec des matrices orthogonales), et en font la méthode de choix pour une implémentation robuste. En pratique, les bibliothèques numériques fournissent toutes une version de la résolution de problèmes de moindres carrés basées sur la factorisation QR. Dans les programmes interactifs comme Scilab où Matlab, la commande `x= A\z` résolvent le problème de moindres carrés quand A est rectangulaire, et sont basées sur la factorisation QR de A.

La décomposition QR possède une importante interprétation en terme d'orthogonalisation. Pour préciser cela, partitionnons la matrice Q en $Q = (Q_1\ Q_2)$, avec $Q_1 \in \mathbf{R}^{m\times n}$, $Q_1 \in \mathbf{R}^{(m-n)\times n}$. Les colonnes de Q_1 forment une base orthogonale de l'image de A, celle de Q_2 une base orthonormale

de $\mathrm{Im}A^{\perp}$, et l'on a la factorisation réduite

$$A = Q_1 R. \tag{5.14}$$

Cette factorisation montre que, pour $k \leq n$, les k premières colonnes de Q forment une base orthonormale du sous-espace engendré par les k premières colonnes de A. De plus, comme R est inversible, ces sous-espace sont égaux pour chaque k.

L'algorithme 5.3.2 construit donc une base orthonormale pour *tous* les sous espaces engendrés par les colonnes de A. On trouve le résultat obtenu habituellement par l'orthogonalisation de Gram–Schmidt. Il y a toutefois deux différences notables :

- l'algorithme de Gram–Schmidt est notoirement *instable* : les vecteurs calculés par cette méthode ne sont pas numériquement orthogonaux. L'algorithme QR calcule le même résultat de façon numériquement stable (voir ci-dessous). Il est possible d'utiliser une variante (l'algorithme de Gram–Schmidt modifié), qui possède de meilleures propriétés de stabilité, mais celles de l'algorithme QR sont encore meilleures.
- L'algorithme de Gram–Schmidt ne calcule que la matrice que nous avons notée Q_1. Comme nous l'avons vu, l'algorithme QR fournit en plus la matrice Q_2. Selon les applications, cette différence peut ou on être importante.

Remarque 5.1. La relation (5.14) montre que R n'est autre que le facteur de Cholesky de A^tA. En effet, $A^tA = R^tQ_1^tQ_1R = R^tR$, puisque Q_1 est orthogonale. Cette remarque montre alors que R est déterminée de façon unique par A (si A est de rang n), et donc $Q_1 = AR^{-1}$ l'est également. Par contre, Q_2 n'est déterminée que par la condition que $(Q_1\ Q_2)$ soit orthogonale, et n'est donc pas unique.

La méthode de Householder possède d'excellentes propriétés numériques. Dans la proposition suivante, nous notons ε_{M} la précision machine (c'est-à-dire le plus petit nombre positif tel que, sur l'ordinateur, $1+\varepsilon_{\mathrm{M}} \neq 1$).

Proposition 5.2. *Soit $\overline{R}$ la matrice calculée numériquement par l'algorithme 5.3.2. Il existe une matrice exactement orthogonale $\tilde{Q}$ (qui n'est* pas *la matrice calculée numériquement) telle que*

$$A + E = \tilde{Q}\begin{pmatrix}\overline{R}\\ 0\end{pmatrix}, \quad \|E\|_2 \leq c(m,n)\varepsilon_M \|A\|_2, \tag{5.15}$$

où $c(m,n)$ est une fonction ń lentement ż croissante de m et de n.

Preuve. Voir [47]. □

Ce résultat exprime la ń stabilité rétrograde ż de l'algorithme QR. Ce concept, introduit par J. Wilkinson, s'est avéré particulièrement fécond dans l'analyse de stabilité des algorithmes. L'idée de base est de ne pas chercher à comparer le résultat de l'algorithme (ce que nous avons noté R) et la valeur calculée (que nous avons noté $\overline{R}$), qui peuvent être très différentes, mais de ń rejeter l'erreur sur les données ż. C'est ce que nous avons fait dans la proposition 5.2, où $\overline{R}$ est le ń facteur R ż exact d'une matrice proche de A.

5.3.3 Résolution du problème de moindres carrés

Une fois la factorisation QR obtenue, il est facile de résoudre le problème de moindres carrés (4.2).

Théorème 5.3. *Soir* $A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$ *la décomposition QR de la matrice A. La solution du problème de moindres carrés est donnée par la résolution du système linéaire*

$$Rx = z_1, \tag{5.16}$$

avec $Q^t z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$.

Preuve. Du fait que la matrice Q est orthogonale, nous avons

$$\|Ax - z\|_2^2 = \left\| Q \begin{pmatrix} R \\ 0 \end{pmatrix} x - z \right\|_2^2 = \left\| \begin{pmatrix} R \\ 0 \end{pmatrix} x - Q^t z \right\|_2^2$$

Posons $Q^t z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ comme indiqué dans l'énoncé du théorème, il vient

$$\|Ax - z\|_2^2 = \|Rx - z_1\|_2^2 + \|z_2\|_2^2 .$$

Le deuxième terme est constant (indépendant de x), et le premier est minimisé en prenant $x = R^{-1} z_1$, ce qui est possible puisque R est inversible. □

Ajoutons quelques mots sur la mise en œuvre effective de la méthode.

Le second membre $Q^t b$ se calcule implicitement en utilisant la représentation de Q^t comme produit de matrices de Householder et (5.9) :

for $k = 1, \ldots, n$ **do**
  $z_{k:m} = z_{k:m} - 2(u_k^t z_{k:m}) u_k$
end for

en notant $z_{k:m}$ le sous vecteur composé des éléments z_k à z_m.

Une fois ce calcul effectué, la résolution du système triangulaire (5.16) se fait en $O(n^2)$ opérations. Le coût principal de la méthode est donc dans la décomposition QR initiale.

Comme conséquence simple de la proposition 5.2, nous avons un résultat de stabilité rétrograde pour le problème de moindres carrés.

Proposition 5.3. *La solution* $\bar{x}$ *calculée au théorème 5.3 est la solution exacte du problème perturbé*

$$\min_{x \in \mathbf{R}^n} \|(A+E)x - (z+f)\|_2 \tag{5.17}$$

où les perturbations E et f vérifient

$$\|f\|_2 \leq c(m,n)\varepsilon_M \|z\|_2 \ \textit{et} \ \|E\|_2 \leq c(m,n)\varepsilon_M \|A\|_2$$

.

En combinant les résultats de la proposition 5.3 et du théorème 5.1, nous pouvons majorer *l'erreur* commise sur la solution :

Corollaire 5.1. *La solution* $\bar{x}$ *calculée au théorème 5.3 vérifie*

$$\|x - \bar{x}\|_2 \leq c(m,n)\varepsilon_M \left\{ \kappa(A) \left(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{1}{\cos\theta} \frac{\|\delta z\|_2}{\|z\|_2} \right) + \kappa(A)^2 \operatorname{tg}\theta \frac{\|\delta A\|_2}{\|A\|_2} \right\} \|x\|_2 , \tag{5.18}$$

x étant la solution exacte de (4.2)

5.4 SVD et méthodes numériques

La méthode que nous avons vue au paragraphe 4.2.1 donne une méthode numérique ayant d'excellentes propriétés de stabilité. Le seul inconvénient en est son coût, plus élevé que la méthode basée sur la factorisation QR. Rappelons, sous forme matricielle, la méthode du paragraphe 4.2.1, en supposant connue la factorisation en valeurs de $A = U_n \Sigma V^t$ (pour simplifier, nous nous bornons au cas où A est de rang n, Σ_1 est donc $n \times n$, inversible) :

Algorithme 5.3 Résolution d'un problème de moindres carrés par SVD

1: Calculer $w = U_n^t z$,
2: Résoudre le système (diagonal) $\Sigma_1 y = w$,
3: Poser $x = Vy$.

Pour que cette méthode soit effectivement implémentable, il reste à voir comment calculer la décomposition (ou la factorisation) en valeurs singulières de A. Cela fait appel à des techniques plus spécialisées, et différentes, de celles que nous avons vue dans ce cours, et nous ne donnerons que des indications générales. Des détails sont disponibles dans les références citées au début du chapitre (en particulier [21] et [9]).

Contrairement aux factorisations LU et QR, un algorithme de calcul de la SVD est nécessairement itératif. Cela est une conséquence de ce que les valeurs singulières de A sont les valeurs propres de A^tA, c'est-à-dire les racines du polynôme caractéristiques de cette dernière matrice. Et il n'existe pas d'algorithme *fini* permettant de calculer les racines d'un polynôme de degré supérieur à 5 (ceci est connu depuis Abel et Galois au XIX$^{\text{ème}}$ siècle).

Au passage, signalons que comme pour les équations normale, la méthode esquissée ci-dessus (calculer les valeurs propres de A^tA) n'est *pas* recommandable.

La méthode la plus courante de calcul de la SVD procède en deux phases :

– on réduit tout d'abord A à une forme bidiagonale

$$B = PAQ$$

où P et Q sont deux matrices orthogonales (comme $B^tB = Q^tA^tAA$, A et B ont les même valeurs singulières). Ce calcul s'effectue en multipliant A à gauche et à droite par des matrices de Householder, et nécessite un nombre fini d'opérations.

– on calcule ensuite les valeurs singulières de B par une variante de l'algorithme QR pour les matrices tridiagonales. Cette phase, qui est en principe itérative, est en pratique plus rapide que la précédente, sauf si on veut calculer également les vecteurs singuliers.

Chapitre 6

Problèmes inverses linéaires

Nous abordons dans ce chapitre les méthodes de régularisation pour les problèmes inverses linéaires. Régulariser un problème mal posé, c'est le ń remplacer ż par un autre, bien posé, de sorte que l'erreur commise soit compensée par le gain de stabilité. Bien entendu, ceci demande a être quantifié, ce que nous ferons. Ce chapitre présente une introduction aux méthodes de régularisation les plus courantes, à savoir la *méthode de Tikhonov*, et la *troncature spectrale*. Bien entendu, il n'est pas possible de reconstituer une information manquante, et la régularisation va conduire à une perte de précision sur la solution, que nous essayerons de quantifier en fonction de l'erreur sur les données. Nous verrons qu'il est possible d'analyser la méthode de Tikhonov dans un cadre (variationnel) très général, mais que la SVD permet de jeter un éclairage particulier sur ces méthodes.

La principale difficulté dans l'application d'une méthode de régularisation à un problème particulier est la détermination du paramètre de régularisation lui-même. Nous dirons quelques mots sur des solutions possibles. Nous terminerons par quelques mots sur les méthodes itératives, en analysant la plus simple d'entre elles, la méthode de Landweber.

Dans ce chapitre, A désigne un opérateur linéaire continu d'une espace de Hilbert E dans un espace de Hilbert F, et nous supposerons que ce problème est *mal posé*, c'est-à-dire que A n'est pas inversible dans $\mathcal{L}(E,F)$. Comme nous l'avons vu au chapitre 4, cela peut être du à ce que A n'est pas injectif, mais le cas le plus intéressant est celui ou l'image de A n'est pas fermée, ce qui sera toujours le cas si A est compact. Dans ce cas A peut où non être bijectif, mais, s'il existe, l'inverse ne sera pas continu.

6.1 La méthode de Tikhonov

Pour résoudre l'instabilité évoquée au paragraphe précédent, nous allons introduire une information *a priori*.

Nous nous donnons donc un estimé *a priori* $x_0 \in E$. Pour un nombre $\varepsilon > 0$ (le coefficient de régularisation), nous remplaçons (4.2) par le *problème régularisé* :

$$\min\left\{\frac{1}{2}\|Ax-\hat{z}\|_F^2+\frac{\varepsilon^2}{2}\|x-x_0\|_E^2\right\} \tag{6.1}$$

Nous allons voir que ce problème admet une solution unique, qui dépend continûment de $\hat{z}$, et qui converge, lorsque $\varepsilon \to 0$, vers la solution la plus proche de x_0 de (4.2). Évidemment, si ε est choisi trop petit, (6.1) sera proche de (4.2), donc mal posé, alors que si ε est trop grand (6.1) ne sert qu'à

forcer x à être proche de x_0. Le choix ń optimal ż de ε est donc délicat. Nous reviendrons sur ce point au paragraphe 6.1.1.

Commençons par le résultat suivant, qui montre que (6.1) est encore un problème de moindre carrés :

Lemme 6.1. *Posons* $\tilde{A} = \begin{pmatrix} A \\ \varepsilon I \end{pmatrix}, \quad \tilde{z} = \begin{pmatrix} \hat{z} \\ \varepsilon x_o \end{pmatrix}, \quad \tilde{A} \in \mathcal{L}(E, F \times E)$

Alors (6.1) est équivalent à

$$\min_{x \in E} \frac{1}{2} \|\tilde{A}x - \tilde{z}\|^2_{F \times E} \tag{6.2}$$

Preuve. En effet, on calcule simplement :

$$\tilde{A}x - \tilde{z} = \begin{pmatrix} Ax - \hat{z} \\ \varepsilon(x - x_0) \end{pmatrix}$$

Il suffit ensuite de calculer le carré de la norme. □

Ce résultat va nous permettre facilement de montrer que (6.1) possède une solution unique, mais il sert également de base aux méthodes numériques pour résoudre (6.1) (voir [25, 37] et le paragraphe 6.1.2). Une conséquence du lemme (6.1) et du théorème (4.2) est la formulation suivante de (6.1).

Proposition 6.1. *Le problème (6.1) est équivalent à :*

$$(A^*A + \varepsilon^2 I)x = A^*z + \varepsilon^2 x_0. \tag{6.3}$$

Ce problème (et donc (6.1)) admet une solution unique, qui dépend continûment de $\hat{z}$.

Preuve. En effet, (6.3) n'est autre que l'équation normale pour (6.1). On l'obtient en remarquant que :

$$\tilde{A}^* = (A^*, \varepsilon I).$$

En ce qui concerne l'existence et l'unicité de solutions à (6.3), notons que

$$((A^*A + \varepsilon^2 I)x, x) = \|Ax\|^2 + \varepsilon^2 \|x\|^2 \geq \varepsilon^2 \|x\|^2.$$

On applique alors le lemme de Lax-Milgram. D'après le théorème de l'application ouverte A.2, l'opérateur $A^*A + \varepsilon^2 I$, continu et bijectif, a un inverse continu. En fait, nous pouvons obtenir unn estimation explicite en prenant le produit scalaire de l'équation (6.3) avec x, il vient :

$$\|Ax\|^2 + \varepsilon^2 \|x\|^2 \leq \|A^\star \hat{z}\| \|x\| + \varepsilon^2 \|x_0\| \|x\|$$

c'est-à-dire :

$$\|x_\varepsilon\| \leq \frac{1}{\varepsilon^2} \|A^\star \hat{z}\| + \|x_0\| \tag{6.4}$$

□

Remarque 6.1. L'estimation (6.4) ń explose ż quand $\varepsilon \to 0$. Ceci est normal, puisque la solution $\hat{x}$ de (4.2) ne dépend pas de façon continue de $\hat{z}$.

Nous voulons maintenant aborder la question de savoir dans quelle mesure la méthode de Tikhonov a bien régularisé le problème de départ. Pour cela, il sera naturel de se placer dans le cas d'une donnée bruitée, puisqu'alors nous n'avons pas d'estimation sur l'erreur commise sur la solution. Nous allons montrer que la méthode de Tikhonov donne une telle estimation, même si elle est ń non optimale ż, c'est-à-dire d'une ordre plus faible que l'erreur sur la donnée.

Nous supposons connue une observation ń idéale ż $\hat{z} \in \mathrm{Im}A$, et également une suite de mesures bruitées $z_n \in F$, $z_n \notin \mathrm{Im}A$, avec $\delta_n = \|z_n - \hat{z}\|_F \underset{n\to\infty}{\longrightarrow} 0$. Le nombre $\|z_n - \hat{z}\|_F / \|\hat{z}\|_F$ est le *rapport signal sur bruit*. L'hypothèse sous-jacente dans ce paragraphe est que ce rapport tend vers 0, c'est-à-dire que l'on est capable de le réduire arbitrairement, ce qui est évidemment irréaliste.

Considérons tout d'abord la suite de problèmes :

$$\text{Trouver } x_\varepsilon^n \in E \text{ réalisant le minimum de } \frac{1}{2}\|Ax - z_n\|_F^2 + \varepsilon^2 \|x - x_0\|_E^2 , \tag{6.5}$$

(x_ε^n existe et est unique d'après la proposition 6.1). Remarquer que nous ne cherchons pas, pour l'instant, à adapter le paramètre de régularisation au niveau de bruit.

Pour comprendre comment ń fonctionne ż la méthode de régularisation, cherchons à estimer l'erreur entre la solution du problème bruité et la solution exacte. Pour simplifier, nous ferons l'hypothèse (de régularité) que $\hat{x} \in \mathrm{Im}A^\star$. Le cas général est traité par Baumeister [7].

Proposition 6.2. *Supposons que $\hat{x} \in \mathrm{Im}A^\star$, soit $\hat{x} = A^\star w$. Alors,*

$$\forall n,\ \|x_\varepsilon^n - \hat{x}\|_E \leq \|A^\star\| \frac{\delta_n}{\varepsilon^2} + \frac{\varepsilon}{\sqrt{2}} \|w\|_F$$

Preuve. Introduisons la quantité intermédiaire x_ε solution du problème

$$A^\star A x_\varepsilon + \varepsilon^2 x_\varepsilon = A^\star \hat{z} + \varepsilon^2 x_0 \tag{6.6}$$

L'inégalité triangulaire donne :

$$\|x_n - \hat{x}\| \leq \|x_n - x_\varepsilon\| + \|x_\varepsilon - \hat{x}\| \tag{6.7}$$

Nous allons estimer séparément chaque terme du second membre ci-dessus. La première partie correspond à l'erreur sur les données, et est amplifié par le caractère mal posé du problème sous-jacent, alors que le second est du à l'approximation de la solution exacte, et tend vers 0 avec ε.

En soustrayant (6.6) de l'équation normale associée à (6.5), il vient :

$$\varepsilon^2(x_\varepsilon^n - x_\varepsilon) + A^\star A(x_\varepsilon^n - x_\varepsilon) = A^\star(z_n - \hat{z}),$$

puis en prenant le produit scalaire avec $x_\varepsilon^n - x_\varepsilon$

$$\varepsilon^2 \|x_\varepsilon^n - x_\varepsilon\|^2 + \|A(x_\varepsilon^n - x_\varepsilon)\|^2 \leq \|A^\star\|\ \|z_n - \hat{z}\|\ \|x_\varepsilon^n - x_\varepsilon\|$$

et en particulier

$$\varepsilon^2 \|x_\varepsilon^n - x_\varepsilon\| \leq \|A^\star\|\ \delta_n \tag{6.8}$$

c'est une majoration de la première partie de (6.7).

Pour la second partie, écrivons :

$$\begin{aligned}
\|Ax_\varepsilon - A\hat{x}\|^2 &= (Ax_\varepsilon - A\hat{x},\ Ax_\varepsilon - A\hat{x}) \\
&= (A^\star Ax_\varepsilon - A^\star A\hat{x},\ x_\varepsilon - \hat{x}) && \text{par définition de } A^\star \\
&= (-\varepsilon^2 x_\varepsilon - \varepsilon^2\hat{x} + \varepsilon^2\hat{x},\ x_\varepsilon - \hat{x}) && \text{en utilisant les équations normales} \\
&= -\varepsilon^2\|x_\varepsilon - \hat{x}\|^2 + \varepsilon^2(\hat{x},\ x_\varepsilon - \hat{x}) \\
&= -\varepsilon^2\|x_\varepsilon - \hat{x}\|^2 + \varepsilon^2(w,\ A(x_\varepsilon - \hat{x})) && \text{d'après la définition de } w \\
&\leq -\varepsilon^2\|x_\varepsilon - \hat{x}\|^2 + \varepsilon^2\|w\|\ \|A(x_\varepsilon - \hat{x})\| && \text{par Cauchy-Schwarz} \\
&\leq -\varepsilon^2\|x_\varepsilon - \hat{x}\|^2 + \frac{\varepsilon^4}{2}\|w\|^2 + \frac{1}{2}\|A(x_\varepsilon - \hat{x})\|^2 && \text{par l'inégalité élémentaire } ab \leq \frac{a^2+b^2}{2}
\end{aligned}$$

d'où nous tirons :

$$\frac{1}{2}\|A(x_\varepsilon - \hat{x})\|^2 + \varepsilon^2\|x_\varepsilon - \hat{x}\|^2 \leq \frac{\varepsilon^4}{2}\|w\|^2$$

et en particulier

$$\|x_\varepsilon - \hat{x}\| \leq \frac{\varepsilon}{\sqrt{2}}\|w\| \tag{6.9}$$

En regroupant (6.8) et (6.9), nous obtenons l'estimation du théorème. □

Ce résultat montre que, comme nous l'avons signalé, l'erreur se compose de deux termes :
- Un premier terme du aux erreurs sur les données, multiplié par un ń nombre de conditionnement ż, qui tend vers l'infini lorsque $\varepsilon \to 0$.
- Un second terme du à l'approximation de la solution exacte, et qui tend vers 0 avec ε ;

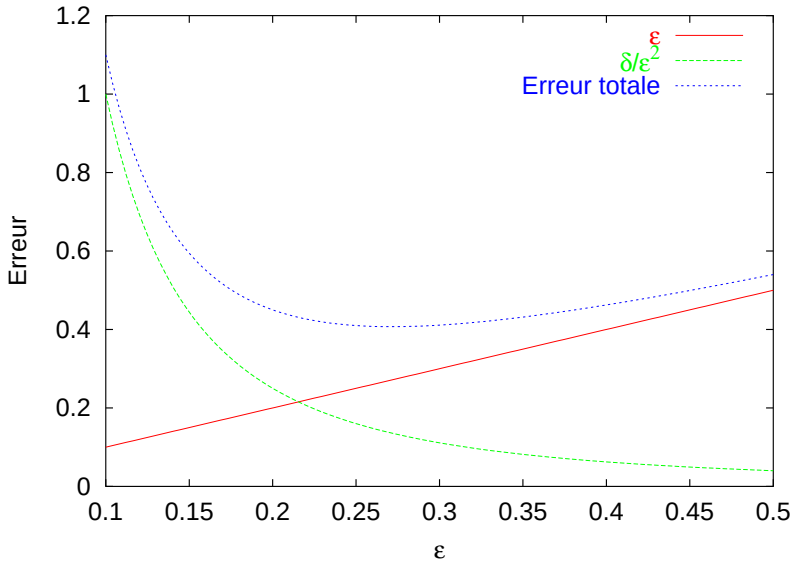

FIGURE 6.1 – Comportement de l'erreur totale

Nous voyons donc bien la nécessité d'adapter le paramètre de régularisation au niveau de bruit présent dans les données. Une telle *stratégie de régularisation* peut se concevoir de deux façons :

– Si l'on possède une estimation du niveau de bruit, on peut en déduire comment il faut choisir ε pour obtenir la convergence de x_ε^n vers $\hat{x}$. C'est ce que nous faisons au théorème suivant. Une telle stratégie s'appelle une stratégie de régularisation *a priori*. Elle suppose que l'on sache estimer le bruit présent dans les données, ce qui n'est pas forcément possible ;
– L'autre stratégie, appelée *a posteriori*, consiste à estimer au cours du calcul la valeur convenable du paramètre, en utilisant uniquement les données disponibles. De telles stratégies existent (voir [25, 37]). Nous en verrons un exemples au paragraphe 6.1.1.

Dans la proposition 6.2, nous avons laissé ε tendre vers 0 indépendamment de δ, et nous avons vu qu'une telle stratégie (ou plutôt absence de stratégie) ne permettait pas la convergence de la solution régularisée vers la ń vraie ż solution. Nous allons donc adapter le paramètre de régularisation au niveau de bruit. Pour cela, nous modifions le problème (6.5) en

$$\text{Trouver } x_n \in E \text{ réalisant le minimum de } \frac{1}{2}\|Ax - z_n\|_F^2 + \varepsilon_n^2 \|x - x_0\|_F^2 , \tag{6.10}$$

où nous cherchons comment choisir la suite ε_n est pour assurer la convergence de x_n vers *une* solution de (4.2). Il se trouve que cette solution sera toujours la solution la plus proche de x_0. Ce résultat la conséquence du lemme suivant.

Lemme 6.2. *Pour $x \in E$, notons la décomposition (unique) :*

$$x = x^K + x^I, \quad x^K \in \mathrm{Ker}A, \quad x^I \in \mathrm{Ker}A^\perp = \overline{\mathrm{Im}A^*}$$

La solution du problème (6.10) *vérifie :*

$$\forall n, \quad x_n^K = x_0^K$$

Preuve. Projetons orthogonalement l'équation (6.3) sur $\mathrm{Ker}A$:
– $A^*Ax_n \in \mathrm{Im}A^* \subset (\mathrm{Ker}A)^\perp$
– $A^*z_n \in \mathrm{Im}A^* \subset (\mathrm{Ker}A)^\perp$
d'où $\varepsilon^2 x_n^K = \varepsilon^2 x_0^K$. □

Nous pouvons maintenant énoncer le théorème de convergence, dont la démonstration élémentaire est empruntée à G. Chavent [14].

Théorème 6.1. *Soit $\hat{z} \in \mathrm{Im}A$. Supposons $\delta_n = \|z_n - \hat{z}\| \underset{n\to\infty}{\longrightarrow} 0, \quad \varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$. Alors :*

i) $\|Ax_n - \hat{z}\|_F \underset{n\to\infty}{\longrightarrow} 0$

ii) *Si $\delta_n/\varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$, alors $\|Ax_n - \hat{z}\|_F = O(\varepsilon_n)$ et $x_n \underset{n\to\infty}{\longrightarrow} \hat{x}$ où $\hat{x}$ est la solution de (4.2) la plus proche de x_0.*

iii) *Si de plus $\hat{x} - x_0 \in \mathrm{Im}A^*$ (hypothèse de régularité), et si $\delta_n/\varepsilon_n^2 \underset{nto\infty}{\longrightarrow} 0$ alors $\|Ax_n - \hat{z}\|_F = O(\varepsilon_n^2)$ et $\|x_n - \hat{x}\|_E = O(\varepsilon_n)$*

Preuve. Notons $\hat{x}$ la solution de (4.3) qui minimise $\|x - x_0\|_E$ ($\hat{x}$ est bien défini, par une variante du corollaire 4.1).

i) Par définition de x_n, nous avons :

$$\|Ax_n - z_n\|_F^2 + \varepsilon_n^2 \|x_n - x_0\|_E^2 \leq \|A\hat{x} - z_n\|_E^2 + \varepsilon_n^2 \|\hat{x} - x_0\|_E^2$$

d'où (en ajoutant et retranchant une quantité positive) :

$$\begin{aligned}\|Ax_n-\hat{z}\|_F^2+\varepsilon_n^2\|x_n-\hat{x}\|_E^2 \le{}& \|Ax_n-z\|_F^2+\varepsilon_n^2\|x_n-\hat{x}\|_E^2\\ &+\|A\hat{x}-z_n\|_F^2+\varepsilon_n^2\|\hat{x}-x_0\|_E^2\\ &-\|Ax_n-z_n\|_F^2-\varepsilon^2\|x_n-x_0\|_E^2.\end{aligned}$$

En utilisant les identités

$$Ax_n-z_n=(Ax_n-\hat{z})+(\hat{z}-z_n),\quad \text{et}\quad x_n-x_0=(x_n-\hat{x})+(\hat{x}-x_0)$$

nous obtenons :

$$\begin{aligned}\|Ax_n-\hat{z}\|_F^2+\varepsilon_n^2\|x_n-\hat{x}\|_E^2 &\le -2(Ax_n-\hat{z},\hat{z}-z_n)-2\varepsilon_n^2(x_n-\hat{x},\hat{x}-x_0)\\ &\le 2\delta_n\|A\hat{x}_n-z\|_F+2\varepsilon_n^2\|x_n-\hat{x}\|_E\|\hat{x}-x_0\|_E\end{aligned}\tag{6.11}$$

par l'inégalité de Cauchy-Schwarz. Si nous notons, dans l'inégalité ci-dessus :

$$\begin{aligned}a&=\|Ax_n-\hat{z}\|_F, & b&=\varepsilon_n\|x_n-\hat{x}\|_E,\\ \alpha&=\delta_n, & \beta&=\varepsilon_n\|\hat{x}-x_0\|_E,\end{aligned}$$

cette inégalité se réécrit :

$$a^2+b^2\le 2a\alpha+2b\beta$$

et nous en déduisons :

$$(a-\alpha)^2+(b-\beta)^2\le\alpha^2+\beta^2\le(\alpha+\beta)^2$$

puis :

$$\begin{cases} a-\alpha\le\alpha+\beta \implies a\le 2\alpha+\beta,\\ b-\beta\le\alpha+\beta \implies b\le\alpha+2\beta,\end{cases}$$

c'est-à-dire finalement :

$$\begin{cases}\|Ax_n-\hat{z}\|_F\le 2\delta_n+\varepsilon_n\|\hat{x}-x_0\|_E\\ \|x_n-\hat{x}\|_E\le\dfrac{\delta_n}{\varepsilon_n}+2\|\hat{x}-x_0\|_E.\end{cases}\tag{6.12}$$

Nous obtenons donc la convergence des observations sous la seule hypothèse que $\hat{z}\in\mathrm{Im}A$. Par contre, et c'est normal puisque à ce stade nous n'avons pas encore lié ε_n à δ_n, nous ne pouvons pas conclure quant à la convergence de x_n vers $\hat{x}$.

ii) Nous supposons de plus que $\delta_n/\varepsilon_n\xrightarrow[n\to\infty]{}0$. Soit alors $\eta>0$ fixé. La définition de $\hat{x}$ entraîne que $\hat{x}-x_0\in\mathrm{Ker}A^\perp=\overline{\mathrm{Im}A^*}$, donc :

$$\exists w\in F,\quad \|\hat{x}-x_0-A^*w\|\le\eta.$$

Reprenons (6.11), en notant que $(x_n-\hat{x},A^*w)=(A(x_n-\hat{x}),w)=(Ax_n-\hat{z},w)$.

$$\begin{aligned}\|Ax_n-\hat{z}\|_F^2+\varepsilon_n^2\|x_n-\hat{x}\|_E^2 &\le -2(A\hat{x}_n-\hat{z},\hat{z}-z_n)-2\varepsilon_n^2(\hat{x}_n-x_0,\hat{x}-x_0)\\ &\quad+2\varepsilon_n^2(Ax_n-\hat{z},w)-2\varepsilon_n^2(x_n-\hat{x},A^*w)\\ &\le 2\|Ax_n-\hat{z}\|_F\left(\delta_n+\varepsilon_n^2\|w\|_F\right)+2\varepsilon_n^2\|x_n-\hat{x}\|_E\eta\end{aligned}$$

Comme précédemment, nous en déduisons :

$$\begin{cases}\bullet\ \|Ax_n-\hat{z}\|_F\le 2\delta_n+2\varepsilon_n^2\|w\|_F+\varepsilon_n\eta=\varepsilon_n\left(2\dfrac{\delta_n}{\varepsilon_n}+2\varepsilon_n\|w\|_F+\eta\right)\\ \bullet\ \|x_n-\hat{x}\|_E\le\dfrac{\delta_n}{\varepsilon_n}+\varepsilon_n\|w\|_F+2\eta\end{cases}\tag{6.13}$$

et, puisque nous avons supposé que $\dfrac{\delta_n}{\varepsilon_n} \to 0$, les deux quantités ci-dessus peuvent être majorées par $3\varepsilon_n\eta$ (resp. 3η) pour n assez grand, et comme η est arbitraire, cela prouve la convergence de x_n vers $\hat{x}$.

iii) Enfin, si nous supposons que $\hat{x} - x_0 \in \mathrm{Im}A^*$, nous pouvons prendre $\eta = 0$ dans les inégalités (6.13). En supposant de plus que $\dfrac{\delta_n}{\varepsilon_n^2} \underset{n\to\infty}{\longrightarrow} 0$, nous obtenons :

$$\begin{cases} \bullet\ \|Ax_n - \hat{z}\|_F \leq \varepsilon_n^2 \left(\dfrac{\delta n}{\varepsilon_n^2} + 2\|w\|_F \right) = O(\varepsilon_n^2) \\ \bullet\ \|x_n - \hat{x}\|_E \leq \varepsilon_n \left(\dfrac{\delta n}{\varepsilon_n^2} + \|w\|_F \right) = O(\varepsilon_n) \end{cases}$$

□

Ce théorème met une fois de plus en évidence le compromis stabilité–précision inhérent aux problèmes mal-posés. Il exprime que la suite ε_n doit tendre vers 0 *moins vite* que le niveau du bruit si l'on veut obtenir la convergence des solutions régularisées. De plus, cette suite doit tendre d'autant moins vite vers 0 que la solution est plus régulière. En ce qui concerne l'erreur sur la solution, elle est d'ordre ε_n, qui est donc plus grand que le niveau du bruit. On a donc une perte de précision, due bien évidemment à l'instabilité. Comme nous l'avons annoncé plus haut, la méthode de Tikhonov n'est pas une méthode de régularisation optimale. Pour une discussion plus approfondie de point, et pour l'analyse d'autre méthode, voir [25].

Remarque 6.2.

i) L'hypothèse $\hat{x} - x_0 \in \mathrm{Im}A^*$ correspond typiquement à un résultat de régularité pour $\hat{x}$, solution de (4.3).

ii) Ce théorème a un intérêt essentiellement théorique. En effet, pour pouvoir l'appliquer, il faudrait connaître la suite δ_n, c'est-à-dire le *niveau de bruit* contenu dans les données, ce qui est difficile en pratique. Le choix de ε_n donné par le théorème 6.1 est ce que nous avons appelé plus haut un choix *a priori*.

Exemple 6.1.

Nous appliquons la méthode de Tikhonov à l'opérateur ń injection canonique ż de l'exemple 4.1. Le problème original est : étant donné $\hat{z} \in L^2(0,1)$, trouver $u \in H_0^1(0,1)$ tel que $u = \hat{z}$. Comme on peut identifier $u \in H_0^1(0,1)$ au couple (u,u'), ce problème revient à retrouver u' à partir d'une mesure de u.

D'après l'équation (4.4), la solution régularisée u_ε est solution du problème elliptque suivant :

$$\begin{cases} -\varepsilon u'' + u = \hat{z} \\ u(0) = u(1) = 0. \end{cases} \tag{6.14}$$

Une calcul avec Maple donne :

$$u(t) = \frac{\sinh(t/\varepsilon)}{\sinh(1/\varepsilon)} \int_0^1 \frac{1}{\varepsilon} \sinh\left(\frac{1-s}{\varepsilon}\right) f(s)\, ds - \int_0^t \frac{1}{\varepsilon} \sinh\left(\frac{t-s}{\varepsilon}\right) f(s)\, ds$$

Comme prévu, u dépend de façon continue de ε, mais le passage à la limite est singulier.

6.1.1 Choix du paramètre de régularisation

Nous allons donner un exemple de méthode de choix à posteriori du paramètre de régularisation. Nous nous conterons d'exposer la plus classique de celles-ci, le ń discrepancy principle ż de Morozov [51], en suivant Kirsch [43].

On suppose toujours que l'on dispose d'une donnée bruitée, mais on ne fait plus l'hypothèse que l'on connaît le niveau de bruit δ. Tout ce dont on dispose est la donnée z^δ. Le discrepancy principle propose de chercher la solution x^ε comme réalisant le minimum de la fonction coût définie en (6.1), en ajoutant la contrainte

$$\left\| Ax_\varepsilon^\delta - z^\delta \right\|_2 = \delta, \tag{6.15}$$

ce qui fournit une équation liant ε à δ.

Une justification heuristique de ce choix est qu'il ne sert à rien de réduire l'erreur en dessous du niveau du bruit.

Commençons par donner quelques propriétés supplémentaires de la solution du problème régularisé. Dans cette proposition δ est fixée, et nous ne ferons pas apparaître explicitement la dépendance par rapport à δ.

Proposition 6.3. *La solution x^ε dépend de façon continue de ε. La fonction $\varepsilon \mapsto \|x^\varepsilon\|_E$ est décroissante et tend vers 0 lorsque $\varepsilon \to \infty$.*

La fonction $\varepsilon \mapsto \|Ax^\varepsilon - z\|_F$ est croissante, et $\lim_{\varepsilon\to 0} Ax^\varepsilon = z$.

Preuve. Par définition de x^ε, on a

$$\varepsilon^2 \|x^\varepsilon\|_E^2 \leq \|Ax^\varepsilon - z\|_F + \varepsilon^2 \|x^\varepsilon\|_E^2 \leq \|z\|_F^2$$

(en prenant $x = 0$ dans la fonction à minimiser), c'est-à-dire

$$\|x^\varepsilon\|_E \leq \sqrt{\|z\|_F}/\varepsilon.$$

Pour la suite de la démonstration, établissons tout d'abord l'égalité :

$$\varepsilon^2 \|x^\varepsilon - x^\eta\|^2 + \|A(x^\varepsilon - x^\eta)\|^2 = (\eta^2 - \varepsilon^2)(x^\eta, x^\varepsilon - x^\eta). \tag{6.16}$$

Cette identité s'obtient en soustrayant les équations normales pour x^ε et x^η :

$$\varepsilon^2(x^\varepsilon - x^\eta) + A^*A(x^\varepsilon - x^\eta) + (\varepsilon^2 - \eta^2)x^\eta = 0$$

et en prenant le produit scalaire avec $(x^\varepsilon - x^\eta)$.

On déduit de (6.16) que

$$\varepsilon^2 \|x^\varepsilon - x^\eta\|_E^2 \leq \left|(\eta^2 - \varepsilon^2)\right| |(x^\eta, x^\varepsilon - x^\eta)| \leq \left|(\eta^2 - \varepsilon^2)\right| \|x^\eta\|_E \|x^\varepsilon - x^\eta\|_E$$

c'est-à-dire

$$\varepsilon^2 \|x^\varepsilon - x^\eta\|_E \leq \left|(\eta^2 - \varepsilon^2)\right| \|x^\eta\|_E \leq \left|(\eta^2 - \varepsilon^2)\right| \frac{\|z\|_F}{\varepsilon}$$

ce qui montre la continuité de la fonction $\varepsilon \to x^\varepsilon$.

Soit maintenant $\eta > \varepsilon$. On déduit de (6.16) que $(x^\eta, x^\varepsilon - x^\eta)$ est positif. Par conséquent,

$$\|x^\eta\|_E^2 \leq (x^\eta, x^\varepsilon) \leq \|x^\eta\|_E \|x^\varepsilon\|_E,$$

d'où la décroissance de l'application $\varepsilon \to \|x^\varepsilon\|_E$

Soit maintenant $\varepsilon > \eta$. Prenons le produit scalaire de l'équation normale pour x^η par $x^\varepsilon - x^\eta$, pour obtenir

$$\eta(x^\eta, x^\varepsilon - x^\eta) + (Ax^\eta - z, A(x^\varepsilon - x^\eta)) = 0.$$

Toujours en utilisant (6.16), nous voyons que cette fois $(x^\eta, x^\varepsilon - x^\eta) \leq 0$, et donc $0 \leq (Ax^\eta - z, A(x^\varepsilon - x^\eta)) = (Ax^\eta - z, Ax^\varepsilon - z) - \|Ax^\eta - z\|_F^2$, et l'inégalité de Cauchy-Schwarz montre que l'application $\varepsilon \to \|Ax^\varepsilon - z\|_F$ est croissante.

Enfin, soit $\alpha > 0$. Comme l'image de A est dense dans F, il existe $x \in E$ tel que $\|Ax - z\|_F \leq \alpha^2/2$. Prenons ε_0 de sorte que $\varepsilon_0^2 \|x\|_E^2 \leq \alpha^2/2$. Alors,

$$\|Ax^\varepsilon - z\|_F^2 \leq \|Ax^\varepsilon - z\|_F^2 + \varepsilon^2 \|x^\varepsilon\|_E^2 \leq \|Ax - z\|_F^2 + \varepsilon^2 \|x\|_E^2 \leq \alpha^2,$$

d'où $\|Ax^\varepsilon - z\|_F^2 \leq \alpha$, pour $\varepsilon \leq \varepsilon_0$. □

En remplaçant z par z^δ, notons x_ε^δ la solution de l'équation (6.3). Nous cherchons $\varepsilon = \varepsilon(\delta)$ tel que la relation (6.15) soit vérifiée.

Proposition 6.4. *Sous l'hypothèse* $\left\|z^\delta - z\right\|_F \leq \delta < \left\|z^\delta\right\|_F$, *l'équation* (6.15) *admet une solution unique.*

Soit $x \in E$ tel que $Ax = z$, et notons x^δ la solution du problème régularisée correspondant. On a

$$\lim_{\delta \to 0} x^\delta = x. \tag{6.17}$$

Preuve. D'après la proposition 6.3, la fonction $\varepsilon \to \left\|Ax_\varepsilon^\delta\right\|_F$ est continue, décroissante avec $\lim_{\varepsilon \to \infty} = \left\|z^\delta\right\|_F > \delta$ et $\lim_{\varepsilon \to 0} = 0 < \delta$. Il existe donc une unique valeur de ε solution de l'équation (6.15).

Nous démontrerons la seconde partie de la proposition sous l'hypothèse supplémentaire (de régularité) $x \in \mathrm{Im} A^*$, avec $x = A^* w$. Pour le cas général, voir [43].

Par définition de x^δ,

$$\left\|Ax^\delta - z\right\|_F^2 + \varepsilon(\delta)^2 \left\|x^\delta\right\|_E^2 \leq \|Ax - z\|_F^2 + \varepsilon(\delta)^2 \|x\|_E^2 = \varepsilon(\delta)^2 \|x\|_E^2 + \left\|z - z^\delta\right\|_F^2 \leq \varepsilon(\delta)^2 \|x\|_E^2 + \delta^2$$

et comme $\left\|Ax^\delta - z\right\|_F = \delta$, on a $\left\|x^\delta\right\|_E \leq \|x\|_E$, pour tout $\delta > 0$. On en déduit

$$\left\|x^\delta - x\right\|_E^2 = \left\|x^\delta\right\|_E^2 - 2(x^\delta, x) + \|x\|_E^2 \leq 2(\left\|x^\delta\right\|_E^2 - (x^\delta, x)) = 2(x - x^\delta, x).$$

Ensuite,

$$\left\|x^\delta - x\right\|_E^2 \leq 2(x - x^\delta, A^* w) = 2(z - Ax^\delta, w) \leq 2(z - z^\delta, w)(z^\delta - Ax^\delta, w) \leq 2\delta \|w\|_F + 2\delta \|w\|_F,$$

c'est-à-dire $\left\|x^\delta - x\right\|_E^2 \leq \sqrt{2\delta \|w\|_F}$. □

L'ordre de convergence obtenu dans la preuve précédente est en $O(\sqrt{\delta})$. On a donc perdu un demi ordre de convergence, et la méthode précédente n'est donc pas optimale.

Pour une valeur de δ donnée, la détermination de ε se ramène à la résolution de l'équation (6.15), et nous avons vu que la fonction considérée est monotone. Une première idée est d'utiliser la méthode de Newton, mais en fait il est préférable d'utiliser la formulation alternative suivante due à Hebden :

$$h(\varepsilon) = \frac{1}{\|Ax^\varepsilon - \hat{z}\|_2} - \frac{1}{\delta^2}.$$

On peut vérifier que la dérivée de l'application $\varepsilon \to x_\varepsilon^\delta$ s'obtient par la résolution de l'équation

$$A^* Ax + \varepsilon^2 x = -\varepsilon x_\varepsilon^\delta,$$

et il est alors facile d'en déduire la dérivée de h.

Il est donc essentiel de pouvoir résoudre un grand nombre de problèmes comme (6.6), pour différentes valeurs de ε. Nous esquissons une méthode numérique efficace au paragraphe suivant.

6.1.2 Méthodes numériques

Nous allons donner quelques indications sur les méthodes numériques utilisées pour résoudre le problème régularisé (6.1). La méthode la plus simple, mais pas la plus efficace, est d'utiliser la forme (6.2). On s'est ramené à un problème de moindres carrés standard, auquel on peut appliquer les méthodes du chapitre 5, en particulier la méthode QR.

Cette approche possède le mérite de la stabilité numérique, et de la simplicité. Elle a par contre le désavantage de demander la formation de la matrice $\tilde{A} = \begin{pmatrix} A \\ \varepsilon I \end{pmatrix}$, de taille $(m+n) \times n$. Dans le cas où ce calcul devra être effectué plusieurs fois (pour estimer le paramètre de régularisation), l'application de cette méthode devient très coûteuse. Nous allons présenter brièvement une méthode alternative, dont le coût est essentiellement indépendant du nombre de systèmes à résoudre, et qui devient particulièrement intéressante si l'on veut déterminer le paramètre de régularisation, comme au paragraphe 6.1.1.

Cette approche procède en deux étapes :

- la première consiste à transformer A en une forme bidiagonale, de la même façon que pour le calcul des valeurs singulières de A (voir le paragraphe 5.4).
- on résout ensuite les différents problèmes, pour chaque valeur du paramètre de régularisation, par une méthode utilisant des rotations de Givens (voir [28], [21]).

La première étape transforme A en une matrice

$$A = U \begin{pmatrix} B \\ 0 \end{pmatrix} V^t$$

où $U \in \mathbf{R}^{m \times m}$ et $V \in \mathbf{R}^{n \times n}$ sont orthogonales, et B est bidiagonale. Le problème (6.2) devient alors :

$$\min \left\| \begin{pmatrix} B \\ \varepsilon I \end{pmatrix} \xi - \begin{pmatrix} U^t z \\ \varepsilon \xi_0 \end{pmatrix} \right\|_2, \tag{6.18}$$

où $\xi = V^t x$ et $\xi_0 = V^t x_0$.

Comme les matrices U et V sont orthogonales, on a les relations

$$\|\xi\|_2 = \|x\|_2, \quad \|Ax - z\|_2 = \left\|B\xi - U^t z\right\|_2.$$

Nous ne décrirons pas la second étape en détail ici. Signalons simplement que l'on transforme orthogonalement la matrice $\begin{pmatrix} B \\ \varepsilon I \end{pmatrix}$ en $\begin{pmatrix} \hat{B} \\ 0 \end{pmatrix}$:

$$\begin{pmatrix} B \\ \varepsilon I \end{pmatrix} = Q_\varepsilon \begin{pmatrix} \hat{B} \\ 0 \end{pmatrix} \tag{6.19}$$

où $Q_\varepsilon \in \mathbf{R}^{2n \times 2n}$ est orthogonale, et $\hat{B} \in \mathbf{R}^{n \times n}$ est encore bidiagonale. Notons que la matrice Q_ε n'est pas formée explicitement, mais sous une forme factorisée analogue à ce que nous avons vu au paragraphe 5.3.2. Le problème initial (6.2) est finalement équivalent à

$$\min \left\| \begin{pmatrix} \hat{B} \\ 0 \end{pmatrix} \xi_\varepsilon - \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \right\|_2, \tag{6.20}$$

dont la solution est simplement $\xi_\varepsilon = \hat{B}^{-1}\hat{\beta}_1$. Notons que la réduction (6.19) et la résolution de (6.20) doivent être effectuées pour chaque valeur de ε, mais leur complexité est seulement $O(n)$.

La solution du problème régularisé initial (6.2) s'obtient alors simplement par $x = V\xi_\varepsilon$.

Pour plus de détails on se reportera à [9] ou [37].

6.2 Applications de la décomposition en valeurs singulières

Nous supposons dans ce paragraphe que l'opérateur A est compact. D'après le théorème 4.5, il admet un développement en valeurs singulières. Nous reprenons les notations du paragraphe 4.3.

6.2.1 Décomposition en valeurs singulières et méthode de Tikhonov

Nous allons maintenant utiliser l'information fournie par la DVS pour régulariser notre problème. Pour cela, nous réinterprétons la méthode de Tikhonov.

Proposition 6.5. *Soit $z \in F$. La solution de l'équation* (6.3) *se développe sur la base (e_j) comme suit :*

$$x = \sum_{j=1}^{+\infty} \frac{\sigma_j}{\sigma_j^2 + \varepsilon^2}(z, f_j)e_j \tag{6.21}$$

Preuve. Comme pour le théorème 4.6, nous développons le second membre et la solution selon leurs bases hilbertiennes respectives, puis nous utilisons les équations de (4.17). On obtient

$$\sum_{j=1}^{+\infty} x_j(\sigma_j^2 + \varepsilon^2)f_j = \sum_{j=1}^{+\infty} \sigma_j(z, f_j)f_j$$

Il suffit ensuite d'identifier les coefficients. □

Remarque 6.3. Contrairement à la suite $(1/\sigma_j)$, la suite $\sigma_j/\sigma_j^2 + \varepsilon^2$ reste bornée quand $j \to \infty$, pour $\varepsilon > 0$, *fixé*. Nous retrouvons ainsi le fait que le problème régularisé est bien posé.

On voit que la DVS nous conduit à une interprétation simple de la méthode de Tikhonov. Nous avons remplacé la fonction $\sigma \mapsto 1/\sigma$, non bornée au voisinage de 0, par la fonction $\sigma \mapsto \sigma/\sigma^2 + \varepsilon^2$, qui reste bornée. Bien évidemment, la borne en question tend vers l'infini avec ε (elle vaut $1/2\varepsilon$, pour $\sigma = \varepsilon$). Nous illustrons ce résultat sur la figure 6.2.

Remarque 6.4. Si nous reprenons le calcul de la remarque 4.7, nous obtenons cette fois :

$$\left\|x' - x\right\|_E = \eta\frac{\sigma_i}{\sigma_i^2 + \varepsilon^2} \leq \frac{\eta}{2\varepsilon} = \frac{1}{2\varepsilon}\left\|z - z'\right\| \tag{6.22}$$

On a donc cette fois une borne indépendante des valeurs singulières de A.

Remarque 6.5. On peut également réécrire la formule (6.21) sous la forme

$$x = \sum_{j=1}^{+\infty} \frac{\sigma_j^2}{\sigma_j^2 + \varepsilon^2}\frac{(z, f_j)}{\sigma_j}e_j, \tag{6.23}$$

les coefficients $\dfrac{\sigma_j^2}{\sigma_j^2 + \varepsilon^2}$ s'appellent les *facteurs de filtrage* de la méthode de Tikhonov. Ils sont évidemment compris entre 0 et 1. Nous pouvons tracer (figure 6.3) l'analogue de la figure 6.2 pour cette représentation. On constate que les facteurs de filtrage correspondant aux grandes valeurs singulières sont peu amortis, ce qui est souhaitable puisque ces valeurs singulières sont les plus représentatives, alors que ceux qui correspondent aux petites valeurs singulières (plus précisément celles qui vérifient $\sigma_j < \varepsilon$) sont effectivement filtrées.

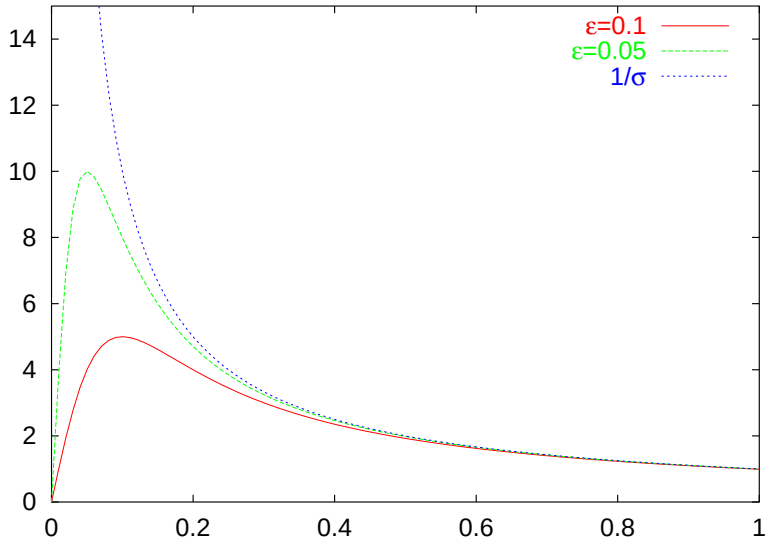

FIGURE 6.2 – Évolution de la fonction ń filtre ż de la DVS avec ε

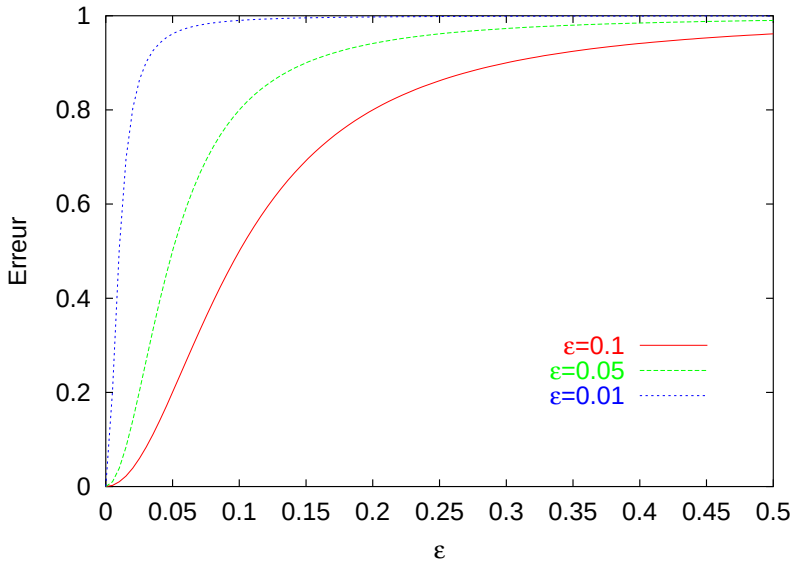

FIGURE 6.3 – Évolution des facteurs de filtrage de la DVS avec ε

6.2.2 Régularisation par troncature spectrale

Si l'on connaît la DVS de l'opérateur A, on peut proposer une autre méthode de régularisation, appelée troncature spectrale. Cette méthode consiste à tronquer le développement (4.16) à un certain rang. Ce rang joue le rôle du paramètre de régularisation ε dans la méthode de Tikhonov.

Pour $n > 0$, posons

$$x_n = \sum_{j=1}^{n} \frac{(z, f_j)}{\sigma_j} e_j. \tag{6.24}$$

Puisque la série converge, il est clair que $x_n \underset{j\to\infty}{\longrightarrow} \hat{x}$ (où $\hat{x}$ est la solution de norme minimale de (4.2)). La méthode de troncature converge donc sur des données non bruitées. Examinons maintenant la cas de données bruitées, c'est-à-dire que nous remplaçons $\hat{z}$ par z_δ, avec $\|\hat{z} - z_\delta\| = \delta$.

Notons

$$x_n^\delta = \sum_{j=1}^{n} \frac{(z^\delta, f_j)}{\sigma_j} e_j$$

qui est ce que l'on peut effectivement calculer, et cherchons à estimer la différence $x_n - x_n^\delta$.

$$x_n - x_n^\delta = \sum_{j=1}^{n} \frac{(z - z^\delta, f_j)}{\sigma_j} e_j$$

et donc, puisque le système e_j est orthonormal,

$$\left\| x_n - x_n^\delta \right\|^2 = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \left| (z - z^\delta, f_j) \right|^2 \leq \frac{1}{\sigma_n^2} \sum_{j=1}^{n} \left| (z - z^\delta, f_j) \right|^2 \leq \frac{1}{\sigma_n^2} \delta^2. \tag{6.25}$$

En ce qui concerne l'erreur totale, écrivons :

$$\left\| \hat{x} - x_n^\delta \right\| \leq \|\hat{x} - x_n\| + \left\| x_n - x_n^\delta \right\| \leq \|\hat{x} - x_n\| + \frac{\delta}{\sigma_n}. \tag{6.26}$$

Le premier terme tend vers 0, et pour que la somme tende vers 0, il faut une fois de plus choisir n en fonction de δ de sorte que si, par exemple, on considère une suite $z_n \underset{n\to\infty}{\longrightarrow} \hat{z}$, avec $\|\hat{z} - z_n\| = \delta_n$, alors $\dfrac{\delta_n}{\sigma_n} \underset{n\to\infty}{\longrightarrow} 0$.

Pour obtenir un taux de convergence, nous devrons ici aussi faire une hypothèse de régularité. Supposons donc que $\hat{x} \in \mathrm{Im} A^\star$, soit $\hat{x} = A^\star w,\ w \in F$. Dans ce cas, en utilisant (4.17) :

$$x_n - \hat{x} = \sum_{j=1}^{n} \frac{(z, f_j)}{\sigma_j} e_j - \sum_{j=1}^{+\infty} \sigma_j (w, f_j) e_j = \sum_{j=1}^{n} \left[\frac{(z, f_j)}{\sigma_j} - \sigma_j (w, f_j) \right] e_j + \sum_{j=n+1}^{+\infty} \sigma_j (w, f_j) e_j$$

et donc (les valeurs singulières forment une suite décroissante) :

$$\|x_n - \hat{x}\|_E \leq \sigma_{n+1} \|w\|_F .$$

L'équation (6.26) devient alors :

$$\left\| \hat{x} - x_n^\delta \right\| \leq \sigma_{n+1} \|w\|_F + \frac{\delta_n}{\sigma_n}. \tag{6.27}$$

Une façon simple d'assurer la convergence est, par exemple, de choisir n de sorte que $\delta_n/\sigma_n = C\sigma_{n+1}$ (où C est une constante). L'inégalité (6.27) donne alors $\left\| \hat{x} - x_n^\delta \right\| \leq C' \sigma_{n+1} = O(\sqrt{\delta_n})$.

Remarque 6.6. Comme pour la méthode de Tikhonov, le choix du paramètre de régularisation (ici n) est à la fois important et délicat. Si n est choisi trop grand, l'approximation ressemble trop à celle instable donnée par la formule (4.20), et l'influence des petites valeurs singulières conduit à l'instabilité. Si n est trop petit, la solution obtenue n'aura que peu en commun avec la solution ń réelle ż. Dans cette discussion, il faut bien entendu prendre en compte la rapidité plus ou moins grande avec laquelle les valeurs singulières tendent vers 0. Plus cette décroissance est rapide, plus n devra être choisi petit, c'est-à-dire que peu de composantes de la solution pourront être retrouvées.

6.3 Méthodes itératives

Dans ce paragraphe, nous voulons introduire une classe de méthodes différentes pour résoudre les problèmes inverses linéaires : les méthodes itératives. Les méthodes (régularisation de Tikhonov, troncature spectrale) que nous avons vues précédemment sont qualifiées de ń directes ż parce qu'elles donnent (dans le cas de la dimension finie) la solution *exacte* du problème régularisé (aux erreurs d'arrondis près) en un nombre fini dt'opérations. Pour des problèmes de taille modérée, ces méthodes sont les plus utilisées. Toutefois, pour des problèmes de grande taille, ou les matrices obtenues après discrétisation sont souvent creuses, ces méthodes s'avèrent inadaptées car d'une part le nombre d'opérations devient trop grand, et d'autre part, ces méthodes ne respectent pas la structure des matrices.

Une alternative est alors de se tourner vers des méthodes itératives, qui construisent une *suite* de solutions approchées qui (dans le cas non bruité) convergent vers la solution désirée. Nous verrons que dans le contexte des problèmes inverses la situation est plus compliquée : en présence de bruit, la suite construite par la méthode itérative ne converge pas, en général, vers une solution du problème de départ. Il est, encore une fois, nécessaire de régulariser le processus itératif, et c'est l'indice d'itération lui-même qui joue le rôle de paramètre de régularisation. En d'autres termes, il convient d'arrêter les itérations plus tôt qu'on ne le ferait dans un cas non bruité.

Nous n'examinerons dans ce paragraphe que la plus simple des méthodes itératives : la méthode de Landweber [45], qui a pour principal avantage de se prêter à une analyse simple. Malheureusement, elle converge trop lentement pour être utilisable en pratique, d'autant plus que des méthodes beaucoup plus performantes existent. Les deux plus importantes sont la méthode ń ν ż de Brakhage (voir [25, par 6.3], et surtout la méthode du gradient conjugué et ses variantes. Cette dernière méthode est celle qui est le plus communément employée. Dans le contexte des problèmes mal posés, un exposé accessible se trouve dans le livre de Kirch [43], des exposés plus complets sont dans [25, 35] (cette dernière référence est consacrée entièrement à l'analyse des méthodes de type gradient conjugué pour les problèmes mal posés).

Étant donné un paramètre de relaxation ω, la méthode de Landweber pour résoudre le problème de moindres carrés (4.3) est définie par la formule de récurrence

$$x_{n+1} = x_n + \omega(A^\star \hat{z} - A^\star A x_n), \tag{6.28}$$

Nous prendrons $x_0 = 0$ pour simplifier (la cas général est discuté dans [43]). Dans ce cas, on voit immédiatement par récurrence sur n que $x_n \in \mathrm{Im}A \subset \mathrm{Ker}A^\perp$. Par conséquent, si x_n converge vers un élément $x \in E$, on doit avoir $x \in \mathrm{Ker}A^\perp$ (puisque ce ce sous-espace est fermé). De plus, par continuité, x satisfait l'équation (4.3), et donc $x = \hat{x}$ la solution de norme minimale de (4.2). Il reste donc à trouver une condition sur ω qui assure la convergence de x_n vers $\hat{x}$, puis à montrer que, dans le cas bruité l'on peut choisir n en fonction du niveau de bruit.

Pour simplifier, nous nous restreignons au cas où A est un opérateur compact, et nous notons $(\sigma_j)_{j\in\mathbf{N}}$ ses valeurs singulières, et $(u_j)_{j\in\mathbf{N}}$, $(v_j)_{jn\mathbf{N}}$ ses vecteurs singuliers. L'analyse qui suit est essentiellement celle de Groetsch [31].

Proposition 6.6. **i)** *On a*

$$x_n = \omega \sum_{j=0}^{n-1} (I - \omega A^\star A)^j A^\star \hat{z}; \tag{6.29}$$

ii) *Si on choisit* $0 < \omega < 2/\sigma_1^2$*, la suite des itérés* x_n *de la méthode de Landweber converge vers* $\hat{x}$ *pour* $\hat{z} \in \mathrm{Im}A$.

iii) *Sous l'hypothèse (de régularité) supplémentaire* $\hat{x} \in \mathrm{Im}A^\star$*, on a de plus l'estimation*

$$\|x_n - \hat{x}\| = O(1/n) \tag{6.30}$$

Preuve. **i)** L'équation (6.29) se démontre aisément par récurrence.

ii) Notons $e_n = x_n - \hat{x}$ l'erreur. En soustrayant (6.28) de (4.3), nous obtenons pour $n > 0$

$$e_n = (I - \omega A^\star A) e_{n-1} = (I - \omega A^\star A)^n e_0. \tag{6.31}$$

Si nous introduisons le développement en valeurs singulières de $e_0 = \sum_{j=1}^{\infty} (e_0, u_j) u_j$, l'équation (6.31) se réécrit

$$e_n = \sum_{j=1}^{\infty} (e_0, u_j)(1 - \omega\sigma_j^2)^n u_j$$

Comme les vecteurs singuliers sont orthogonaux, on en déduit

$$\|e_n\|^2 = \sum_{j=1}^{\infty} \left|(e_0, u_j)\right|^2 |1 - \omega\sigma_j^2|^{2n}, \tag{6.32}$$

et si $\omega < 2/\|A\|^2 = 2/\sigma_1^2$, chaque terme $|1 - \omega\sigma_j^2|$ est en valeur absolue strictement inférieur à 1. Pour passer à la limite dans (6.32) nous utilisons le théorème de convergence dominée de Lebesgue. Tout d'abord, chaque terme de la somme tend vers 0, et est majoré par $\left|(e_0, u_j)\right|$. Or d'après l'inégalité de Bessel

$$\sum_{j=1}^{\infty} \left|(e_0, u_j)\right|^2 \leq \|e_0\|^2 .$$

Le théorème de Lebesgue permet donc de conclure que $e_n \to 0$, et donc que la suite x_n converge vers $\hat{x}$.

iii) Par hypothèse, il existe $w \in E$, tel que $\hat{x} = A^\star w$ et (6.32) devient

$$\|e_n\|^2 = \sum_{j=1}^{\infty} \sigma_j^2 \left|(w, v_j)\right|^2 \left|1 - \omega\sigma_j^2\right|^{2n}$$

En utilisant alors l'inégalité (démontrée plus bas)

$$\left|1 - \omega\sigma_j^2\right|^n \leq \frac{1}{n\omega} \frac{1}{\sigma_j^2} \tag{6.33}$$

nous obtenons

$$\|e_n\|^2 \leq \frac{1}{n^2\omega^2} \sum_{j=1}^{\infty} \left|(w, v_j)\right|^2 = \frac{1}{n^2\omega^2} \|w\|_F^2 \tag{6.34}$$

c'est l'estimation du théorème.

□

Démontrons maintenant l'inégalité (6.33), ainsi qu'une autre qui nous sera utile pour la démonstration du théorème suivant.

Lemme 6.3. *On a les inégalités, valable pour tout* $x \in]0,1[$

$$(1-x)^n \leq \frac{1}{nx} \tag{6.35}$$

$$1-(1-x)^n \leq nx \tag{6.36}$$

Preuve. Pour la première inégalité, il suffit de montrer que

$$nx(1-x)^n \leq 1, \quad \text{pour } x \in [0,1]$$

La fonction en question est positive, les valeurs en 0 et en 1 sont toutes deux nulles, et la dérivée s'annule au point $1/(n+1)$, ou la fonction vaut $\left(\frac{n}{n+1}\right)^{n+1} \leq 1$.

La seconde inégalité est une conséquence de la concavité de la fonction $x \mapsto (1-x)^n$:

$$1-(1-x)^n \leq 1-(1-nx) = nx$$

□

Nous voyons donc que la méthode itérative converge pour des données non bruitées. Considérons maintenant ce qui se passe si nous remplaçons z par z^δ avec $\left\|z-z^\delta\right\| = \delta$. Nous noterons $x^{n,\delta}$ la suite des itérés correspondants, et nous introduirons la suite x^n construite à partir de la donnée non bruitée z. Nous voulons estimer l'erreur $x^{n,\delta} - \hat{x}$, et montrer que nous pouvons choisir n en fonction de δ pour obtenir la convergence.

Théorème 6.2. **i)** *Si n est choisi tel que* $n\delta \underset{n\to\infty}{\to} 0$*, on a* $\lim_{\delta\to 0} x^{n(\delta),\delta} = \hat{x}$.

ii) *Sous l'hypothèse supplémentaire* $\hat{x} \in \mathrm{Im}A^\star$*, on peut choisir* $n(\delta) = E(1/\sqrt{\delta})$ *et alors* $\left\|x^{n(\delta),\delta} - \hat{x}\right\| = O(\sqrt{\delta})$ *(E est la fonction partie entière).*

Preuve. Nous utilisons encore une fois l'inégalité triangulaire pour écrire :

$$\left\|x^{n,\delta} - \hat{x}\right\| \leq \|x^n - \hat{x}\| + \left\|x^{n,\delta} - x^n\right\|$$

Dans le premier cas, nous savons que le premier terme tend vers 0, dans le second nous savons de plus qu'il se comporte (au moins) comme $O(1/n)$.

Pour estimer le second terme, soustrayons l'équation définissant $x^{n,\delta}$ de celle définissant x^n. En notant $d_n = x^{n,\delta} - x^n$, il vient :

$$d^{n+1} = (I - \omega A^\star A)d^n + \omega A^\star(z^\delta - z),\ d_0 = 0.$$

Cette récurrence est la même que celle qui définit la suite x_n originale. Nous pouvons donc utiliser (6.29). Un raisonnement semblable à celui de la proposition 6.6 conduit à la représentation :

$$\left\|d^{n+1}\right\|^2 = \sum_{j=1}^{\infty} \frac{1-(1-\omega\sigma_j^2)^{2n}}{\sigma_j^2} \left|(z^\delta - z, v_j)\right|^2. \tag{6.37}$$

soit, d'après le lemme 6.3 :

$$\left\|d^{n+1}\right\|^2 \leq \omega n \sum_{j=1}^{\infty} \left|(z^\delta - z, v_j)\right|^2 = n\omega \left\|z^\delta - z\right\|^2 \leq n\omega\delta. \tag{6.38}$$

i) Dans le cas général, on a seulement

$$\left\| x^{n,\delta} - \hat{x} \right\| \leq \|x^n - \hat{x}\| + n\omega\delta$$

et le premier terme tend vers 0. La condition $n\delta \to 0$ suffit à assurer la convergence.

ii) Dans ce cas on peut préciser la convergence, en utilisant ii) du théorème précédent.

$$\left\| x^{n,\delta} - \hat{x} \right\| \leq \frac{1}{n\omega} \|w\|_F^2 + n\omega\delta,$$

La somme au second membre est minimale quand $n\omega = \|w\|_F \delta^{-1/2}$, et le minimum vaut $\|w\|_F \delta^{1/2}$.

□

Comme nous l'avions annoncé au début de ce paragraphe, pour obtenir la convergence de la méthode itérative, nous devons arrêter les itérations à un niveau dépendant du niveau de bruit : d'autant plus tôt que le bruit est plus fort. Par ailleurs, comme pour la méthode de Tikhonov, ce résultat suppose que l'on sache évaluer le niveau de bruit. Enfin, attirons une nouvelle fois l'attention du lecteur sur le fait que la méthode de Landweber converge trop lentement pour être utilisable en pratique. Il existe d'autres méthodes itératives, comme la méthode du gradient conjugué, qui converge bien plus rapidement pour un coût essentiellement équivalent.

Troisième partie

Problèmes non–linéaires

Chapitre 7

Problèmes inverses non-linéaires — généralités

Nous abordons maintenant les problèmes inverses non-linéaires, et nous nous concentrerons sur l'identification de paramètres dans les équations différentielles ou aux dérivées partielles.

Dans la situation générale, nous serons en présence d'un phénomène physique dont la *structure* est connue, mais dont les paramètres précis de fonctionnement ne le sont pas. Il est possible de mesurer certaines propriétés de ce système, correspondant à des entrées connue. Le système fonctionne comme un boite noire, et nous voudrions en connaître le contenu, sans ń ouvrir ż la boite.

Nous avons vu des exemples au chapitre 1. Reprenons l'exemple 1.2 : on *sait* (plus précisément, on fait l'hypothèse) que la conduction de la chaleur obéit à l'équation (2.7), mais nous ne connaissons pas le coefficient c (qui peut être un scalaire, ou une fonction de la position). Par contre, nous supposons que nous avons accès à une *mesure* de la température T, dans une partie du domaine (ou de la frontière). Avec ces informations, nous voulons déterminer le coefficient c qui permette de reproduire ces mesures.

La première différence avec les chapitres précédents est que l'application entre le paramètre (c dans notre exemple) et la mesure est non seulement non-linéaire, mais est exprimé par l'intermédiaire d'une équation comme (2.7), et la mesure est alors une ń partie ż de la solution de cette équation. Une équation jouant le rôle de (2.7) s'appelle une *équation d'état*, et la variable T, solution de cette équation, s'appelle l'*état* du système. Il sera en général irréaliste de supposer que l'état tout entier du système est connu : toujours dans notre exemple thermique, il ne sera pas possible de mesurer la température en tout point du domaine.

Une seconde différence, plus pratique, et qu'il est plus difficile d'obtenir des résultats théoriques que dans le cas linéaire. Les résultats qui ont été obtenus sont souvent liés à un problème particulier. Nous pouvons citer le livre [6] pour des éléments d'une théorie générale, ainsi que les articles de G. Chavent (par exemple [16]). En tout état de cause, dans le cadre de ce cours nous n'aborderons pas ces questions (ce qui ne veut pas dire qu'elles soient moins importantes), et nous nous concentrerons sur les méthodes numériques, et tout particulièrement sur la formulation aux moindres carrés.

7.1 Les trois espaces fondamentaux

Pour donner une formulation abstraite des problèmes que nous considérerons dans la suite de ce cours, nous allons introduire 3 espaces de Hilbert, ainsi que des applications entre ces espaces. Dans

toutes les applications que nous considérerons par la suite, ces espaces seront tous de dimension finie. Il nous paraît toutefois utile de donner les définitions dans ce cadre plus général.

- l'espace des *modèles* (ou paramètres) M ;
- l'espace *d'état* U ;
- l'espace des *données* (ou observations) D ;

Comme nous l'avons signalé, l'introduction de l'espace U permet de rendre explicite la dépendance entre le paramètre et les données. Par contre, l'existence de l'état ne dispense pas d'introduire l'observation, puisqu'il est en général non mesurable.

Deux application mettent en évidence les relations entre ces 3 espaces :

L'équation d'état relie de façon implicite le paramètre et l'état (tous deux peuvent évidemment être des vecteurs). Nous l'écrirons

$$F(a,u) = 0, \quad a \in M,\ u \in U,\ F(a,u) \in Z \tag{7.1}$$

où Z est un autre espace de Hilbert. Nous supposons qu'il existe un sous-espace $M_{\text{ad}} \subset M$ tel que pour tout $a \in M_{\text{ad}}$, F définit localement un état unique $u = u_a$. Cela était le cas pour tous les exemples du chapitre 2 avec $M_{\text{ad}} = M$, puisque l'application F était linéaire par rapport à u. Dans les exemples 2.7 et 2.4, l'équation d'état était respectivement (2.15) et le système (2.12), (2.13).

Il sera pratique de noter

$$u = S(a) = u_a \tag{7.2}$$

la solution de l'équation d'état (7.1). La première égalité sera commode quand nous voudrons opérer sur cette solution (la dériver par exemple), alors que la seconde est un abus de notation suggestif.

L'équation d'observation extrait de l'état la partie correspondant aux mesures. Cela sera souvent une injection, rarement l'identité (sauf dans des exemples purement pédagogiques). Elle s'écrit

$$d = Hu, \quad u \in U \tag{7.3}$$

Nous avons fait l'hypothèse simplificatrice que l'observation est un opérateur linéaire, indépendant du paramètre. L'extension à une situation plus générale n'est pas difficile, et est laissée au lecteur.

Si nous injectons la solution de (7.1) dans 7.3, nous obtenons l'application qui relie le paramètre à l'observation. Nous la noterons

$$d = \Phi(a) = H\big(S(a)\big) = H(u_a) \tag{7.4}$$

Le problème inverse est alors, étant donnée une observation d_{obs}, de résoudre l'équation :

$$\Phi(a) = d_{\text{obs}}. \tag{7.5}$$

Donnons maintenant quelques exemples pour illustrer les concepts précédents. Nous reviendrons en détail sur ces exemples plus tard dans ce chapitrem puis au chapitre suivant, pour determiner les fonctionnelles correspondantes, puis calculer leurs gradients.

Nous commençons par un exemple simple, sans réelle signification.

Exemple 7.1.
Nous considérons le problème aux limites en une dimension d'espace :

$$\begin{cases} -bu''(x) + cu'(x) = f(x) \quad 0 < x < 1 \\ u(0) = 0,\ u'(1) = 0. \end{cases} \tag{7.6}$$

Dans ce cas, le paramètre est le couple $m = (b,c)$ et $M = \mathbf{R}^2$. Le théorème de Lax–Milgram montre que l'on peut prendre $M_{\text{ad}} = \{(b,c) \in M,\, b > 0\}$, et qu'alors un choix naturel pour U est $U = H^1(0,1)$. Nous pouvons envisager plusieurs possibilités pour l'observation :

- **i)** On suppose que l'on mesure l'état u en tout point de l'intervalle $]0,1[$. L'espace D est alors $L^2(0,1)$, et d_{obs} est une fonction définie sur $]0,1[$. Dans ce cas le problème inverse est (très) *surdéterminé*. On a ici $Hu = u$.
- **ii)** Inversement, on peut supposer que l'on ne mesure u qu'à l'extrémité droite de l'intervalle. Dans ce cas, $D = R$, et d_{obs} est un nombre. Le problème est ici *sous-déterminé*. Cette fois, nous avons $Hu = u(1)$.
- **iii)** Un cas intermédiaire est celui où l'on mesure u non-seulement à l'extrémité de l'intervalle, mais aussi en un point intérieur. Pour fixer les idées, nous supposerons que u est connu en $1/2$. L'espace D est ici $D = \mathbf{R}^2$, et d_{obs} est un couple, et $Hu = (u(1/2), u(1))$. Dans ce cas, il y a exactement autant de données que d'inconnues. Cela ne veut pas dire que le problème soit plus simple, puisqu'il faut toujours prendre en compte le caractère mal-posé du problème inverse.

Dans les trois cas, l'application Φ est définie par (rappelons que $u_{(b,c)}$ désigne la solution de l'équation d'état (7.6) correspondant aux paramètres (b,c)) :

- **i)** $\Phi(b,c) = u_{(b,c)}$ (fonction définie sur $]0,1[$) ;
- **ii)** $\Phi(b,c) = u_{(b,c)}(1)$;
- **iii)** $\Phi(b,c) = (u_{(b,c)}(1/2), u_{(b,c)})(1)$.

Dans les deux derniers cas, l'application Φ va d'un espace de dimension finie dans un autre, mais sa définition fait intervenir la solution du problème aux limites (7.6). □

Exemple 7.2.
Nous prenons maintenant un système instationnaire, que nous supposerons régi par un système d'équations différentielles :

$$\begin{aligned} y'(t) &= f(y(t), a) \quad t \in [0,T] \\ y(0) &= y_0 \end{aligned} \tag{7.7}$$

où y est à valeur dans $\mathbf{R}^d$, le paramètre a est dans $\mathbf{R}^p$ et $f \in C^1(\mathbf{R}^d, \mathbf{R}^p)$. Nous supposerons que l'on mesure y à certains instants $\tau_1, \ldots, \tau_Q$. Le problème est de retrouver a.

On a donc $M = \mathbf{R}^d$. D'après le théorème de Cauchy-Lipschitz [2], le problème de Cauchy (7.7) possède une solution unique dans $C^1(0,T)^m$. Cet espace n'est pas un espace de Hilbert, mais nous pouvons prendre $U = H^1(0,T)^d$.

L'espace d'observation est $D = \mathbf{R}^Q$, et l'opérateur d'observation est

$$H :\ y \mapsto (y(\tau_1, a), \ldots, y(\tau_Q, a)).$$

L'observation sera un Q-uplet de m vecteurs, chaque élément représentant un instant de mesure.

Dans cet exemple, l'application Φ va encore d'un espace de dimension finie dans un autre (en l'occurrence de $\mathbf{R}^p$ dans $\mathbf{R}^Q$), mais sa définition passe, comme dans l'exemple précédent, par l'intermédiaire de la résolution de l'équation différentielle (7.7). □

L'exemple suivant est une généralisation de 7.1, et cette fois correspond à une application réaliste.

Exemple 7.3 (Suite de l'exemple 2.2)**.**
Dans ce cas, le paramètre est la conductivité K. Cette fois, le paramètre ne ń vit ż plus dans un espace de dimension finie, mais est une fonction de la position spatiale. Bien entendu, après discrétisation, nous aurons un nombre fini de paramètres, mais ce nombre dépend justement de la finesse de la discrétisation, qui est censée tendre vers 0 (et le nombre de paramètres tend vers l'infini). Il est important de prendre en compte le caractère intrinsèquement *distribué* du paramètre à identifier, et de repousser le choix d'une représentation en dimension finie le plus tard possible.

Un choix possible pour l'espace M est $L^2(\Omega)$, et

$$M_{\text{ad}} = \left\{K \in L^2(\Omega), \exists (K_m, K_M)\ 0 < K_m \leq K(x) \leq K_M < \infty (\text{pp})\right\}$$

qui est un sous-ensemble (mais pas un sous-espace vectoriel) convexe de M. L'espace U est naturellement $H^1(\Omega)$, de sorte que pour $K \in M_{\text{ad}}$, l'existence de $T = T_K$ est assurée par le théorème de Lax–Milgram [19, vol. 4].

Pour pouvoir préciser l'opérateur d'observation nous envisagerons plusieurs situations :

i) le cas le plus simple (mais nous l'avons déjà souligné, irréaliste) est celui ou l'état T est connu partout. Même dans ce cas, l'opérateur H ne sera pas l'identité : à cause du choix de U, cela voudrait dire que nous mesurons T *et sa dérivée*. Nous prendrons donc pour D l'espace $L^2(\Omega)$, et pour H l'injection canonique de U dans D. L'observation est alors une fonction $d_{\text{obs}} \in L^2(\Omega)$, c'est-à-dire une fonction définie sur Ω.

ii) Un cas plus réaliste est celui où la température n'est mesurée que sur le bord du domaine. Nous prenons dans ce cas $D = L^2(\Gamma)$, et l'opérateur d'observation est l'opérateur trace sur le bord (qui est bien défini sur U). L'observation est cette fois une fonction définie sur $\partial\Omega$.

iii) Enfin, pour un troisième exemple nous supposons que nous mesurons la température en des points fixés $x_q, q = 1, \ldots, Q$. Une telle observation n'est pas définie dans notre cadre mathématique, puisque les fonctions de U ne sont pas continues. D'ailleurs, une telle mesure n'est jamais réellement ponctuelle en pratique, il s'agit toujours d'une moyenne de la température autour du point considéré. Soit donc ω_q un voisinage du point x_q, nous prendrons comme mesure $\int_{\omega_q} T(x)\,dx$ (ce qui est bien défini si ω_q est borné). L'espace D est alors $\mathbf{R}^Q$, et l'observation est l'application

$$u \mapsto \left(\int_{\omega_q} T(x)\,dx\right)_{q=1,\ldots,Q}.$$

Dans ce cas, l'observation d_{obs} est un vecteur de fonctions, chacune étant définie sur l'un des ouverts Ω_q.

Dans les deux premiers cas, l'observation est *continue*. Elle est distribuée dans le premier cas, frontière dans le second. Enfin, dans le troisième cas, l'observation est *discrète*.

□

Le dernier exemple réunit des caractéristiques ds deux précédents : il s'agit d'une équation aux dérivées partielles d'évolution, dans laquelle le paramètre à identifier est une fonction.

Exemple 7.4.
L'équation de la chaleur est le modèle de base qui régit les phénomènes de diffusion et intervient dans un grand nombre de domaines de la physique. Étant donné un ouvert $\Omega \subset \mathbf{R}^2$ (pour fixer les idées),

dont nous notons Γ la frontière, et un réel $T > 0$, nous considérons le problème :

$$\text{(7.8)} \qquad \begin{cases} \dfrac{\partial u}{\partial t} - \operatorname{div}(a \operatorname{grad} u) = f & \text{dans } \Omega \times]0,T[\\ u(x,t) = 0 & \text{sur } \Gamma_D \times]0,T[\\ a\dfrac{\partial u}{\partial n} = g & \text{sur } \Gamma_N \times]0,T[\\ u(x,0) = u_0(x) & \text{sur } \Omega, \end{cases}$$

où $f \in L^2(0,T;L^2(\Omega))$, $g \in L^2(0,T;L^2(\Gamma_N))$, et $u_0 \in L^2(\Omega)$ sont des fonctions données et supposées connues (Γ_N et Γ_D forment une partition de Γ), et nous cherchons à identifier la fonction a. Avec les réserves maintenant habituelles sur son caractère non-hilbertien, le choix naturel pour l'espace M est $M = L^\infty(\Omega)$, et $M_{\text{ad}} = \left\{ a \in M,\ a(x) \geq a_* > 0 \right\}$.

Nous ferons l'hypothèse que u est mesurée sur la partie Γ_N du bord (et que cette observation est continue en temps), et également que $u(x,T)$ est connue sur tout Ω à l'instant final. Dans ces conditions, les données consistent en deux fonctions : $\hat{d}_N \in L^2(0,T;\Gamma_N)$ et $\hat{d}_T \in L^2(\Omega)$, et l'espace $D = L^2(0,T;\Gamma_N) \times L^2(\Omega)$. □

7.2 Formulation par moindres carrés

Dans les exemples que nous vennons d'examiner, l'application Φ est définie implicitement. Elle est non-linéaire, même si l'équation d'état et l'équation d'observation sont linéaires. Cela rend évidemment plus difficile la résolution du problème inverse. Par contre, dans les exemples 2.12 et 2.9, l'application Φ était définie explicitement, sans recours à un état interne. De plus, ces deux applications sont linéaires. Cela rend évidemment plus difficile la résolution du problème inverse.

Ce que nous avons dit précédemment laisse à penser que l'équation (7.4) peut ne pas avoir de solution, et que même si elle en a, l'application inverse n'est pas nécessairement continue. Nous allons donc introduire une formulation, à priori plus faible, qui a fait la preuve de son utilité. Nous *remplaçons* l'équation (7.4) par le problème de *minimisation* suivant,

$$\text{(7.9)} \qquad \text{minimiser } J(m) = \frac{1}{2} \left\| \Phi(m) - d_{\text{obs}} \right\|_D^2 \text{ pour } m \in M_{\text{ad}}$$

Cette formulation s'appelle une méthode de *moindres carrés*, et J est la *fonction coût*, ou *fonctionnelle d'erreur* (la littérature anglo-saxonne dira : output least squares method, for the cost function, or functional, J). Il est important de comprendre comment ń fonctionne ż cette fonctionnelle. L'observation étant donnée une fois pour toute, pour évaluer la fonctionnelle J en un paramètre p, on commence par résoudre l'équation d'état (7.1), puis l'équation d'observation (7.3), et l'on compare l'observation simulée à celle mesurée.

Remarque 7.1 (L'erreur d'équation). Il existe une autre façon de reformuler un problème inverse comme un problème d'optimisation. Il consiste à remplacer l'état par l'observation (quitte à interpoler cette dernière). Cela conduit à une fonctionnelle

$$\text{(7.10)} \qquad J_{\text{eqn}}(m,d) = \frac{1}{2} \left\| F(p,\tilde{d}) \right\|^2$$

où $\tilde{d}$ est un interpolé de d. Comme elle est quadratique par rapport aux paramètres, cette méthode est très populaire auprès des physiciens et des ingénieurs. Son principal désavantage est de nécessiter d'interpoler l'observation.

Nous allons maintenant revenir sur les exemples du paragraphe 7.1, et proposer pour chacun d'eux une formulation en terme de minimisation de fonctionnelle.

Exemple 7.5 (suite de l'exemple 7.1)**.**
Nous reprenons les différentes observations considérées quand nous avons introduit cet exemple, et proposons une fonctionnelle dans chaque cas.

i) Quand on observe l'état sur tout l'intervalle, $D = L^2(0,1)$, et il est naturel de prendre

$$J_1(b,c) = \frac{1}{2}\int_0^1 |u(x) - d_{\text{obs}}(x)|^2 \, dx. \tag{7.11}$$

ii) Si l'on mesure seulement u à l'extrémité droite de l'intervalle, on prendra

$$J_2(b,c) = \frac{1}{2}|u(1) - d_{\text{obs}}(1)|^2 . \tag{7.12}$$

iii) Enfin, si l'on mesure u en deux points, on prendra :

$$J_3(b,c) = \frac{1}{2}\left(|u(1/2) - d_{\text{obs}}(1/2)|^2 + |u(1) - d_{\text{obs}}(1)|^2\right). \tag{7.13}$$

□

Exemple 7.6 (suite de l'exemple 7.2)**.**
On prend comme fonction coût

$$J(a) = \frac{1}{2}\sum_{q=1}^{Q} \left\| y_a(\tau_q) - d_{\text{obs}}^q \right\|_{\mathbf{R}^d}^2 \tag{7.14}$$

□

Exemple 7.7 (suite de l'exemple 7.3)**.**
Nous allons proposer une fonction coût pour chacunes des trois situations vues à l'exemple 7.3.

i) Dans le cas où l'on mesure u partout (l'opérateur d'observation est l'injection canonique de $H^1(\Omega)$ dans $L^2(\Omega)$), on prendra

$$J_1(a) = \frac{1}{2}\int_\Omega |u_a(x) - d_{\text{obs}}(x)|^2 \, dx. \tag{7.15}$$

ii) Dans le cas où l'on mesure u sur le bord, le choix naturel est :

$$J_2(a) = \frac{1}{2}\int_{\partial\Omega} \left|u_{a|\partial\Omega}(x) - d_{\text{obs}}(x)\right|^2 d\gamma(x). \tag{7.16}$$

iii) Enfin, si l'on mesure u au voisinages de points de Ω, on prendra :

$$J_3(x) = \frac{1}{2}\sum_{q=1}^{Q}\int_{\Omega_q} \left|u_a(x) - d_{\text{obs}}^q(x)\right|^2 dx. \tag{7.17}$$

□

Exemple 7.8 (suite de l'exemple 7.4)**.**
Comme dans les cas précédents, nous agrégeons les erreurs de mesure en une fonction coût unique :

$$J(a) = \frac{1}{2}\int_0^T \int_{\Gamma_N} \left|u - \hat{d}_N\right|^2 \, dx\,dt + \frac{1}{2}\int_\Omega \left|u(x,T) - \hat{d}_T\right|^2 \, dx \tag{7.18}$$

□

Qu'apporte une telle reformulation ? Il est clair qu'elle ne peut pas changer comme par magie un problème mal posé en problème bien posé. Par contre, elle permet de rétablir l'existence. En effet, si il n'existe pas de solution à l'équation (7.4), le problème de minimisation aura forcément une solution (la fonction coût J est positive). Par contre, rien ne garantit que le minimum soit atteint en un point $p \in M_{\text{ad}}$. Il existe des contre-exemples, que l'on pourra trouver dans [6]. Une autre question essentielle est celle de l'unicité. On voit facilement qu'elle est liée à la convexité de la fonctionnelle J. Encore une fois, rien ne garantit cette propriété. La formulation (7.9) présente toutefois des avantages :

- elle donne une façon systématique pour poser les problèmes inverses ;
- dans certains cas, on peut démontrer des propriétés de la fonctionnelle J ;
- Comme nous le verrons au chapitre suivant, cette formulation permet de régulariser le problème, c'est-à-dire de l'approcher par une famille de problèmes bien posés, dont la solution converge vers la solution du problème original ;
- il existe de méthodes numériques robustes et bien étudiées pour résoudre les problèmes d'optimisation ;
- sous des hypothèses raisonnables sur les données, la fonctionnelle J est différentiable, et se prête à l'attaque par une méthode d'optimisation locale de type gradient.

Remarque 7.2 (Sur le choix des normes). Nous avons travaillé dès le départ avec des espaces de Hilbert, et donc des normes hilbertiennes. En pratique, dans les espaces de fonction cela se traduit par des normes L^2 ou de Sobolev. Il n'y a rien de sacré dans ce choix, qui est essentiellement fait par commodité. En un certain sens, il n'est pas naturel, puisque les paramètres varient souvent dans (un sous-espace de) L^∞. Le principal avantage des normes hilbertiennes est qu'elles conduisent à des situations pour lesquelles on sait faire les calculs. On peut également justifier le choix de normes L^2 par des considérations statistiques dans lesquelles nous n'entrerons pas.

D'un autre coté, une norme hilbertienne présente le désavantage de donner plus de poids aux points aberrants, ce qui ne serait pas le cas pour une norme de type L^1. Cela a conduit à des méthodes d'inversions dites ń robustes ż, que nous essayerons de voir brièvement en exercice.

7.2.1 Difficultés des problèmes inverses

La difficulté des problèmes inverses provient d'une combinaison de facteurs.

- Comme nous l'avons mentionné au paragraphe 7.2, la fonction coût est en général *non convexe*. Cela conduit à l'existence de minima locaux, et la méthode d'optimisation peut converger vers n'importe lequel de ces minima.
- Le problème inverse peut-être *sous-determiné*, du fait d'un manque de données (qui est intrinsèque au problème). Cela conduit à l'existence de plusieurs solutions, autrement dit de plusieurs paramètres produisant les mêmes observations.
- Le manque de continuité produit une *instabilité*. Même si l'on peut (en théorie) résoudre le problème pour des observations exactes, cela ne veut pas dire que l'on pourra le résoudre pour des données bruitées, même si le niveau de bruit est faible.

– Une difficulté de nature différente est liée au coût de la résolution, en supposant que l'on puisse s'affranchir des obstacles précédents. En effet, la simple *évaluation* de la fonction coût demande la résolution de l'équation d'état, c'est-à-dire en général d'une (ou de plusieurs) équation aux dérivées partielles.

7.2.2 Optimisation, paramétrisation, discrétisation

Le problème d'optimisation (7.9) ainsi que l'équation d'état (7.1) sont en général posés ń en dimension infinie ż, c'est-à-dire que les espaces de Hilbert M, U et D sont de dimension infinie. Ce sont en général des espaces de fonctions, ainsi que nous l'avons vu dans les exemples du chapitre 1. Avant de commencer une résolution sur ordinateur il est bien entendu nécessaire de se ramener à un problème en dimension finie. Comme on ne dispose habituellement que d'une nombre fini d'observations, l'espace D est souvent de dimension finie dès le départ.

Il n'en est pas de même pour l'espace des paramètres. Le processus qui consiste à remplacer M par un espace de dimension finie est la *paramétrisation*. Des exemples couramment utilisés sont des fonctions constantes par morceaux (une valeur par cellule, dans le cas d'une grille), des approximations polynômiales par morceaux (fonctions splines), mais d'autres choix sont possibles. Il est souhaitable de conserver le paramètre sous forme ń fonctionnelle ż le plus longtemps possible, de façon à pouvoir changer facilement de paramétrisation. Bien entendu, cela implique que l'algorithme d'optimisation devra fonctionner en dimension infinie. Il est possible de formuler les méthodes de quasi-Newton (dont nous rappellerons le principe au paragraphe 7.3.3), et d'étudier leur convergence, dans le cadre des espaces de Hilbert.

Il est par contre nécessaire de discrétiser l'équation d'état. On remplace donc U par un espace de dimension finie. Ce choix a évidemment une influence capitale, car en changer implique d'avoir à recommencer tout le processus d'analyse. On obtient donc un problème d'optimisation pour lequel l'inconnue est toujours de dimension infinie, mais l'équation d'état est posée en dimension finie. Ceci permet d'obtenir le gradient exact de la fonction coût effectivement utilisée par le programme. Une alternative serait de discrétiser l'équation d'état au dernier moment, et de discrétiser également le gradient continu. L'expérience prouve que cette seconde manière de procéder dégrade la convergence des méthodes d'optimisation, et nous procéderons en calculant le gradient exacte de la fonctionnelle approchée.

Une fois ces choix faits, il faut choisir une discrétisation de m adaptée à la simulation, qui pourra être différente de la paramétrisation évoquée plus haut. Cette discrétisation sera d'une finesse compatible avec celle utilisée pour l'équation d'état, mais ne sera pas nécessairement adaptée à l'optimisation. On calculera tout de même le gradient de la fonction coût par rapport à ces paramètres. Une autre paramétrisation sera simplement une étape supplémentaire bien distincte de la simulation, dont le gradient s'obtient par application de la règle de dérivation composée. Cette méthode permet de séparer les différentes composantes du logiciel : simulation, calcul du gradient, paramétrisation et optimisation sont ainsi des modules séparés avec des interfaces bien spécifiées. Nous reviendrons sur ce point au paragraphe 8.3.

Avant de détailler le calcul du gradient par la méthode de l'état adjoint, nous allons rappeler le principe de quelques méthodes d'optimisation. Pour rester en conformité avec les principes exposés plus haut, nous décrirons les algorithmes d'abord dans le cas de la dimension infinie. Comme cela requiert tout de même un niveau d'abstraction élevé, nous donnerons également une version en dimension finie, qui est bien entendu celle qui sera finalement mise en oeuvre numériquement.

7.3 Rappels d'optimisation

Dans tout le chapitre nous ferons l'hypothèse que J est une application de classe C^2 de M dans $\mathbf{R}$, ce qui sera suffisant pour appliquer les résultats dont nous aurons besoin.

Sans vouloir faire un cours d'optimisation (il existe de très bons livres pour cela, comme par exemple [11, 22, 42, 54] ou encore [26], disponible sur l'Internet), nous rappellerons brièvement les propriétés de la méthode de BFGS, la plus importante des méthodes de quasi-Newton, puis nous présenterons la méthode de Gauss-Newton, souvent utilisée pour résoudre des problèmes de moindres carrés non-linéaires, ainsi que sa variante connue sous le nom de méthode de Levenberg–Marquardt.

7.3.1 Algorithmes locaux et globaux

Pour résoudre le problème (7.9) on peut distinguer deux grandes classes d'algorithmes. Ceux qui sont basés sur la condition nécessaire de la proposition 7.2 ci-dessous sont qualifiés de ń locaux ż. En pratique cela veut dire que ces méthodes exploitent l'information fournie pas le gradient de la fonctionnelle à minimiser et cherchent à résoudre l'équation fournie par la *condition d'optimalité du premier ordre* (voir la proposition 7.2).

Le principal défaut de ce type d'algorithme est qu'il ne peut converger que vers un minimum *local*, ou même un point critique (point où le gradient s'annule). Il faut utiliser les dérivées secondes pour écarter les points critiques qui ne sont pas des minima. Avec une information uniquement locale du premier ordre, il n'est pas possible d'éviter cet inconvénient. Il est contrebalancé par le fait que ces méthodes sont bien comprises du point de vue mathématique, et qu'il en existe des implémentations de qualité (une liste de logiciels d'optimisation existant se trouve dans [50]).

Les algorithmes locaux s'opposent aux méthodes dites ń globales ż qui explorent tout l'espace des paramètres de façon à converger vers *le* minimum global (si il existe). Ces méthodes ont parfois un aspect plus heuristiques, et sont fondamentalement plus coûteuses et plus lentes. Elles présentent l'avantage d'éviter le calcul du gradient de la fonction coût, qui comme nous le verrons est un des points délicats dans la mise en œuvre des méthodes locales. Parmi les méthodes globales citons la méthode du ń recuit simulé ż [64] qui évite les minima locaux en laissant croître la fonction coût avec une certaine probabilité. Citons également les méthodes de recherche directe [62] qui construisent une suite de simplexes se contractant vers le minimum global. Dans les deux cas le nombre d'évaluation de la fonction coût est très important. Les méthodes globales sont en fait à réserver au cas où on l'on cherche à identifier un petit nombre de paramètres, et où le gradient de la fonctionnelle est difficile à calculer.

Dans le reste de ce cours, nous ne considérerons plus que les algorithmes locaux, et nous chercherons essentiellement à montrer comment calculer économiquement le gradient de la fonction coût.

7.3.2 Gradients, hessiens et conditions d'optimalité

À titre de référence, nous avons rassemblé dans cette section quelques définitions et résultats bien connus. Les démonstrations se trouvent dans les références données plus haut.

Définition 7.1. – La *différentielle* de J au point $a \in M$ est la forme linéaire, notée $J'(a)$, caractérisée par la relation :

$$J(a+\delta a) = J(a) + J'(a)\delta a + o(\|\delta a\|), \quad \forall \delta a \in M \tag{7.19}$$

– Le *gradient* de J au point $a \in M$, noté $\nabla J(a)$, est défini par l'égalité :

$$(\nabla J(a), h) = J'(a)h, \quad \forall h \in M. \tag{7.20}$$

– Le *hessien* de J au point $a \in M$, noté $\nabla^2 J(a)$ est l'opérateur auto-adjoint de M défini par l'égalité :

$$(\nabla^2 J(a)h, k) = J''(a)(h,k), \quad \forall (h,k) \in M^2. \tag{7.21}$$

Signalons au passage que si la différentielle J' est une notion topologique (elle est définie par une limite), le gradient est une notion géometrique, puisqu'il est lié au choix d'un produit scalaire. De la même façon, l'identification de la forme bilinéaire $J''(a)$ avec l'opérateur hessien dépend du produit scalaire.

Dans le cas de la dimension finie, avec le produit scalaire usuel, la différentielle est un vecteur ligne, et le gradient est le vecteur colonne associé :

$$J'(a) = \nabla J(a)^t = \left(\frac{\partial J}{\partial a_1}, \ldots, \frac{\partial J}{\partial a_p} \right). \tag{7.22}$$

De même, le hessien s'identifie à une matrice (symétrique), dont les éléments sont

$$\nabla^2 J(a)_{ij} = \frac{\partial^2 J}{\partial a_i \partial a_j}. \tag{7.23}$$

La notion de gradient (et sa distinction avec la différentielle) prend toute son importance en dimension infinie, quand il n'y a plus d'identification canonique entre un espace de Hilbert et son dual. Même en dimension finie, il est utile de conserver les deux notions, et de réécrire (7.19) sous la forme :

$$J(a+\delta a) = J(a) + \nabla J(a)^t \delta a + o(\|\delta a\|), \quad \forall \delta a \in M. \tag{7.24}$$

Dans la suite de ce cours, nous serons principalement concernés par des fonctions coût du type ń moindres carrés non-linéaires ż, c'est-à-dire

$$J(a) = \frac{1}{2} \|\Phi(a) - d\|_D^2, \tag{7.25}$$

où $\Phi : M \mapsto D$ est une application non-linéaire. Pour ce type de fonctions coût, le gradient et le hessien prennent une forme particulière.

Proposition 7.1. *Le gradient et le hessien de la fonctionnelle J sont donnés par les formules suivantes :*

$$\nabla J(a) = \Phi'(a)^* (\Phi(a) - d) \; ; \tag{7.26}$$

et

$$\nabla^2 J(a)\delta a = \Phi'(a)^* \Phi'(a)\delta a + (\Phi''(a)\delta a)^* (\Phi(a) - d), \text{ pour } \delta a \in M. \tag{7.27}$$

Preuve. – Si nous notons $N(u) = 1/2 \|u\|^2$ pour $u \in D$, alors $J(a) = N(\Phi(a) - d)$, et la différentiation d'une fonction composée donne :

$$J'(a)\delta a = N'(\Phi(a) - d))\Phi'(a)\delta a$$

et il est facile de voir que $N'(u)\delta u = (u, \delta u)$ pour $\delta u \in D$. Le résultat découle alors de (7.20).

– On dérive l'équation ci-dessus pour obtenir l'expression du hessien.

□

Rappelons maintenant les liens entre gradients, hessiens et les conditions d'optimalité, tout d'abord dans le cas sans contrainte.

Proposition 7.2 (Condition nécessaire du premier ordre)**.** *Soit â un point où J atteint son minimum. On a :*

$$\nabla J(a) = 0. \tag{7.28}$$

Cette condition n'est évidemment pas suffisante, sauf si J est convexe. Un point critique de J peut être un minimum, un maximum ou bien un point-selle. Dans le sens inverse, on a :

Proposition 7.3 (Condition suffisante du second ordre)**.** *Supposons que $\nabla J(\hat{a}) = 0$ et que $\nabla^2 J(\hat{a})$ soit défini positif. Alors J atteint un minimum local strict en â.*

Une notion importante tant dans la définition des algorithmes que leur analyse est la suivante :

Définition 7.2. Un vecteur $d \in M$ est une *direction de descente* pour J au point a si

$$(\nabla J(a), d) < 0 \tag{7.29}$$

Cette définition assure que la fonction décroît, au moins localement, le long d'une direction de descente. À titre d'exemple, signalons que la direction opposée au gradient est évidemment une direction de descente. C'est sur cette remarque qu'est basé l'algorithme de la plus grande pente, qui malheureusement converge trop lentement pour être utile en pratique.

Comme cela nous sera utile pour présenter la méthode de l'état adjoint, nous donnons également un résultat de base dans le cas de la minimisation avec contraintes d'égalité. On cherche toujours à minimiser une fonction J définie sur une espace de Hilbert M, mais on suppose de plus que l'inconnue a doit satisfaire une contrainte de la forme $G(a) = 0$. Le résultat suivant est connu sous le nom de *théorème des multiplicateurs de Lagrange*.

Proposition 7.4. *Soit â une solution du problème de minimisation vérifiant $G(\hat{a}) = 0$, où $G : M \mapsto P$, (P est un autre espace de Hilbert) est une application que nous supposerons différentiable. Sous l'hypothèse :*

$$G'(\hat{a}) \text{ surjective}$$

il existe $p \in P$, appelé multiplicateur *tel que :*

$$\nabla J(\hat{a}) + G'(\hat{a})^* p = 0 \tag{7.30}$$

Dans le cas de la dimension finie, notons N_a la dimension de M et N_p celle de P. La contrainte G devient un vecteur de $\mathbf{R}^{N_p}$ $G = (G_1, \ldots, G_{N_p})$, il en est de même pour $p = (p_1, \ldots, p_{N_p})$, et la condition (7.30) se réécrit :

$$\nabla J(\hat{a}) + \sum_{j=1}^{N_p} p_j \nabla G_j(\hat{a}) = 0. \tag{7.31}$$

Il est usuel d'introduire le lagrangien, défini par

$$\mathcal{L}(a, p) = J(a) + \sum_{j=1}^{N_p} p_j G_j(a), \tag{7.32}$$

et les équations (7.30) ou (7.31) expriment que $\nabla_m \mathcal{L} = 0$.

7.3.3 Méthodes de quasi-Newton

Une méthode très efficace pour résoudre (7.9), ou plutôt (7.28), est la méthode de Newton. Elle consiste à construire la suite a^n définie par :

$$H^{n+1}(a^{n+1} - a^n) = \nabla J(a^n) \tag{7.33}$$

où H^n désigne le hessien de J au point a^n. Pour que la méthode soit bien définie, il faut évidemment que H^{n+1} soit inversible (ou mieux défini positif, ce qui permet de distinguer les minima des points critiques).

Il est commode de définir d^n comme la solution du système linéaire

$$H^{n+1}d^n = -\nabla J(a^n) \tag{7.34}$$

(on démontre que si H^{n+1} est défini positif, alors d^n est une direction de descente pour J), de sorte que l'équation (7.33) se réécrit :

$$a^{n+1} = a^n + d^n. \tag{7.35}$$

Il est bien connu (voir les références d'optimisation ci-dessus) que la méthode de Newton est localement convergente, avec convergence quadratique, pourvu que le point initial soit choisi assez proche de la solution (inconnue), et que le hessien en la solution ne soit pas singulier. Elle possède toutefois des défauts assez graves :

- la convergence n'est pas globale ;
- l'algorithme n'est pas défini aux points où le hessien est singulier ;
- l'algorithme ne génère pas nécessairement des directions de descente ;
- il faut calculer le hessien à chaque itération, et résoudre le système linéaire (7.34).

Tous ces défauts sont corrigés par les méthodes de quasi-Newton. Nous nous contenterons d'exposer brièvement l'algorithme de BFGS avec recherche linéaire.

La première idée est de garder la direction de descente donnée par (7.34), mais de ne pas ń avancer ż de toute la longueur du pas. On remplace (7.35) par :

$$a^{n+1} = a^n + \alpha^n d^n \tag{7.36}$$

où α^n est déterminé par une *recherche linéaire*, c'est-à-dire que l'on s'assure que entre a^n et a^{n+1} J a décru ń suffisamment ż. On s'est rendu compte qu'il n'est ni nécessaire ni même utile de déterminer la paramètre α^n ń optimal ż (celui qui réalise le minimum en α de $J(a^n + \alpha d^n)$. Il suffit, pour assurer la convergence de l'algorithme, de vérifier les *conditions de Wolfe* :

$$J(a^n + \alpha^n d^n) - J(a^n) \leq \omega_1 \alpha^n (\nabla J(x^n), d^n), \tag{7.37}$$

$$(\nabla J(a^n + \alpha^n d^n), d^n) \geq \omega_2 (\nabla J(x^n), d^n), \tag{7.38}$$

où les constantes ω_1 et ω_2 vérifient $0 < \omega_1 < \omega_2 < 1$.

La première inégalité assure que J aura diminué d'une fraction significative entre l'ancien point et le nouveau. La seconde empêche simplement le pas de devenir trop petit. En effet, elle n'est pas vérifiée pour $\alpha^n = 0$ si d^n est bien une direction de descente, et donc pas non plus pour des valeurs de α^n ń petites ż (par continuité). On démontre par ailleurs que si d^n est une direction de descente, et si la fonction $\alpha \to J(a^n + \alpha d^n)$ est bornée inférieurement, il existe une valeur $\alpha^n > 0$ satisfaisant les conditions de Wolfe. Un algorithme pour trouver une valeur de α satisfaisante est décrit dans [11].

La seconde modification à apporter à la méthode de Newton est de remplacer le choix de la direction d^n de façon à assurer que celle-ci est toujours une direction de descente. La méthode de BFGS (du nom de ses auteurs Broyden, Fletcher, Goldfarb et Shanno) résout d'un seul coup les trois dernières difficultés signalées plus haut.

Cette méthode fonctionne en mettant à jour de façon itérative une approximation du hessien. Il est standard de définir les quantités

$$s^n = a^{n+1} - a^n \text{ et } y^n = \nabla J(a^{n+1}) - \nabla J(a^n). \tag{7.39}$$

Avec ces notations, la mise à jour de BFGS s'écrit :

$$H^{n+1}\delta a = H^n \delta a + \frac{(y^n, \delta a)}{(y^n, s^n)} y^n - \frac{(H^n s^n, \delta a)}{(H^n s^n, s^n)} H^n s^n. \tag{7.40}$$

En dimension finie, si le produit scalaire est le produit euclidien usuel, cette formule prend la forme plus habituelle :

$$H^{n+1} = H^n + \frac{y^n (y^n)^t}{(y^n)^t s^n} - \frac{(H^n s^n)(H^n s^n)^t}{(H^n s^n)^t s^n}. \tag{7.41}$$

Il s'agit d'une mise à jour par ajout d'une matrice de rang 2. Ses principales propriétés sont résumées dans la

Proposition 7.5. *Si H^n est symétrique et définie positive et si $(y^n, s^n) > 0$, alors H^{n+1} est symétrique et définie positive et $d^{n+1} = -(H^{n+1})^{-1}\nabla J(a^{n+1})$ est une direction de descente.*

On peut démontrer (voir [26]) que si l'on utilise la recherche linéaire de Wolfe, la condition précédente est vérifiée.

En combinant les deux idées précédentes, nous obtenons (moyennant certaines hypothèses) une méthode globalement convergente. On peut énoncer le résultat suivant, pris dans [26] :

Théorème 7.1. *Supposons J convexe dans un voisinage de l'ensemble $\{a \in M,\ J(a) \leq J(a^0)\}$, et supposons de plus que la suite $J(a^n)$ soit bornée inférieurement. Alors la suite générée par l'algorithme de BFGS avec la règle de Wolfe pour la recherche linéaire vérifie :*

$$\liminf_{n\to\infty} \|\nabla J(a^n)\| = 0$$

Ce théorème veut dire que, moyennant des hypothèses techniques, tout point d'accumulation de la suite calculée est un point critique de la fonctionnelle J.

En ce qui concerne l'implémentation, il reste une difficulté à résoudre : comment exploiter la formule (7.41) sans stocker de matrice pleine. En fait, il est possible de donner une formule analogue à (7.41), mais sur $(H^n)^{-1}$, ce qui permet de mettre directement d^n à jour, sans inverser (explicitement) de système linéaire. Ceci conduit à un coût par itération faible (en dehors du calcul de $J(a)$ et de son gradient.

7.3.4 Moindres carrés non-linéaires et méthode de Gauss-Newton

Bien que la méthode vue au paragraphe précédent donne toute satisfaction, elle s'applique à des fonctions générales, et il est possible d'essayer d'exploiter la structure particulière du problème (7.9). En effet, comme nous l'avons vu au paragraphe 7.3.2, les fonctions coût qui nous intéressent ici ont

la forme particulière (7.25), et nous avons vu à la proposition 7.1 comment calculer le gradient et le Hessien de cette fonctionnelle, en supposant que le jacobien et la dérivée seconde de l'application Φ soient connus.

Notons que si le problème d'optimisation est consistant, c'est-à-dire que la solution $\hat{a}$ vérifie $\Phi(\hat{a}) = d$, ou si Φ est linéaire, alors le second terme du hessien dans (7.27) disparaît, et nous avons :

$$\nabla^2 J(\hat{a}) = \Phi'(\hat{a})^* \Phi'(\hat{a}).$$

Cette remarque laisse penser que, au moins au voisinage de la solution, il est possible de négliger le second terme du hessien. La méthode obtenue en remplaçant, dans la méthode de Newton, le hessien par l'expression

$$H_{\mathrm{GN}}(a) = \Phi'(a)^* \Phi'(a) \tag{7.42}$$

s'appelle la *méthode de Gauss-Newton*. Une itération de la méthode conduit à la résolution du système linéaire

$$\Phi'(a)^* \Phi'(a) d = -\nabla J(a) = -\Phi'(a)^* (\Phi(a) - d), \tag{7.43}$$

c'est-à-dire un problème de moindres carrés *linéaires* pour l'opérateur $\Phi'(a)$, problème que nous avons abondamment étudié au chapitre 6.

Cette simplification présente plusieurs avantages :

- Tout d'abord, elle évite d'avoir à calculer la dérivée seconde de Φ, ce qui peut-être difficile en pratique ;
- Ensuite, on voit que le coût de calcul du Hessien approché $H_{\mathrm{GN}}(a)$ est essentiellement le même que celui du calcul du gradient. Le coût d'une itération de Gauss-Newton est donc du même ordre que celui d'une itération de descente.
- Enfin, le Hessien approché est semi-défini positif, ce qui n'est pas nécessairement le cas du Hessien complet.

La méthode de Gauss-Newton souffre des mêmes défauts que la méthode de Newton, en particulier le manque de convergence globale, et le fait que la méthode n'est pas définie si $\Phi'(a)^* \Phi'(a)$ n'est pas inversible (défini positif), ce qui est possible. Une idée naturelle est alors de *régulariser* les problèmes linéarisés, selon les techniques vues au chapitre 6, en particulier la régularisation de Tikhonov. La méthode correspondante porte le nom de Levenberg–Marquardt. Cette méthode consiste donc à calculer $a^{n+1} = a^n + d^n$ où d^n résout le système linéaire

$$(\varepsilon^2 I + \Phi'(a)^* \Phi'(a)) d^n = -\Phi'(a)^* (\Phi(a) - d) \tag{7.44}$$

En fait, comme nous l'avons vu au chapitre 6, il n'est pas nécessaire de calculer l'opérateur $\varepsilon^2 I + \Phi'(a)^* \Phi'(a)$, on peut résoudre le problème de moindres carrés :

$$\min \frac{1}{2} \left\| \begin{bmatrix} \Phi'(a^n) \\ \varepsilon I \end{bmatrix} d^n + \begin{bmatrix} \Phi(a^n) - d \\ 0 \end{bmatrix} \right\|^2 .$$

Cette méthode peut bien entendu être couplée avec la règle de Wolfe pour donner une méthode globalement convergente (sous des hypothèses raisonnables). En fait, l'analyse montre (et l'expérience confirme) que la méthode est efficace pour les problèmes où le résidu final $\Phi(\hat{a}) - d$ est petit, et peut ne pas converger si le résidu final est trop grand.

Le choix du paramètre ε n'est pas plus facile qu'au chapitre 6. Une possibilité s'appuie sur les méthodes de *région de confiance* pour ajuster le paramètre en fonction de l'adéquation de la fonctionnelle J à un modèle quadratique (voir [42] pour plus de détails).

Nous pouvons constater que le gradient est un ingrédient essentiel de l'algorithme BFGS. Pour Gauss-Newton, il semble que le jacobien complet de Φ soit nécessaire. Au chapitre suivant, nous développerons une méthode de calcul efficace pour les problèmes auxquels nous nous intéressons dans ce cours. Nous verrons qu'il est possible de calculer le gradient de J sans calculer le jacobien de Φ (ce qui peut sembler surprenant). Ce résultat favorise la méthode BFGS (que nous recommandons en général), mais la méthode que nous exposerons permet de calculer une ligne du jacobien de Φ, ce qui peut permettre d'utiliser la méthode de Gauss-Newton.

Chapitre 8

Calcul du gradient – La méthode de l'état adjoint

Dans ce chapitre, nous allons exposer une méthode de calcul du gradient d'une fonction du type

$$J(a) = \frac{1}{2} \|\Phi(a) - d_{\text{obs}}\|_D^2 \tag{8.1}$$

où l'application non-linéaire Φ est définie par la résolution d'une équation d'état :

$$F(a,u) = 0, \tag{8.2}$$

dont la solution est u_a, puis en extrayant de u_a une observation

$$\Phi(a) = Hu_a. \tag{8.3}$$

La difficulté est clairement dans le calcul de la dérivée de l'application (implicite) $a \to u_a$. Nous allons commencer par passer en revue plusieurs méthodes pour calculer ce gradient. Nous verrons plus en détail la méthode de l'état adjoint qui permet ce calcul pour un coût indépendant du nombre de paramètres. Après avoir exposé la méthode dans un cadre général, abstrait, nous donnerons plusieurs exemples explicites pour voir comment mener à bien ce calcul dans une situation concrète.

8.1 Méthodes de calcul du gradient

Nous présentons dans ce paragraphe plusieurs façons de calculer le gradient de (8.1).

Il s'agit tout d'abord de la méthode des différences finies (qui n'est pas recommandée, mais qui présente tout de même une utilité pour fournir des valeurs de références), de la méthode des sensibilités, et enfin de la méthode de l'état adjoint. Laquelle des deux dernières est la plus adaptée à une situation dépend du nombre de paramètres par rapport au nombres d'observations.

8.1.1 Les différences finies

Cette méthode est, en apparence, très simple à mettre en œuvre, ce qui peut expliquer sa popularité, mais elle n'est pas recommandable, puisque non seulement son coût est proportionnel au nombre de paramètres à identifier, mais elle donne un résultat approché, avec une précision qu'il est difficile d'évaluer. Dans certaines circonstances, elle peut servir à valider un calcul de gradient par l'une des autres méthodes (sensibilités ou état adjoint).

On remplace le calcul d'une dérivée partielle par le quotient aux différences :

$$\frac{\partial J}{\partial a} \approx \frac{J(a+h_j)-J(a)}{h_j}. \tag{8.4}$$

On constate immédiatement que le nombre d'évaluations de J est égal au nombre de paramètres à identifier (plus un, mais ce calcul est toujours nécessaire). Par rapport à la méthode précédente, le coût est équivalent, mais l'on n'obtient que le gradient, pas le jacobien. De plus, le résultat n'est pas exact. D'ailleurs, en précision finie, l'erreur commise se compose de deux termes : l'erreur d'approximation et l'erreur d'arrondi. Nous allons analyser précisément comment se combinent ces deux effets. Pour simplifier, nous nous plaçons dans le cas d'une fonction f d'une variable réelle x.

Dans ce cas, le développement de Taylor de f à l'ordre un donne :

$$f'(x) = \frac{f(x+h)-f(x)}{h} + h/2f''(x) + O(h^2). \tag{8.5}$$

Supposons par ailleurs que f soit calculée avec une précision relative ε_f. Ce peut être simplement l'erreur d'arrondi, auquel cas ε_f est la précision de l'arithmétique de l'ordinateur (de l'ordre de 10^{-16} en double précision), ou bien une valeur bien plus grande si f est le résultat d'un calcul complexe. Dans ce cas, ce qui est effectivement calculé est $\tilde{f} = f(1+\varepsilon_f)$, et la différence entre le quotient calculé et la vraie dérivée vaut, en négligeant l'erreur d'arrondi due à la division) :

$$f'(x) - \frac{\tilde{f}(x+h)-\tilde{f}(x)}{h} = h/2f''(x) + 2\varepsilon_f/h\, f(x) + O(h^2). \tag{8.6}$$

La somme des deux premiers termes est minimisée par le choix

$$h = 2\sqrt{\varepsilon_f \frac{f(x)}{f''(x)}} \tag{8.7}$$

et l'erreur totale est alors proportionnelle à $\sqrt{\varepsilon_f}$. En double précision, cela veut dire que la dérivée aura environ 8 chiffres exacts. Notons que la valeur de la fonction à dériver, et de sa dérivée seconde, influent sur le choix du pas optimal, comme le montre l'équation (8.7). Plus $f(x)$ sera grand, plus on pourra choisir ε grand. De même, plus $f''(x)$ sera grand, c'est-à-dire plus $f'(x)$ varie rapidement, plus le choix de ε devra être petit. Ainsi, le choix effectif du pas reste délicat, même si $\sqrt{\varepsilon_f}$ est une première estimation raisonnable.

Le livre [22] contient un algorithme de choix du pas, dans le cas de plusieurs variables, qui prend en comte les divers facteurs d'échelle qui peuvent intervenir. Toutefois, comme nous déconseillons de calculer le gradient par cette méthode, nous ne donnerons pas de détails.

Il est tout de même des cas où le calcul du gradient par différences finies peut-être utile : en particulier pour vérifier un calcul par état adjoint. Dans ce cas, on choisit un paramètre ń au hasard ż, et l'on calcule le gradient en ce point par différences finies, pour plusieurs valeurs du pas (disons de 10^{-2} à 10^{-15}). Si les deux calculs sont justes, l'erreur doit passer par un minimum au voisinage de 10^{-7}.

8.1.2 Les fonctions de sensibilité

Il s'agit de la méthode la plus naturelle pour calculer le gradient de J. Elle consiste à dériver l'équation d'état explicitement par rapport au paramètre m, puis à utiliser la règle dérivation d'une fonction composée. Insistons sur le fait que cette méthode donne un résultat *exact*.

Rappelons que nous avons vu à la proposition 7.1 que

$$\nabla J(a) = \Phi'(a)^*(\Phi(a) - d_{\text{obs}}). \tag{8.8}$$

Une première idée est donc de calculer le Jacobien de Φ. Comme Φ est définie de façon implicite par la résolution de l'équation d'état (7.1) et l'équation d'observation (7.3), nous devons faire appel au théorème des fonctions implicites (plus exactement à son corollaire qui permet le calcul de la différentielle de l'application implicite une fois que l'on sait que celle-ci est différentiable). Ce résultat dit que l'on obtient le jacobien de Φ en dérivant l'équation d'état :

$$\partial_u F(a,u)\delta u + \partial_a F(a,u)\delta a = 0, \tag{8.9}$$

puis en résolvant l'équation (linéaire) précédente, et en composant avec (la dérivée de) l'observation (que nous avons supposé linéaire) :

$$\Phi'(a) = -H\left(\partial_u F(a,u)\right)^{-1}\partial_a F(a,u) \tag{8.10}$$

En regroupant les équations (8.8) et (8.10), on obtient finalement le gradient de J :

$$\nabla J(a) = \left(\Phi'(a)\right)^*\big(Hu(a) - d_{\text{obs}}\big). \tag{8.11}$$

Le principal désavantage de cette méthode réside dans le fait que le calcul de δu demande la résolution d'une équation d'état (linéarisée) pour chaque valeur de δa. Après passage en dimension finie, cela veut dire que le calcul de chaque dérivée partielle $\partial J/\partial a_j$ demande la résolution d'une équation comme (8.9). Le coût du calcul du gradient est donc *proportionnel* au nombre de paramètres. Or, dans une grande partie des situations d'intérêt, ce nombre peut être très grand : plusieurs centaines, voire plusieurs milliers. Nous verrons plus bas, c'est le principal avantage de la méthode de l'état adjoint, qu'il est possible de réaliser ce calcul à un coût proportionnel à celui d'une seule équation linéarisée, et en particulier, *indépendant* du nombre de paramètres.

En contrepartie, cette méthode fournit plus que le gradient, puisque nous avons vu qu'elle calcule le jacobien de Φ. Un fois ce jacobien disponible, il est possible de l'exploiter en calculant, par exemple, ses valeurs singulières. De plus, la méthode de Gauss-Newton nécessite la connaissance de ce jacobien. Si le nombre de paramètres n'est pas trop élevé, la méthode de Gauss-Newton en calculant le gradient comme en (8.11) peut être plus économique qu'une méthode de quasi-Newton avec calcul du gradient par l'état adjoint.

8.1.3 La méthode de l'état adjoint

Nous avons déjà noté que la méthode des fonctions de sensibilité fournissait plus que le gradient de J. Si nous n'avons besoin que du gradient, nous pouvons réarranger le calcul menant à (8.11) pour éviter le calcul du jacobien complet. En reportant (8.10) dans (8.11), et en transposant le produit, nous obtenons (la notation $-*$ représente l'inverse de l'adjoint) :

$$\nabla J(a) = -\Big(\partial_a F(a,u)^*(\partial_u F(a,u)^*)^{-1}H^*\Big)\big(Hu(a) - d_{\text{obs}}\big). \tag{8.12}$$

La remarque d'apparence triviale qui va nous permettre de simplifier le calcul est qu'il est possible de parenthéser différemment cette expression :

$$\nabla J(a) = -\big(\partial_a F(a,u)\big)^*\Big((\partial_u F(a,u)^*)^{-1}\Big(H^*\big(Hu(a) - \hat{d}\big)\Big)\Big). \tag{8.13}$$

Il est commode de donner un nom à la quantité à l'intérieur de la seconde parenthèse, et d'introduire le vecteur p solution de

$$\partial_u F(a,u)^* p = -H^* \Big(Hu(a) - d_{\text{obs}} \Big), \tag{8.14}$$

On appelle cette équation *l'équation adjointe*, et p est *l'état adjoint*. Une fois cette équation résolue, le gradient se calcule alors par :

$$\nabla J(a) = \partial_a F(a,u)^* p. \tag{8.15}$$

Résumons le résultat de ces calculs dans le

Théorème 8.1. *Si p est la solution de* l'équation adjointe (8.14)*, le gradient de J au point a est donné par* (8.15)*, où $u = u(a)$ est la solution de l'équation d'état* (7.1) *correspondant à a.*

Remarque 8.1. Le théorème 8.1 est d'une grande importance. Il fournit le cadre général sur lequel s'appuie la méthode de l'état adjoint. Toutefois, il est difficile à appliquer tel quel en pratique, et c'est pourquoi nous proposerons une méthode plus simple pour aboutir au même résultat. En effet, il peut être difficile, dans un cas concret d'identifier les différents adjoints concernés, voire même l'opérateur F lui-même. Ce sera en particulier vrai pour les problèmes d'évolution.

Remarque 8.2. Nous voyons, d'après l'équation 8.15 que nous obtenons le gradient complet de J en résolvant la seule équation adjointe (8.14). Ceci implique que la méthode de l'état adjoint permet de calculer le gradient à un coût proportionnel à celui du calcul de la fonction elle-même. Dans la plupart des cas, ce coût est un petit (entre 3 et 5) multiple de celui de la fonction (ce résultat est démontré dans [30]). Ceci justifie, à posteriori, la méthode de l'état adjoint, et l'abandon de la méthode ń naturelle ż vue au paragraphe 8.1.2.

Remarque 8.3. L'opérateur qui intervient dans l'équation adjointe est donc *l'adjoint* de $H\partial_u F(a,u)^{-1}$, qui n'est autre que la dérivée de l'application Φ qui associe le paramètre aux données. C'est cette propriété qui à donné son nom à la méthode. La différence avec le calcul par les fonctions de sensibilité est qu'ici cette transposition est implicite, ń cachée ż dans la définition de l'équation adjointe.

Remarque 8.4. L'équation adjointe (8.14) est toujours linéaire, et ce même si l'équation d'état était non-linéaire. Les sources de l'équation adjointe sont formées par le résidu de l'équation d'observation.

En pratique, nous verrons qu'il est parfois difficile de réaliser concrètement les transpositions indiquées dans le théorème 8.1. Nous allons introduire une autre facon de mener à bien ce calcul qui se révèle d'utilisation plus simple en pratique.

8.1.4 Calcul de l'état adjoint par le lagrangien

Le paragraphe précédent a montré comment calculer le gradient de notre fonction coût en résolvant seulement deux équations : l'équation d'état, suivie de l'équation adjointe (les opérations de la formule (8.15) sont typiquement très simples). Comme nous l'avons déjà signalé, la mise en oeuvre de ce résultat n'est pas aussi simple. Nous allons donner une technique qui conduit au même résultat, mais se révèle plus souple d'emploi, comme nous le verrons dans les exemples qui suivent (voir paragraphe 8.2).

La méthode repose sur ce qui peut être vu comme une astuce de calcul. On commence par prétendre que les variables a et u varient indépendamment, et l'on considère l'équation d'état comme une

contrainte. Dans ces conditions, comme nous l'avons vu au paragraphe 7.3.2, il est naturel d'introduire un lagrangien. Dans notre cas, celui-ci s'écrit (au moins formellement, puisqu'ici nous n'avons pas un nombre fini de contraintes) :

$$\mathcal{L}(a, u, p) = \frac{1}{2} \|Hu - d_{\text{obs}}\|^2 + (p, F(a, u)) \tag{8.16}$$

La remarque (encore une fois en apparence triviale) fondamentale est que, si u vérifie l'équation d'état correspondant au paramètre a, on a l'identité :

$$\mathcal{L}(a, u(a), p) = J(a), \ \forall p \in Z.$$

En dérivant cette relation, il vient :

$$J'(a)\delta a = \partial_a \mathcal{L}(a, u)\, \delta a + \partial_u \mathcal{L}(a,u)\, \partial_a u(a) \delta a. \tag{8.17}$$

La partie difficile à calculer est $\partial_a u(a)$. Si nous pouvons *choisir* $p \in Z$ de façon que ce terme disparaisse, nous aurons une expression simple pour la dérivée de J. Pour cela, nous considérerons que $\delta u = \partial_a u(a)\, \delta a$ est une quantité indépendante, et nous demandons que l'opérateur $\delta u \to \partial_u \mathcal{L}(a,u)$ s'annule. Définissons alors *l'équation adjointe* abstraite par :

$$\partial_u \mathcal{L}(a, u(a))\, \delta u = 0, \quad \forall \delta u \in U, \tag{8.18}$$

soit, en explicitant l'expression de $\partial_u \mathcal{L}$

$$(H\delta u, Hu(a) - d_{\text{obs}}) + (p, \partial_u F(a, u(a))\, \delta u) = 0, \quad \forall \delta u \in U. \tag{8.19}$$

Nous retrouvons précisément l'équation adjointe (8.14), puis la différentielle de J se calcule par la formule :

$$J'(a)\delta a = (p, \partial_a F(a, u(a))\, \delta a) \tag{8.20}$$

et nous voyons qu'en fait nous obtenons le gradient de J :

$$\nabla J(a) = (\partial_a F(a, u(a)))^* p, \tag{8.21}$$

identique à (8.15).

Remarque 8.5. En pratique, la forme la plus utile de l'équation adjointe est la formulation ń variationnelle ż (8.19). En effet, comme nous l'avons déjà signalé, il n'est pas toujours commode de calculer les opérateurs adjoints. Par contre, il est toujours simple (nous le constaterons) de partir de la forme (8.19), et de la manipuler (par des intégrations ou des sommations par parties), pour aboutir à une équation adjointe explicite.

De même, il est souvent plus commode de partir de l'équation (8.20) et de la manipuler pour identifier le gradient que d'utiliser la formule (8.21) telle quelle.

Cette méthode est suffisamment importante pour que nous en résumions les étapes principales.

Théorème 8.2. *Le calcul du gradient de la fonctionnelle* (7.9) *passe par les étapes suivantes :*

- **i)** *Définir le lagrangien par* (8.16)*;*
- **ii)** *Résoudre l'équation adjointe (variationnelle)* (8.19) *qui détermine l'état adjoint p ;*
- **iii)** *La différentielle de J est donnée par* (8.20) *qui permet d'identifier le gradient de J.*

8.1.5 Le test du produit scalaire

Il est notoirement délicat de valider un calcul d'état adjoint. La méthode la plus fiable est basée sur la définition de l'adjoint. Soit donc $b \in Z$ et $c \in U$ deux vecteurs, et considérons les deux problèmes :

$$\partial_u F(a,u)v = b, \tag{8.22}$$

$$\partial_u F(a,u)^* p = c. \tag{8.23}$$

Par définition de l'adjoint, on doit avoir :

$$(\partial_u F(a,u)v,\, p) = (\partial_u F(a,u)^* p,\, v), \tag{8.24}$$

c'est-à-dire

$$(b,\, p) = (c,\, v). \tag{8.25}$$

C'est cette dernière équation qui constitue le *test du produit scalaire*. Son intérêt est qu'elle ne fait plus intervenir que les second membres et les solutions des équations, et non les opérateurs eux-mêmes. C'est une conséquence de la théorie, et elle est facile à vérifier après avoir programmé l'état adjoint et l'état direct (linéarisé). Sa validité numérique doit être totale. En pratique, on prend souvent le second membre b aléatoire, et on prend ensuite $c = v$. Si le test est vérifié pour plusieurs choix successifs de b, on peut avoir confiance dans la résolution de l'équation adjointe.

8.2 Exemples de calcul de gradient

Nous allons appliquer la théorie développée au paragraphe précédent aux exemples que nous avons commencé à examiner : le modèle elliptique en dimension 1 et le système d'équations différentielles. Ces deux exemples sont en un sens typiques, le premier des problèmes stationnaires, le second des problèmes d'évolution. Nous verrons ensuite comment le cas d'une équation aux dérivées partielles, d'abord elliptique, puis parabolique (l'équation de la chaleur) peut se traiter en utilisant l'état adjoint.

8.2.1 Équation elliptique en dimension 1

Rappelons que l'équation d'état est un problème aux limites en dimension 1 :

$$\begin{cases} -bu''(x) + cu'(x) = f(x) \quad 0 < x < 1 \\ u(0) = 0,\ u'(1) = 0. \end{cases} \tag{8.26}$$

et que l'observation peut prendre l'une des trois formes vues à l'exemple 7.1, et que les fonctions coût correspondantes ont été définies par les équations (7.11) à (7.13).

Nous n'aborderons pas ici la question de la différentiabilité des applications considérées. Nous nous contenterons de raisonner formellement, en supposant que les dérivées que nous voulons calculer existent.

Avant de commencer le calcul du gradient proprement-dit, notons que la formulation variationnelle de (8.26) est

$$\int_0^1 bu'(x)s'(x) + cu'(x)s(x)\,dx = \int f(x)s(x)\,dx, \forall s \in V = \left\{ s \in H^1(0,1), s(0) = 0 \right\}. \tag{8.27}$$

L'existence et l'unicité de la solution de ce problème sont des conséquences du théorème de Lax–Milgram, si b et c sont strictement positifs. Il suffit de vérifier la coercivité :

$$\int_0^1 bu'(x)^2 + cu'(x)u(x)\,dx = b\int_0^1 u'(x)^2\,dx + cu(1)^2 \geq \|u\|^2_{H^1_0},$$

d'après l'inégalité de Poincaré.

Calcul direct

Commençons par le calcul direct du gradient, et prenons pour fixer les idées, le cas de $J_1(b,c) = \int_0^1 |u(x) - d_{\text{obs}}|^2\,dx$. Nous allons effectuer un développement de Taylor au premier ordre de J_1 par rapport à b et c :

$$J_1(b+\delta b, c+\delta c) = \int_0^1 |u+\delta u(x) - d_{\text{obs}}|^2 = J_1(b,c) + \int_0^1 (u(x) - d_{\text{obs}})\delta u(x)\,dx,$$

et δu correspond à la dérivée directionnelle de l'application $(b,c) \to u$ dans la direction $\delta b, \delta c$. Pour la calculer nous pouvons dériver formellement l'équation (8.55) par rapport aux paramètres b et c, et obtenir une équation dont les dérivées partielles de u par rapport à b et c sont solutions. Nous obtenons le système suivant, en notant $v = \dfrac{\partial u}{\partial b}$, $w = \dfrac{\partial u}{\partial c}$:

$$\begin{cases} -bv''(x) + cv'(x) = u'' \quad 0 < x < 1 \\ v(0) = 0,\ v'(1) = 0, \\ -bw''(x) + cw'(x) = -u' \quad 0 < x < 1 \\ w(0) = 0,\ w'(1) = 0, \end{cases} \tag{8.28}$$

et $\delta u = v\delta b + w\delta c$.

On en déduit

$$J'(b,c)\begin{pmatrix} b \\ c \end{pmatrix} = \int_0^1 (u(x) - d_{\text{obs}}(x))v(x)\,dx\ \delta b + \int_0^1 (u(x) - d_{\text{obs}}(x))w(x)\,dx\ \delta c,$$

ce qui donne les dérivées partielles de J :

$$\nabla J_1(b,c) = \begin{pmatrix} \displaystyle\int_0^1 (u(x) - d_{\text{obs}}(x))v(x)\,dx \\ \displaystyle\int_0^1 (u(x) - d_{\text{obs}}(x))w(x)\,dx \end{pmatrix} \tag{8.29}$$

si v et w sont solutions du système linéarisé (8.28).

Comme nous l'avons annoncé au paragraphe 8.1.2, ce calcul est exact, et son coût est proportionnel au nombre de paramètres.

Pour les deux autres cas d'observation, nous conservons le calcul des dérivées v et w, et les gradients s'en déduisent facilement par :

$$\nabla J_2(b,c) = \begin{pmatrix} (u(1) - d_{\text{obs}}(1))v(1) \\ (u(1) - d_{\text{obs}}(1))w(1), \end{pmatrix} \tag{8.30}$$

et

$$\nabla J_3(b,c) = \begin{pmatrix} (u(1/2) - d_{\text{obs}}(1/2))v(1/2) + (u(1) - d_{\text{obs}}(1))v(1) \\ (u(1/2) - d_{\text{obs}}(1/2))w(1/2) + (u(1) - d_{\text{obs}}(1))w(1). \end{pmatrix} \tag{8.31}$$

État adjoint

Nous passerons par le Lagrangien, défini (dans le cas de J_1) par :

$$\mathcal{L}((b,c),u,p) = \frac{1}{2}\int_0^1 |u(x) - d_{\text{obs}}(x)|\, dx + \int_0^1 (bu'(x) + cu'(x) - f(x))p(x)\, dx.$$

pour tout $p \in V$.

Remarquons tout d'abord que si $u = u_{(b,c)}$ est la solution de l'équation d'état (8.26) correspondant aux paramètres b et c, on a $\mathcal{L}((b,c),u_{(b,c)},p) = J(b,c)$.

Dérivons l'identité précédente, p étant fixé :

$$\begin{aligned} \frac{\partial J}{\partial b} &= \frac{\partial \mathcal{L}}{\partial b} + \frac{\partial \mathcal{L}}{\partial u} v \\ &= \int_0^1 (u'(x))p'(x)\, dx + \int_0^1 (u(x) - d_{\text{obs}}(x))v(x)\, dx + \int_0^1 bv'(x)p'(x) + cv'(x)p(x)\, dx, \end{aligned} \tag{8.32}$$

v étant la fonction définie en (8.28) (et l'équation analogue si l'on dérive par rapport à c). Cette égalité est valable poiur toute fonction p. Comme nous l'avons vu au paragraphe précédent, si nous calculons v (et w), le coût du calcul du gradient sera proportionnel au nombre de paramètres. Nous allons voir qu'un choix particulier de p va nous permettre d'éliminer v et w.

Supposons que nous puissions choisir p de façon que $\dfrac{\partial \mathcal{L}}{\partial u} = 0$ (nous allons voir plus bas que ce choix est possible). L'identité précédente devient simplement :

$$\frac{\partial J}{\partial b} = \frac{\partial \mathcal{L}}{\partial b} = \int_0^1 u'(x)p(x)\, dx, \tag{8.33}$$

$u = u_{(b,c)}$ étant solution de l'équation d'état, et p donné par le choix que nous venons d'indiquer.

Il nous reste à montrer que nous pouvons choisir p de sorte que $\dfrac{\partial \mathcal{L}}{\partial u} = 0$, ou de manière équivalente (et plus maniable piur notre propos) $\dfrac{\partial \mathcal{L}}{\partial u} s = 0, \forall s \in V$. D'après (8.32), p doit vérifier

$$\int_0^1 (u(x) - d_{\text{obs}}(x))s(x)\, dx + \int_0^1 bs'(x)p'(x) + cs'(x)p(x)\, dx = 0, \forall s \in V \tag{8.34}$$

On voit facilement que (8.34) est la formulation variationnelle d'un problème aux limites, qui a une solution unique d'après le théorème de Lax–Milgram. On vérifie aisément que ce problème aux limites est

$$\begin{cases} -bp''(x) - cp'(x) = -(u(x) - d_{\text{obs}}(x), \quad 0 < x < 1 \\ p(0) = 0,\ bp'(1) + cp(1) = 0. \end{cases} \tag{8.35}$$

Cette équation (ou sa formulation variationnelle (8.34)) s'apelle l'équation adjointe (de l'équation (8.26)). Noter que le signe devant le coefficient c est différent, et que le terme source de cette équation est (au

signe près) le résidu de l'observation correspondant à u. Ce phénomène est général, et nous le vérifierons à plusieurs reprises.

Avec ce choix de p, l'équation (8.33) devient

$$\frac{\partial J}{\partial b} = \frac{\partial \mathcal{L}}{\partial b} = \int_0^1 u'(x)p'(x)\,dx. \tag{8.36}$$

De plus, il est facile de se convaincre que l'on peut obtenir la dérivée partielle de J par rapport à c par un calcul analogue. L'équation adjointe est *la même*, et la dérivée partielle est donnée par

$$\frac{\partial J}{\partial c} = \frac{\partial \mathcal{L}}{\partial c} = \int_0^1 u'(x)p(x)\,dx. \tag{8.37}$$

Comme nous l'avions anoncé, le coût de calcul du gradient par cette méthode est (essentiellement) indépendant du nombre de paramètres, dans la mesure où l'on ne résout qu'une seule équation adjointe.

Il est important de noter que l'équation adjointe doit être résolue après l'équation d'état, puisque u figure au second membre de l'équation adjointe.

Pour conclure, montrons quelles sont les adaptations nécéssaires pour prendre en compte les autres formes de fonction coût proposées à l'exemple 7.1. D'après ce qui précède, nous voyons que seule le second membre de l'équation adjointe change : dans le cas d'une observation à droite de l'intervalle, l'équation adjointe est

$$\begin{cases} -bp''(x) - cp'(x) = 0, \quad 0 < x < 1 \\ p(0) = 0,\ bp'(1) + cp(1) = -(u(1) - d_{\text{obs}}(1)), \end{cases} \tag{8.38}$$

et si l'on mesure en plus en $1/2$, l'équation adjointe devient :

$$\begin{cases} -bp''(x) - cp'(x) = -(u(1/2) - d_{\text{obs}}(1/2)\delta(x - 1/2), \quad 0 < x < 1 \\ p(0) = 0,\ bp'(1) + cp(1) = -(u(1) - d_{\text{obs}}(1)), \end{cases} \tag{8.39}$$

les formules (8.36) et (8.37) qui donnent les dérivées partielles restent valables sans changement (mais p est la solution de l'équation adjointe appropriée).

8.2.2 Équation différentielle

Nous abordons maintenant un exemple de problème d'évolution. Nous reprenons l'exemple 7.2, et nous voulons calculer le gradient de la fonction coût (7.14). Comme au paragraphe précédent, il est possible de mener ce calcul sur le modèle continu, mais nous ne le ferons qu'après discrétisation.

L'état de l'art des méthodes numériques pour les équations différentielles ordinaires est très avancé. Il existe plusieurs logiciels généraux d'excellente qualité (les livres [33] et [34] constituent une référence récente et très complète). Dans la plupart des situations, il est conseillé d'utiliser un de ces codes plutôt que d'écrire soit même un solveur simple. Nous ne suivrons pas ici ce conseil pour les raisons suivantes :

– Tout d'abord, il est préférable de calculer le gradient à partir du schéma discret que de discrétiser un gradient continu. Cela serait possible, mais dans des situations plus complexes que celle envisagée ici, conduit à un schéma discret différent de ce que nous allons obtenir ci-dessous. L'avantage de notre choix est qu'il conduit au calcul du gradient *exact* de la fonction coût discrète. De plus, il évite de faire des choix, pas toujours évidents, pour discrétiser les différentes quantités. Pour une discussion plus étoffée sur ce point, on pourra consulter la discussion dans [15].

– L'équation adjointe dépend donc du schéma d'intégration utilisé pour le problème direct. Si ce schéma est lui-même sophistiqué (ordre élevé, pas variable), il sera plus délicat de former l'équation adjointe. Pour cette raison, nous nous contenterons d'un schéma très simple, en l'occurrence le schéma d'Euler explicite.
– L'autre raison est pédagogique. Une fois que l'on aura compris la structure du calcul dans le cas du schéma d'Euler, il sera (en principe) facile d'étendre les résultats à des méthodes plus efficaces.

Précisons donc la description du schéma d'approximation de (7.7). Étant donné $N \in \mathbf{N}$, notons $\Delta t = T/N$, et introduisons une subdivision de l'intervalle $[0,T]$: $0 = t^0 < t^1 < \cdots < t^N = T$, avec $t^n = nh$, $n = 0,\ldots,N$.

Le paramètre a est déjà discret dans cet exemple.

Par abus de notations, nous noterons y^i une approximation de $y(t^i)$. Pour fixer les idées, nous choisissons donc comme schéma d'approximation le schéma d'Euler explicite, défini par

$$\left\{ \begin{array}{l} \dfrac{y^n - y^{n-1}}{\Delta t} = f(y^{n-1}, a), \quad n = 1,\ldots,N, \\ y^0 = y_0. \end{array} \right. \tag{8.40}$$

En ce qui concerne l'observation, rappelons que nous avons supposé que l'on mesurait y à certains instants, notés $\tau_1,\ldots,\tau_Q$. Nous ferons l'hypothèse supplémentaire que le pas de temps δt est choisi de façon que chaque instant τ_q soit un multiple entier du pas de temps. Il s'agit là d'une simplification très forte, qui devrait être relaxée dans une application réaliste. Cela nous conduirait à introduire un opérateur d'interpolation, mais ne changerait pas fondamentalement le calcul. Nous noterons $n_q = \tau_q/\delta t$, pour $q = 1,\ldots,Q$.

La fonction coût a été définie en (7.14) comme :

$$J(a) = \frac{1}{2}\sum_{q=1}^{Q} \left| y_a(\tau_q) - d^q_{\mathrm{obs}} \right|^2, \tag{8.41}$$

où y_a désigne la solution de (8.40).

Dans cet exemple, il est difficile d'appliquer directement la théorie du paragraphe 8.1.3, en particulier de calculer la transposée de l'opérateur dérivée. Il est plus simple, et c'est ce que nous ferons, de passer par un lagrangien.

Celui-ci est défini par

$$\mathcal{L}_h(a,y,p) = \frac{1}{2}\sum_{q=1}^{Q} |y^{n_q} - d^q|^2 \Delta t + \sum_{n=1}^{N} \left(p^{n-1}\right)^t \left(\frac{y^n - y^{n-1}}{\Delta t} + f(y^{n-1},a) \right) \Delta t. \tag{8.42}$$

Commençons par former l'équation adjointe, en partant de (8.19) :

$$\frac{\partial \mathcal{L}}{\partial y}\delta y = \sum_{q=0}^{N} (y^q - d^q)^t \delta y^q \Delta t + \sum_{n=1}^{N} \left(p^{n-1}\right)^t \left(\frac{\delta y^n - \delta y^{n-1}}{\Delta t} - \partial_y f(y^{n-1},a)\delta y^{n-1} \right) \Delta t = 0, \tag{8.43}$$
$$\forall \delta y = (\delta y^0,\ldots,\delta y^N).$$

Nous allons rendre cette équation plus explicite. Pour cela, nous cherchons à y faire apparaître δy^n en facteur. Effectuons une ń intégration par parties discrète ż, c'est-à-dire un simple changement d'indices :

$$\sum_{n=1}^{N} \delta y^{n-1} p^{n-1} = \sum_{n=0}^{N-1} \delta y^n p^n,$$

ce qui nous permet de réécrire l'équation adjointe :

$$\sum_{q=1}^{Q}(y^{n_q}-d^q)^t\delta y^{n_q}\,\Delta t+\sum_{n=1}^{N}\left(p^{n-1}\right)^t\delta y^n-\sum_{n=0}^{N-1}(p^n)^t\,\delta y^n-\sum_{n=0}^{N-1}(p^n)^t\,\partial_y f(y^n,a)\delta y^n\Delta t=0 \tag{8.44}$$
$$\forall \delta y=\delta y^0,\ldots,\delta y^N$$

Définissons le vecteur r^n, $n=0,\ldots,N$ par

$$r^{n_q}=\begin{cases}d^q-y^{n_q}, & q=1,\ldots,Q,\\ 0 & \text{sinon.}\end{cases} \tag{8.45}$$

Puisque la condition initiale ne dépend pas de a, on doit avoir $\delta y^0=0$ (plus précisément y varie dans un espace affine de fonctions telles que $y^0=y_0$, et δy varie dans l'espace vectoriel associé). Par ailleurs, nous pouvons réécrire les deux dernières sommes de l'équation (8.44) comme des sommes de 0 à N, à condition d'introduire un nouveau multiplicateur p^N, et d'imposer $p^N=0$. Nous obtenons alors :

$$-\sum_{n=0}^{N}(r^n)^t\delta y^n\Delta t+\sum_{n=0}^{N}\left(p^{n-1}\right)^t\delta y^n-\sum_{n=0}^{N}(p^n)^t\,\delta y^n-\sum_{n=0}^{N}(p^n)^t\,\partial_y f(y^n,a)\delta y^n\Delta t=0 \tag{8.46}$$
$$\forall \delta y=(\delta y^0,\ldots,\delta y^N).$$

Comme cette équation est valable pour tout choix de $\delta y^0,\ldots,\delta y^N$, elle est valable si nous prenons toutes les variations nulles sauf celles correspondant à un pas de temps à la fois. Nous en déduisons, après transposition :

$$\frac{p^{n-1}-p^n}{\Delta t}+\partial_y f(y^n,a)^t p^n=r^n,\quad n=N,\ldots,1. \tag{8.47}$$

auquel nous devons ajouter la condition finale $p^N=0$.

L'équation adjointe apparaît donc comme un schéma aux différences *en temps rétrograde*. Ici, l'effet de la transposition est de renverser le temps. Comme nous l'avons déjà noté, cette équation est linéaire, puisque y^n est connu lorsque l'on calcule p^{n-1}.

Nous pouvons maintenant calculer la différentielle par l'équation (8.20) :

$$J'(a)\,\delta a=\sum_{i=0}^{N}(p^n)^t\,\partial_a f(y^n,a)\,\delta a, \tag{8.48}$$

et en déduire le gradient de J :

$$\nabla J(a)=\sum_{n=0}^{N}\partial_a f(y^n,a)^t p^n. \tag{8.49}$$

Remarque 8.6. Nous avons vu que l'équation (8.47) doit être intégrée en temps rétrograde. Cette remarque est une source de difficulté lors de la mise en œuvre informatique (pas pour cet exemple très simple, mais cela serait le cas pour des exemples plus réalistes, tel que celui du paragraphe suivant). En effet, la formule (8.49) montre que nous devons connaître simultanément y^n et p^{n-1} pour calculer le gradient de J. Or ces deux quantités se calculent en faisant évoluer le ń temps ż dans deux

sens différents. De plus, y^n est nécessaire au calcul de p^{n-1} dans l'équation adjointe. La manière naturelle de calculer le gradient est d'accumuler la somme dans (8.49) en même temps que l'on calcule p^{n-1}. Cela veut dire que nous devrons être capable de régénérer l'état direct y^n, soit en l'ayant stocké (éventuellement sur un support externe), soit en le recalculant.

Il est clair qu'aucune de ces deux solutions n'est satisfaisante.

– La première requiert un espace de stockage proportionnel au nombre d'étapes de calcul de l'état direct, ce qui est inacceptable pour des problèmes réalistes (penser que nous avons en vue la discrétisation de problèmes d'évolutions, et que dans ce cas, le nombre de pas de temps est généralement lié à la taille en espace du problème) ;
– Dans le cas général, la seconde demandera un temps de calcul inacceptable. En effet, le problème direct n'a aucune raison d'être réversible en temps, et donc le calcul de u^n demande d'intégrer l'équation directe jusqu'à l'instant n. De même, pour calculer u^{n-1} il faut recommencer l'intégration jusqu'à l'instant $n-1$, et ainsi de suite.

Un compromis ń optimal ż a été trouvé par A. Griewank [56], et consiste en une stratégie de reprise. Les points de reprise sont déterminés de façon à minimiser le nombre de recalculs de trajectoire, la mémoire disponible pour les sauvegardes étant fixée. Cette stratégie consiste à introduire une notion de ń niveau ż, chaque niveau correspondant à une sauvegarde de l'état direct. L'ensemble s'organise en une sorte d'arbre, que l'on parcourt de façon récursive. Griewank appelle cette procédure *treeverse*, et fournit un sous-programme en Fortran.

Concluons ce paragraphe par un complément important : le test du produit scalaire pour valider le calcul de l'état adjoint. Ici encore, il est peu commode d'appliquer directement la formule générale (8.25). Il est plus facile de repartir de l'équation adjointe (8.47). On prend le produit scalaire de chaque équation avec un vecteur (qui sera précisé plus bas) δy^n, et on somme toutes les équations obtenues. Nous obtenons :

$$\sum_{n=1}^{N} \Big(\frac{p^{n-1}-p^n}{\Delta t} + \partial_y f(y^n,a)^t p^n \Big)^t \delta y^n = \sum_{n=1}^{N} (r^n)^t \delta y^n. \tag{8.50}$$

Nous effectuons encore une ń intégration par parties discrète ż, et il vient :

$$\sum_{n=0}^{N-1} \big(\frac{p^n}{\Delta t}\big)^t \delta y^{n+1} + \sum_{n=1}^{N} \Big(-\frac{p^n}{\Delta t} + \partial_y f(y^n,a)^t p^n \Big)^t \delta y^n = \sum_{n=1}^{N} (r^n)^t \delta y^n \tag{8.51}$$

soit (puisque $p^n = 0$, et en choisissant $\delta y^0 = 0$),

$$\sum_{n=0}^{N} \big(p^n\big)^t \left(\frac{\delta y^{n+1}-\delta y^n}{\delta t} + \partial_y f(y^n,a)\delta y^n \right) = \sum_{n=1}^{N} (r^n)^t \delta y^n. \tag{8.52}$$

Supposons maintenant donné un second membre arbitraire $(b^n)_{n=0,\dots,N}$, et choisissons δy^n solution de *l'équation d'état linéarisée* :

$$\frac{y^{n+1}-y^n}{\delta t} + \partial_y f(y^n,a)\delta y^n = b^n, \quad n = 0,\dots,N \tag{8.53}$$

avec la condition initiale $\delta y^0 = 0$, l'équation (8.52) devient simplement :

$$\sum_{n=0}^{N} \big(p^n\big)^t b^n = \sum_{n=0}^{N} r^{nt} \delta y^n. \tag{8.54}$$

L'intéret de cette égalité vient de ce qu'il est difficile de valider un calcul d'état adjoint, dont la solution n'a aucune signification physique. En contôlant cette égalité avec plusieurs seconds membres choisis aléatoirement, on peut gagner un niveau de confiance élevé dans un calcul d'état adjoint. Pour mener à bien ce calcul, il faudra programmer la résolution de l'équation linéarisée, ce qui dans notre cas n'aurait pas été nécessaire. Dans un cas plus réaliste, en tout cas avec un schéma implicite, nous aurions du de toutes façons calculer le jacobien $\partial_y f(y^n, a)$.

Il sera utile de compléter cette ń validation ż (qui formellement n'en est pas une) par une vérification du gradient par différences finies.

8.2.3 Équation elliptique en dimension 2

Nous allons considérer une généralisation de l'exemple vu en détail au paragraphe 8.2.1. Étant donné un ouvert $\Omega \subset \mathbf{R}^2$ (pour fixer les idées), nous considérons le problème :

$$\left\{\begin{array}{rl} -\operatorname{div}(a \operatorname{grad} u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \Gamma_D \\ a\dfrac{\partial u}{\partial n} = g & \text{sur } \Gamma_N \end{array}\right. \tag{8.55}$$

où Γ_D et Γ_N forment une partition de $\partial\Omega$, $f \in L^2(\Omega)$ et $g \in L^2(\Gamma_N)$ sont données.

Le paramètre à identifier est a, et nous ne considérons ici que le cas où l'on mesure u sur la partie Γ_N du bord (puisque la valeur de u sur la partie Γ_D est imposée, cela n'aurait pas de sens de la mesurer). La fonction coût est alors

$$J(a) = \frac{1}{2}\int_{\Gamma_N} \left|u_{a|\Gamma_N}(x) - d_{\text{obs}}(x)\right|^2 d\gamma(x). \tag{8.56}$$

L'adaptation de la méthode aux autres situations de l'exemple 7.3 est facile, et laissée au lecteur.

Concernant le problème direct, c'est une conséquence classique du théorème de Lax–Milgram que le problème (8.55) admet une solution unique dans l'espace $U = \{u \in H^1(\Omega),\ u = 0 \text{ sur } \Gamma_D\}$. La formulation variationnelle de ce problème est :

$$\int_\Omega a(x) \operatorname{grad} u(x) \operatorname{grad} v(x)\, dx = \int_\Omega f(x)v(x)\, dx + \int_{\Gamma_N} g(x)v(x)\, d\gamma(x), \quad \forall v \in U \tag{8.57}$$

Il est possible de mener la théorie dans le cadre continu. Toutefois, ce calcul est quelque peu abstrait, et comme il faudra finalement passer en discret, nous ne mènerons le calcul qu'après discrétisation. Nous approcherons bien entendu l'équation d'état par éléments finis, que nous prendrons linéaires par morceaux pour simplifier. Nous supposerons que l'ouvert Ω est polygonal, de sorte que nous pouvons le recouvrir par une triangulation régulière $\mathcal{T}_h$. Nous approchons alors U par $U_h = \{u_h \in C^0(\bar{\Omega}),\ u_h|_K \in P^1, \forall K \in \mathcal{T}_h,\ u_h = 0 \text{ sur } \Gamma_D\} \subset U$. Nous noterons N_h le nombre de sommets du maillage *qui ne sont pas situés sur* Γ_D, et N_Γ le nombre de sommets sur Γ_N.

Nous devons également choisir une paramétrisation pour le paramètre a. Pour suivre la méthode générale décrite au paragraphe 7.2.2, nous prenons pour a la discrétisation la plus fine possible : nous prendrons une valeur par triangle. Nous noterons donc a_h le vecteur représentant la fonction approchée, $a_h = (a_K,\ K \in \mathcal{T}_h)$. La discrétisation de la mesure est naturellement une valeur par noeud du bord : d_{obs} sera approché par le vecteur $d_h = (d_M,\ M \text{ sommet de } \Gamma_N)$.

Il est classique qu'après discrétisation, le problème (8.55) se ramène à un système linéaire

$$F(a,u_h) = K(a)\,u_h - L = 0, \tag{8.58}$$

où le vecteur inconnu u_h, la matrice de rigidité $K(a)$ et le second membre L sont définis de la façon suivante :

Pour u_h : Nous identifierons la fonction inconnue $u_h \in U_h$ avec le vecteur de ses composantes dans la base canonique des éléments finis, c'est-à-dire le vecteur de ses valeurs aux sommets du maillage qui ne sont pas situés sur le bord Γ_D.

Pour $K(a)$: La matrice $K(a)$ peut-être définie implicitement par l'égalité :

$$\begin{aligned} \forall (u_h, v_h) \in U_h,\ v_h^t K(a) u_h &= \int_\Omega a_h(x)\,\mathrm{grad}\,u_h(x)\,\mathrm{grad}\,v_h(x)\,dx \\ &= \sum_{T\in\mathcal{T}_h} a_T \int_T \mathrm{grad}\,u_h(x)\,\mathrm{grad}\,v_h(x)\,dx. \end{aligned} \tag{8.59}$$

Plus explicitement, chaque élément de la matrice $K(a)$ est une intégrale du type précédent où les fonctions u_h et v_h sont des fonctions de base de l'espace U_h. Cette matrice se calcule habituellement par une procédure d'assemblage (voir par exemple[49]).

Pour L : Enfin, le second membre se calcule également par assemblage à partir de l'égalité

$$\forall v_h \in U_h,\ v_h^t L = \int_\Omega f(x) v_h(x)\,dx + \int_{\Gamma_N} g(x) v_h(x)\,d\gamma(x). \tag{8.60}$$

Il est en général nécessaire d'approcher cette intégrale par une formule de quadrature numérique.

Passons maintenant à l'observation. Elle consiste ici simplement à extraire d'un vecteur u_h ses composantes sur Γ_N. L'espace des observations D_h est l'espace de fonctions éléments finis sur le bord Γ_N, et nous identifierons également une telle fonction avec le vecteur de ses valeurs aux sommets de Γ_N.

L'opérateur d'observation peut (et doit, du point de vue numérique) se représenter par une matrice booléenne $H \in \mathbf{R}^{N_h \times N_\Gamma}$, définie par

$$\forall u_h \in U_h, \forall d_h \in D_h, \quad d_h^t (H u_h) = \int_{\Gamma_N} u_h(x) d_h(x)\,d\gamma(x). \tag{8.61}$$

La fonction coût peut se réécrire sous la forme :

$$\begin{aligned} J(a_h) &= \frac{1}{2}\int_{\Gamma_N} \left| u_h(a_h)(x) - d^h(x) \right|^2 d\gamma(x) \\ &= \frac{1}{2}(H u_h(a_h) - d_h)^t M (H u_h(a_h) - d_h), \end{aligned} \tag{8.62}$$

où la matrice M est définie par

$$\forall (d_h, e_h) \in D_h, e_h^t M d_h = \int_{\Gamma_N} d_h\, e_h\, d\gamma(x).$$

Il s'agit cette fois d'une matrice ń de masse ż, dont les éléments sont les produits scalaire de deux fonctions de base de D_h. Remarquons que cette fonction coût est bien une norme sur D_h mais n'est pas la norme euclidienne usuelle sur cet espace.

Finalement, après discrétisation, le problème discret se résume à chercher le minimum de la fonction coût (quadratique) définie par (8.62), $u_h(a_h)$ étant solution de l'équation d'état (linéaire) (8.58). Pour simplifier les notations, nous noterons u_h pour la solution de l'équation d'état (8.58) $u_h(a_h)$.

Il s'agit là d'un cas où l'on peut appliquer directement la théorie du paragraphe 8.1.3. Pour cela nous devons calculer les différentielles partielles de la fonction F définie en (8.58).

La dérivée par rapport à u_h est particulièrement simple, puisque l'application F est linéaire par rapport à u_h. On à donc :

$$\partial_{u_h} F(a_h, u_h)\delta u_h = K(a_h)\delta u_h.$$

F est également linéaire par rapport à a_h, et on a donc

$$\partial_{a_h} F(a_h, u_h)\delta a_h = \big(K'(a_h)\delta a_h\big)u_h, \tag{8.63}$$

mais l'expression de la différentielle est compliquée à écrire parce que l'application $a_h \to K(a_h)$ est à valeur matricielle. Pour calculer facilement cette différentielle, il est pratique de repartir de la définition variationnelle de $K(a_h)$ (8.59). Pour deux vecteurs (u_h, v_h) *fixés*, nous pouvons dériver l'expression (8.59), et nous obtenons :

$$v_h^t(K'(a_h)\delta a_h)u_h = \sum_{T\in\mathcal{T}_h} \delta a_T \int_T \operatorname{grad} u_h(x) \operatorname{grad} v_h(x)\, dx. \tag{8.64}$$

D'après (8.14), l'équation adjointe s'écrit :

$$K(a_h)^t p_h = -H^t\big(Hu_h - d_h\big), \tag{8.65}$$

où, rappelons le, u_h est solution de l'équation d'état, et le gradient s'en déduit alors par

$$\begin{aligned}
\delta a_h^t \nabla J(a_h) &= \delta a_h^t\, \partial_{a_h} F(a_h, u_h)^t p_h && \text{par (8.21)}\\
&= p_h^t \partial_{a_h} F(a_h, u_h)\delta a_h && \text{parce que l'expression est un scalaire}\\
&= p_h^t\big(K'(a_h)\delta a_h\big)u_h && \text{d'après (8.63)}\\
&= \sum_{T\in\mathcal{T}_h} \delta a_T \int_T \operatorname{grad} u_h(x) \operatorname{grad} p_h(x)\, dx && \text{par (8.64).}
\end{aligned}$$

Pour la dernière égalité, il est important de noter que (8.64) a été établi dans le cas où u_h et v_h sont des vecteurs *indépendants* de a_h. Toutefois, nous l'utilisons a posteriori, et le choix de u_h et p_h est alors légitime.

Finalement, les dérivées partielles de J sont données par

$$\frac{\partial J}{\partial a_T} = \int_T \operatorname{grad} u_h(x) \operatorname{grad} p_h(x)\, dx = |T|\,(\operatorname{grad} u_h \operatorname{grad} p_h)_{|T}, \quad \forall T \in \mathcal{T}_h, \tag{8.66}$$

puisque les gradients sont constants sur chaque triangle du maillage.

Il reste à interpréter l'équation adjointe comme un problème variationnel. Pour cela, prenons le produit scalaire de (8.65) avec un vecteur $v_h \in U_h$:

$$v_h^t K(a_h)^t p_h = -v_h^t H^t\big(Hu_h - d_h\big) \tag{8.67}$$

En transposant cette égalité, nous obtenons

$$p_h^t K(a_h) v_h = -\big(Hu_h - d_h\big)^t H v_h, \quad \forall v_h \in U_h. \tag{8.68}$$

Utilisons alors (8.59) pour le premier membre et (8.61) pour le second, et nous voyons que cette égalité est équivalente à l'équation variationnelle :

$$\int_\Omega a_h(x)\,\mathrm{grad}\,v_h(x)\,\mathrm{grad}\,p_h(x)\,dx = \int_{\Gamma_N} (Hu_h(x) - d_h(x))v_h(x)\,d\gamma(x), \quad \forall v_h \in U_h. \tag{8.69}$$

Il est facile de se convaincre que cette dernière équation est une formulation variationnelle discrétisée de l'équation aux dérivées partielles

$$\begin{cases} -\,\mathrm{div}(a\,\mathrm{grad}\,p) = 0 & \text{dans } \Omega \\ p = 0 & \text{sur } \Gamma_D \\ a\dfrac{\partial p}{\partial n} = d_{\mathrm{obs}} - u & \text{sur } \Gamma_N \end{cases} \tag{8.70}$$

Comme prévu, l'équation adjointe est linéaire (mais ici l'équation d'état l'était aussi), et son second membre provient du résidu de l'observation.

Dans ce cas, le test du produit scalaire prend la forme suivante : On se donne un second membre arbitraire b_h, et on résout d'abord le problème direct

$$K(a)\,u_h = b_h, \tag{8.71}$$

puis le problème adjoint

$$K(a)^t\,p_h = u_h, \tag{8.72}$$

et l'on doit vérifier que

$$p_h^t b_h = \|u_h\|_2^2. \tag{8.73}$$

Pour résumer, l'équation adjointe fait intervenir la *transposée* de la matrice de rigidité (qui ici est symétrique), puis le gradient s'obtient facilement par l'équation (8.66).

8.2.4 Équation de la chaleur

Cet exemple est en quelque sorte la synthèse des deux précédents : nous considérons une discrétisons de l'équation de la chaleur par éléments finis en espace et différences finies en temps. Nous suivrons dans ce paragraphe des notes de G. Chavent [14]. Nous partirons de la formulation variationnelle de l'équation de la chaleur, avant de donner un schéma discret.

L'équation de la chaleur est le modèle de base qui régit les phénomènes de diffusion et intervient dans un grand nombre de domaines de la physique. Étant donné un ouvert $\Omega \subset \mathbf{R}^2$ (pour fixer les idées) et un réel $T > 0$, nous considérons le problème :

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mathrm{div}(a\,\mathrm{grad}\,u) = f & \text{dans } \Omega\times]0,T[\\ u(x,t) = 0 & \text{sur } \Gamma_D\times]0,T[\\ a\dfrac{\partial u}{\partial n} = g & \text{sur } \Gamma_N\times]0,T[\\ u(x,0) = u_0(x) & \text{sur } \Omega, \end{cases} \tag{8.74}$$

où $f \in L^2(0,T;\, L^2(\Omega))$, $g \in L^2(0,T;\, L^2(\Gamma_N))$, et $u_0 \in L^2(\Omega)$ sont des fonctions données (et supposées connues), et nous cherchons à identifier la fonction a. Comme au paragraphe 8.2.3, nous ferons l'hypothèse que u est mesurée sur la partie Γ_N du bord, mais également que $u(x,T)$ est connue sur tout Ω à l'instant final. Dans ces conditions, les données consistent en deux fonctions : $\hat{d}_N \in L^2(0,T;\, \Gamma_N)$ et $\hat{d}_T \in L^2(\Omega)$. Avec les réserves maintenant habituelles sur son caractère non-hilbertien, le choix naturel pour l'espace M est $M = L^\infty(\Omega)$, et $M_{\text{ad}} = \left\{ a \in M,\ a(x) \geq a_* > 0 \right\}$.

La fonction coût est

$$J(a) = \frac{1}{2}\int_0^T \int_{\Gamma_N} \left| u - \hat{d}_N \right|^2 \, dx\, dt + \frac{1}{2}\int_\Omega \left| u(x,T) - \hat{d}_T \right|^2 \, dx \tag{8.75}$$

Il est standard de mettre (8.74) sous forme variationnelle, et d'en déduire à la fois un théorème d'existence et la méthode d'approximation. Il est usuel de poser

$$H = L^2(\Omega), \quad V = \{ v \in H^1(\Omega), v = 0 \text{sur } \Gamma_D \} \tag{8.76}$$

La formulation faible est alors

$$\left\{ \begin{array}{l} \text{Chercher } u \in L^2(0,T;V) \cap C^0(0,T;H) \quad \text{tel que} \\ \dfrac{d}{dt}\displaystyle\int_\Omega u(t)v + \int_\Omega a \operatorname{grad} u(t) \operatorname{grad} v = \int_\Omega f(t)v + \int_{\Gamma_N} g(t)v, \quad \forall v \in V, \quad \text{p.p. sur }]0,T[\\ u(0) = u_0 \end{array} \right. \tag{8.77}$$

On peut énoncer un résultat d'existence pour le problème (8.77). Nous renvoyons le lecteur à [19, vol. 8] pour plus de détails et nous passerons directement au schéma d'approximation. En ce qui concerne la discrétisation en espace, nous utiliserons une méthode d'éléments finis, en reprenant les notations du paragraphe 8.2.3, et en ce qui concerne la discrétisation en temps, nous prendrons le schéma de Crank-Nicolson (voir [19, vol. 9], [55, 49] pour plus de détails tant sur les éléments finis que les schémas pour les problèmes d'évolution). Nous ne supposerons pas que la discrétisation en espace est nécessairement uniforme, et nous poserons donc, pour une partition $0 = t^0 < t^1 < \cdots < t^K = T$, $\Delta t^{k+1/2} = t^{k+1} - t^k$. Ce choix est justifié par la physique du problème : on sait, en effet, que les solutions de (8.74) deviennent de plus en plus régulières au cours du temps, et il sera naturel de vouloir augmenter le pas de temps au cours de la simulation. De plus nous verrons que dans ce cas, le problème adjoint discret ne sera pas nécessairement celui qui aurait été obtenu par discrétisation d'un état adjoint continu. Il sera commode, dans la suite de poser :

$$u_h^{k+1/2} = \frac{u_h^k + u_h^{k+1}}{2}, \quad k = 0, \ldots, K-1$$

Avec ces notations, le problème direct est :

$$\left\{ \begin{array}{l} \text{Chercher } u_h = (u_h^0, u_h^1, \ldots, u_h^K) \in U_h^{K+1}, \quad \text{tel que, pour tout } v_h \in U_h \\ I_\Omega^h\left(\dfrac{u_h^{k+1} - u_h^k}{\Delta t^{k+1/2}} v_h \right) + I_\Omega^h\left(a_h \operatorname{grad} u_h^{k+1/2} \operatorname{grad} v_h \right) = I_\Omega^h\left(f_h^{k+1/2} v_h \right) + I_{\Gamma_N}^h\left(g_h^{k+1/2} v_h \right), \\ I_\Omega^h(u_h^0 v_h) = I_\Omega^h(u_0 v_h), \quad \forall v_h \in U_h \end{array} \right. \tag{8.78}$$

où $f_h^{k+1/2}$ et $g_h^{k+1/2}$ sont des approximations de f_h et g_h à l'instant $t^{k+1/2}$. La dernière équation de (8.78) définit la condition initiale discrète par *projection* sur $L^2(\Omega)$, ce qui se signifie simplement

que u_h^0 est la fonction de U_h qui prend les mêmes valeurs que u_0 aux points du maillage situés hors de Γ_D.

Le dernier point à préciser concerne l'approximation des intégrales en temps. Par analogie avec le schéma numérique, nous poserons, pour θ définie sur $[0,T]$ (c'est la règle du trapèze) :

$$I_{\Delta t}(\theta) \approx \int_0^T \theta(t)\,dt = \sum_{k=0}^{K-1} \frac{\theta(t^k)+\theta(t^{k+1})}{2}\,\Delta t^{k+1/2}. \tag{8.79}$$

Nous pouvons alors définir le lagrangien, par :

$$\begin{aligned}\mathcal{L}_h(a_h,u_h,p_h) &= \frac{1}{2}\sum_{k=0}^{K-1} I_\Omega^h\left\{\frac{1}{2}\left(\left|u_h^k-\hat{d}_h^k\right|^2+\left|u_h^{k+1}-\hat{d}_h^{k+1}\right|^2\right)\right\}\Delta t^{k+1/2}+\frac{1}{2}I_{\Gamma_N}^h\left(\left|u_h^K-\hat{d}_T\right|^2\right)\\ &+\sum_{k=0}^{K-1}\left\{I_\Omega^h\left(\frac{u_h^{k+1}-u_h^k}{\Delta t^{k+1/2}}p_h^{k+1/2}\right)+I_\Omega^h\left(a_h\,\mathrm{grad}\,u_h^{k+1/2}\,\mathrm{grad}\,p_h^{k+1/2}\right)\right\}\Delta t^{k+1/2}\\ &-\sum_{k=0}^{K-1}\left\{I_\Omega^h\left(f_h^{k+1/2}p_h^{k+1/2}\right)+I_{\Gamma_N}^h\left(g_h^{k+1/2}p_h^{k+1/2}\right)\right\}\Delta t^{k+1/2}\end{aligned} \tag{8.80}$$

Suivant la procédure habituelle, nous commençons par écrire l'équation adjointe. Nous noterons $p_h=(p_h^{1/2},\ldots,p_h^{K-1/2})$ le multiplicateur. L'équation adjointe ń abstraite ż est

$$\frac{\partial\mathcal{L}_h}{\partial u_h}\delta u_h = 0, \quad \forall \delta u_h = (0,\delta u_h^1,\ldots,\delta u_h^K)\in U_h^{K+1}.$$

Nous obtenons donc, pour $\delta u_h \in U_h^{K+1}$:

$$\begin{aligned}&\sum_{k=1}^{K}\frac{1}{2}I_\Omega^h\left\{\left(u_h^k-\hat{d}_h^k\right)\delta u_h^k+\left(u_h^{k+1}-\hat{d_h^{k+1}}\right)\delta u_h^{k+1}\right\}\Delta t^{k+1/2}+I_{\Gamma_N}^h\left((u_h^K-\hat{d}_T)\delta u_h^K\right)\\ &+\sum_{k=0}^{K-1}\left\{I_\Omega^h\left(\frac{\delta u_h^{k+1}-\delta u_h^k}{\Delta t^{k+1/2}}p_h^{k+1/2}\right)+I_\Omega^h\left(a_h\,\mathrm{grad}\,\delta u_h^{k+1/2}\,\mathrm{grad}\,p_h^{k+1/2}\right)\right\}\Delta t^{k+1/2}=0\end{aligned} \tag{8.81}$$

Nous pratiquons ensuite une intégration par parties discrète, c'est-à-dire un décalage d'indices, de façon à faire apparaître δu_h^k en facteur. L'équation précédente devient

$$\begin{aligned}&\sum_{k=1}^{K}\frac{1}{2}I_\Omega^h\left\{\left(u_h^k-\hat{d}_h^k\right)\delta u_h^k+\left(u_h^{k+1}-\hat{d_h^{k+1}}\right)\delta u_h^{k+1}\right\}\Delta t^{k+1/2}+I_{\Gamma_N}^h\left((u_h^K-\hat{d}_T)\delta u_h^K\right)\\ &+\sum_{k=1}^{K}I_\Omega^h\left(\delta u_h^k p_h^{k-1/2}\right)+\frac{1}{2}\sum_{k=1}^{K}I_\Omega^h\left(a_h\,\mathrm{grad}\,\delta u_h^k\,\mathrm{grad}\,p_h^{k-1/2}\right)\Delta t^{k-1/2}\\ &-\sum_{k=0}^{K-1}I_\Omega^h\left(\delta u_h^k p_h^{k+1/2}\right)+\frac{1}{2}\sum_{k=0}^{K-1}I_\Omega^h\left(a_h\,\mathrm{grad}\,\delta u_h^k\,\mathrm{grad}\,p_h^{k+1/2}\right)\Delta t^{k+1/2}=0\end{aligned} \tag{8.82}$$

Rappelons que nous avons $\delta u_h^0=0$, ce qui élimine les termes avec $k=0$ dans l'égalité précédente.

Pour interpréter l'équation adjointe, nous procédons en deux étapes : nous allons choisir des perturbations δu_h d'un type particulier, n'agissant que sur un seul instant. Nous traiterons d'abord le cas général $k<K$, puis l'instant final $k=K$.

Choisissons donc $\delta u_h = (0,\dots,v_h,\dots 0)$, où v_h est en position k, pour $k = 1,\dots,K-1$. En posant $\Delta t^k = 1/2(\Delta t^{k-1/2} + \Delta t^{k+1/2})$, nous obtenons, pour tout $k = 1,\dots,K-1$:

$$\Delta t^k I^h_{\Gamma_N}(u_h^k - \hat{d}_h^{\,k}) + I^h_\Omega(v_h p_h^{k+1/2}) - I^h_\Omega(v_h p_h^{k-1/2}) + \frac{\Delta t^{k+1/2}}{2} I^h_\Omega(a_h \operatorname{grad} v_h \operatorname{grad} p_h^{k+1/2}) + \frac{\Delta t^{k-1/2}}{2} I^h_\Omega(a_h \operatorname{grad} v_h \operatorname{grad} p_h^{k-1/2}) = 0, \tag{8.83}$$

que nous pouvons réécrire sous une forme analogue à (8.78) :

$$I^h_\Omega\left(\frac{p_h^{k-1/2} - p_h^{k+1/2}}{\Delta t^k} v_h\right) + I^h_\Omega\left\{a_h\left(\frac{\Delta t^{k-1/2}}{2\Delta t^k}\operatorname{grad} p_h^{k-1/2} + \frac{\Delta t^{k+1/2}}{2\Delta t^k}\operatorname{grad} p_h^{k+1/2}\right)\operatorname{grad} v_h\right\} = -I^h_{\Gamma_N}\left((u_h^k - \hat{d}_h^{\,k})v_h\right), \quad \forall v_h \in U_h,\ \forall k = 1,\cdots,K-1. \tag{8.84}$$

Il reste un degré de liberté : choisissons $\delta u_h = (0,\cdots,v_h)$ dans (8.82). Il vient :

$$\frac{\Delta t^{K-1/2}}{2} I^h_{\Gamma_N}\left((u_h^K - \hat{d}_h^{\,K})v_h\right) + I^h_\Omega\left((u_h^K - \hat{d}_T)v_h\right) + I^h_\Omega\left(v_h p_h^{K-1/2}\right) + \frac{1}{2} I^h_\Omega\left(a_h \operatorname{grad} v_h \operatorname{grad} p_h^{K-1/2}\right) = 0, \quad \forall v_h \in U_h.$$

Cette équation définit $p_h^{K-1/2}$. Nous pouvons la mettre sous la même forme que pour $k < K$, en introduisant un multiplicateur fictif. Posons

$$p_h^{K+1/2} = -(u_h^K - \hat{d}_T), \quad \text{et} \quad \Delta t^K = 1/2\Delta t^{K-1/2}.$$

Remarque 8.7. L'équation (8.84), y compris la condition finale ci-dessus, correspond à une discrétisation (consistante) de l'équation de la chaleur rétrograde :

$$\begin{cases} -\dfrac{\partial p}{\partial t} - \operatorname{div}(a \operatorname{grad} p) = 0 & \text{dans } \Omega\times]0,T[\\ p = 0 & \text{sur } \Gamma_D\times]0,T[\\ a\dfrac{\partial p}{\partial n} = -(u - \hat{d}_N) & \text{sur } \Gamma_N\times]0,T[\\ p(x,T) = -(u(x,T) - \hat{d}_T) & \text{sur } \Omega, \end{cases}$$

mais les différents facteurs Δt rendent cette discrétisation non-standard. Il est peu probable que nous ayons pu obtenir ce schéma particulier en discrétisant le schéma (8.78).

Remarque 8.8. Comme l'observation avait lieu à l'instant final, la condition finale sur l'état adjoint n'est pas homogène.

Nous pouvons enfin calculer le gradient de J_h :

$$\delta J_h = \frac{\partial \mathcal{L}}{\partial a_h} = \sum_{k=0}^{K-1} I^h_\Omega\left(\delta a_h \operatorname{grad} u_h^{k+1/2} \operatorname{grad} p_h^{k+1/2}\right) \Delta t^{k+1/2}, \tag{8.85}$$

ce qui donne, en revenant à la définition de I^h_Ω, la dérivée partielle par rapport aux éléments de a_h :

$$\frac{\partial J_h}{\partial a_T} = |K| \sum_{k=0}^{K-1} \Delta t^{k+1/2}\left(\operatorname{grad} u_{h|T}^k + \operatorname{grad} u_{h|T}^{k+1}\right)\operatorname{grad} p_{h|T}^{k+1/2} \tag{8.86}$$

8.3 Paramétrisation et organisation générale

Nous avons maintenant tous les éléments pour décrire l'organisation générale d'un programme cherchant à résoudre un problème inverse du type de ceux vus ci-dessus. Nous insisterons sur une organisation *modulaire*, pour toutes les raisons citées habituellement et aussi pour quelques raisons spécifiques, en particulier pour assurer l'indépendance du calcul de gradient vis-à-vis de la paramétrisation. Nous supposerons que nous disposons d'un sous-programme résolvant le problème d'optimisation. Ce sous-programme appellera le *simulateur* qui résout le problème direct. La séquence d'appel du simulateur est fixée par le sous-programme d'optimisation. Pour fixer les idées, nous supposerons que la séquence d'appel est :

```
simul(x, f, g)
```

où `x` représente le point courant de l'optimisation (a^n dans les notation du paragraphe 7.3), `f` est la valeur de la fonction coût au point `x` ($J(a^n)$) et `g` est le (vecteur) gradient au même point ($\nabla J(a^n)$). En fait, ce sous-programme sera nécessairement une interface entre l'optimiseur et la vraie routine de simulation dont la séquence d'appel est forcément plus complexe. Notre description de haut niveau oublie en effet un grand nombre de détails dont il conviendrait de se préoccuper dans une mise-en-œuvre réelle. Par exemple, la taille du problème (taille des vecteurs `x` et `g`) n'apparaît pas explicitement. Selon le langage utilisé, cet oubli est ou non raisonnable. En Fortran ou en C, il faut préciser la taille des vecteurs passés en arguments. Dans les langages comme C++, et plus encore avec les environnements de haut niveau comme Matlab ou Scilab la taille du vecteur peut s'obtenir à partir du vecteur lui-même. Un point plus important est que dans les cas qui nous intéressent ici (et en fait dans l'immense majorité des situations), le simulateur requiers d'autres données que simplement le point `x`. En pensant aux exemples vus ci-dessus, on constate qu'il faut pouvoir spécifier la géométrie du problème, les pas de temps, les sources, etc. Encore une fois cet aspect, fort important en pratique, est entièrement dépendant du langage utilité, et déborde largement du cadre de ce cours. Une proposition élégante pour le langage C++ a été formulée par Symes et Gockenbach [27].

Comme nous l'avons déjà signalé plusieurs fois, la *représentation* m_h du paramètre à identifier dans le logiciel de simulation sera en général différente de la *paramétrisation* m_{opt} utilisée pour le problème inverse. Si nous suivons les recommandations de Chavent [15, 16], nous devons prendre la représentation la plus fine compatible avec l'équation d'état. Une conséquence est qu'il existe alors une application de paramétrisation $M : \quad a_{\text{opt}} \mapsto a_h$ permettant de passer du paramétrage ń grossier ż utilisé pour l'optimisation à la représentation ń fine ż pour la simulation. Il est d'ailleurs souvent utile d'introduire une seconde application de projection P qui force le paramètre $M(a_{\text{opt}})$ à satisfaire les contraintes du sous-programme de modélisation (par exemple de positivité).

L'application à minimiser (ce que doit réaliser le sous-programme `simul`) est alors

$$j(a_{\text{opt}}) = J(P(M(a_{\text{opt}}))) \tag{8.87}$$

et son gradient se calcule par application du théorème des fonctions composées. Ceci se traduit par la post-multiplication du gradient de J par la matrice jacobienne *transposée* de P, puis de M. Ces calculs simples se font analytiquement. L'avantage de procéder de cette manière est qu'il est simple de changer de paramétrisation, ou de rajouter une projection. Il est également simple de rajouter un terme de régularisation $1/2\varepsilon^2 \left\| a_{\text{opt}} - a_0 \right\|^2$, et de rajouter le terme $\varepsilon^2(a_{\text{opt}} - a_0)$ au gradient. Le calcul coûteux du gradient par état adjoint est de toute façon nécessaire, mais est indépendant de la paramétrisation.

Exemple 8.1.
Prenons comme exemple le cas du paragraphe 8.2.4, et prenons une paramétrisation de la conductivité a comme fonction affine de la position :

$$a(x,y) = \alpha_0 + \alpha_1 x + \alpha_2 y. \tag{8.88}$$

La représentation utilisée pour la simulation est une valeur de la conductivité par maille, et il est naturel d'associer cette valeur au centre de gravité de la maille considérée. L'application M est donc ici

$$M(\alpha_0, \alpha_1, \alpha_2) = (\alpha_0 + \alpha_1 x_K + \alpha_2 y_K, \quad K \in \mathcal{T}_h) \tag{8.89}$$

où $\mathcal{T}_h$ est la triangulation utilisée, et (x_K, y_K) sont les coordonnées du centre de gravité de la maille K. La projection est dans ce cas une troncature, visant à assurer que le paramètre a n'est ni trop grand ni trop petit. Elle est définie par

$$P(a_h) = \begin{cases} a_{\min} & \text{si } a_h \leq a_{\min} \\ a_h & \text{si } a_{\min} \leq a_h \leq a_{\max} \\ a_{\max} & \text{si } a_{\max} \leq a_h. \end{cases} \tag{8.90}$$

où $a_{\min}$ et $a_{\max}$ sont des bornes ń convenables ż.

En posant $b_h = P(a_h)$, le calcul des gradients se fait dans ce cas particulier par :

$$\frac{\partial J}{\partial b_K} = \begin{cases} \dfrac{\partial J}{\partial a_K} & \text{si } a_{\min} \leq a_h \leq a_{\max} \\ 0 & \text{sinon.} \end{cases} \tag{8.91}$$

puis par

$$\begin{cases} \dfrac{\partial j}{\partial \alpha_0} = \displaystyle\sum_{K \in \mathcal{T}_h} \frac{\partial J}{\partial b_K} \\ \dfrac{\partial j}{\partial \alpha_1} = \displaystyle\sum_{K \in \mathcal{T}_h} x_K \frac{\partial J}{\partial b_K} \\ \dfrac{\partial j}{\partial \alpha_0} = \displaystyle\sum_{K \in \mathcal{T}_h} y_K \frac{\partial J}{\partial b_K} \end{cases} \tag{8.92}$$

Nous pouvons maintenant décrire la structure générale du code d'optimisation sous la forme d'un bloc diagramme (figure 8.1), suivant Chavent [16]. Bien entendu, la boite correspondant à ń modélisation ż est celle où va la plus grande partie de l'effort, tant en développement qu'en calcul.

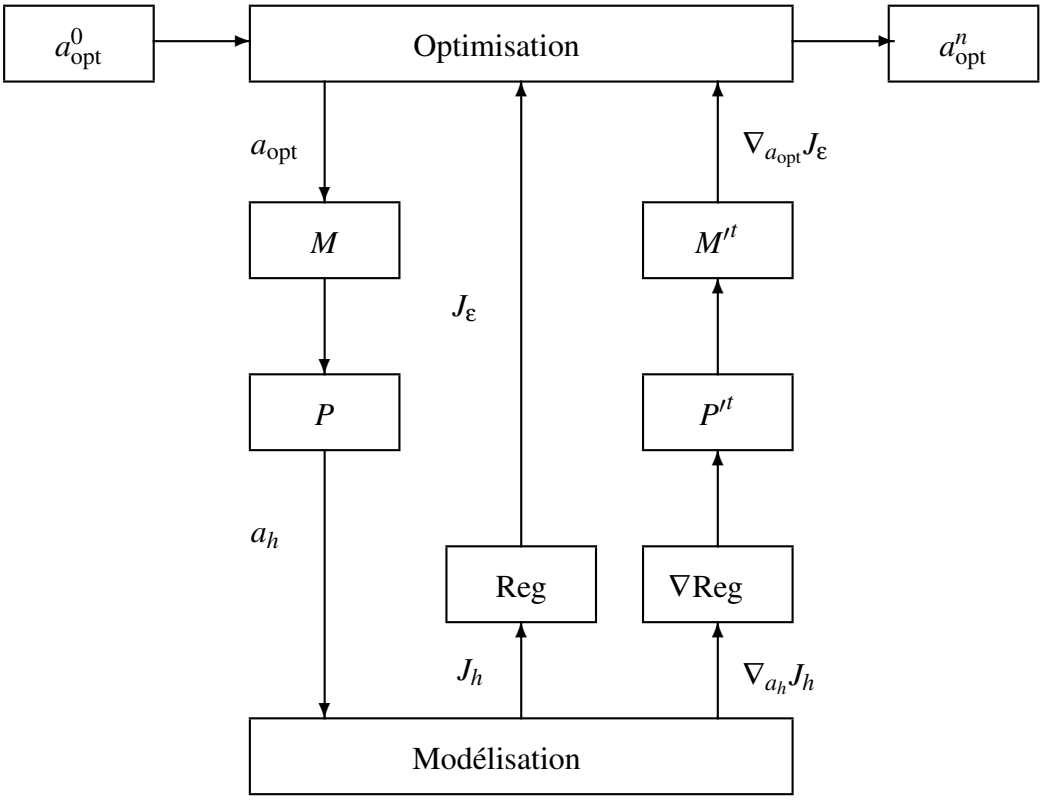

FIGURE 8.1 – Organisation du code d'inversion

Annexe A

Rappels et compléments d'analyse fonctionelle

Nous rappelons dans cet appendice les principaux résultats d'analyse fonctionelle dont nous aurons besoin, ainsi que des compléments concernant les opérateurs dans les espaces de Hilbert.

Nous ne donnerons que peu de démonstrations, renvoyant le lecteur aux (nombreux) ouvrages sur le sujet. Citons particulièrement [12]. Les ouvrages [43], [7], [44] et [31] contiennent également des introductions orientées vers les applications aux problèmes inverses.

Pour simplifier, nous ne considererons que des espaces vectoriels sur $\mathbf{R}$. De plus nous ne considérerons que des espaces *séparables*, c'est-à-dire contenant une partie dénombrable et dense.

A.1 Espaces de Hilbert

Commençons par rappeler quelques définitions.

A.1.1 Définitions et exemples

Définition A.1. Soit E un espace vectoriel sur $\mathbf{R}$. Une *norme* sur E est une application de E dans $\mathbf{R}$, possédant les propriétés suivantes :

- $\forall x \in E,\ \|x\|_E \geq 0$ et $\|x\|_E = 0 \Rightarrow x = 0$;
- $\forall x \in E, \forall \alpha \in \mathbf{R},\ \|\alpha x\|_E = |\alpha|\,\|x\|_E$;
- $\forall (x,y) \in E^2,\ \|x+y\|_E \leq \|x\|_E + \|y\|_E$.

Exemple A.1.
Dans le cas où E est de dimension n (nous l'identifiions alors à $\mathbf{R}^n$), les normes suivantes sont les plus utilisées :

- $\|x\|_1 = \sum_{i=1}^{n} |x_i|$;
- $\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$;
- $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$;

Définition A.2. Soit E un espace vectoriel sur $\mathbf{R}$. Un *produit scalaire* sur E est une application de $E \times E$ dans R, notée $(.,.)$, possédant les propriétés suivantes :

– $\forall (x,y,z) \in E^3, \forall (\alpha,\beta) \in \mathbf{R}^2,\ (\alpha x + \beta y,\, z) = \alpha(x,z) + \beta(y,z)$;
– $\forall (x,y) \in E^2,\ (x,y) = (y,x)$;
– $\forall x \in E, (x,x) \geq 0$,
– $(x,x) = 0 \Rightarrow x = 0$.

Un espace vectoriel muni d'un produit scalaire est appelé un espace *préhilbertien*.

Exemple A.2.
Sur $\mathbf{R}^n$, le produit scalaire euclidien usuel est :

$$(x,y) = \sum_{i=1}^{n} x_i\, y_i \tag{A.1}$$

Exemple A.3.
Soit Ω un ouvert de $\mathbf{R}^n$. L'espace vectoriel des fonctions de carré integrable sur Ω est :

$$L^2 9\Omega) = \{f : \Omega \mapsto \mathbf{R}, \int_\Omega |f(x)|^2\ dx \leq \infty\} \tag{A.2}$$

est un espace préhilbertien si on le munit du produit scalaire :

$$(f,g) = \int_\Omega f(x)\, g(x)\, dx \tag{A.3}$$

Un produit scalaire sur E définit une norme sur E par la formule suivante :

$$\|x\|_E = \sqrt{(x,x)}. \tag{A.4}$$

Parmi les trois normes de l'exemple A.1, seule la seconde provient d'un produit scalaire (celui de l'exemple A.2).

Définition A.3. Un *espace de Hilbert* est un espace vectoriel muni d'un produit scalaire, et qui est complet pour la norme associée à ce produit scalaire.

Exemple A.4.
L'espace vectoriel $\mathbf{R}^n$, muni du produit scalaire euclidien usuel, est un espace de Hilbert.

Le résultat suivant est fondamental.

Proposition A.1. *L'espace vectoriel $L^2(\Omega)$, muni du produit scalaire défini en* (A.3)*, est un espace de Hilbert.*

Exemple A.5 (Espace de Sobolev)**.**
Plaçons nous pour simplifier en dimension 1, sur l'intervalle $[0,1]$. L'espace de Sobolev d'ordre 1 est

$$H^1(0,1) = \Big\{ u \in L^2(0,1), \exists u_1 \in L^2(0,1) \text{ tel que } \forall \varphi \in \mathbf{C}^1_c(0,1), \int_0^1 u(t)\, \varphi'(t)\ dt = -\int_0^1 u_1(t)\varphi(t)\ dt \Big\} \tag{A.5}$$

(où $C^1_c(0,1)$ désigne l'espace des fonctions continûment dérivables, à support compact dans $[0,1]$). Cette définition est équivalente à celle, plus usuelle utilisant la théorie des distributions. Pour $u \in H^1(0,1)$, on note $u' = u_1$.

On démontre (voir [12]) que $H^1(0,1)$ est un espace de Hilbert si on le munit du produit scalaire :

$$(u,v)_{H^1} = \int_0^1 u(t)\,v(t)\,dt + \int_0^1 u'(t)\,v'(t)\,dt. \tag{A.6}$$

Dans les applications, on a souvent besoin du sous-espace de H^1 correspondant aux fonctions ń nulles au bord ż (au sens de la trace). Ce sous-espace est noté $H_0^1(0,1)$, et on peut le munir du produit scalaire suivant :

$$(u,v)_{H_0^1} = \int_0^1 u'(t)\,v'(t)\,dt. \tag{A.7}$$

On démontre (c'est une conséquence de l'inégalité de Poincaré, voir toujours [12]) que la norme correspondante est équivalente à la norme induite par celle de l'espace H^1.

A.1.2 Propriétés des espace de Hilbert

Proposition A.2 (Inégalité de Cauchy-Schwarz)**.** *Pour tous $(x,y) \in E^2$, on a l'inégalité :*

$$|(x,y)|^2 \leq \|x\|_E^2\,\|y\|_E^2\,. \tag{A.8}$$

L'égalité n'a lieu que si x et y sont proportionnels.

Proposition A.3 (Identité du parallelogramme)**.** *Pour tous $(x,y) \in E^2$, on a l'idenntité :*

$$\|x+y\|_E^2 + \|x-y\|_E^2 = 2\left(\|x\|_E^2 + \|y\|_E^2\right) \tag{A.9}$$

Le résultat suivant est l'un des plus importants de la théorie.

Théorème A.1 (de projection)**.** *Soit F un sous-ensemble fermé, convexe de E, et $z \in E$ donné. Il existe un unique élément de $x_0 \in K$ tel que :*

$$\|z-x_0\|_E = \inf y - x,\ x \in F. \tag{A.10}$$

Le point x_0 est caractérisé par la condition :

$$x_0 \in F \ \ et \ \ (z-x_0, x-x_0) \leq 0,\ \forall x \in F \tag{A.11}$$

Le point x_0 mis en évidence au théorème A.1 s'appelle la *projection* de z sur F. Dans le cas où F est un sous-espace vectoriel, on peut préciser ce résultat :

Corollaire A.1. *Soit F un sous-espace vectoriel fermé de E, et soit $z \in E$. La projection de z sur F est caracterisée par :*

$$x_0 \in F \ \ et \ \ (z-x_0, x) = 0,\ \forall x \in F \tag{A.12}$$

Dans un espace de Hilbert, on dit que deux vecteurs sont *orthogonaux* si leur produit scalaire est nul. L'orthogonal d'un sous-espace vectoriel F est :

$$F^\perp = \{x \in E,\ (x,y) = 0, \forall y \in F\}\,.$$

Une conséquence des résultats précédents est :

Corollaire A.2. *Soit F un sous-espace vectoriel de E (non nécessairement fermé). On a*

$$F^\perp + \overline{F} = E \tag{A.13}$$

A.1.3 Bases hilbertiennes

Définition A.4. Une *base hilbertienne* d'un espace de Hilbert E est une suite $(e_n)_{n\in\mathbf{N}^*}$ telle que :
- $\|e_n\|_E = 1, \forall n,$ et $(e_n, e_m) = 0, \forall n \neq m$;
- l'espace vectoriel engendré par les (e_n) est dense dans E.

Précisons la deuxième condition : soit $F_n = \text{vect}\{e_1, \ldots, e_n\}$. Les sous-espaces F_n sont emboités ($F_n \subset F_m$ pour $n \leq m$), donc $F = \cup_{n\in N^*} F_n$ est un sous-espace vectoriel. La seconde condition de la définition exprime que ce sous-espace est dense dans E, c'est-à-dire que tout élément de E peut être approché arbitrairement par un élément de F.

On démontre alors que tout espace vectoriel (séparable) admet une base hilbertienne. Étant donné une base hilbertienne $(e_n)_{n\in\mathbf{N}}$ de E, tout élément de E s'écrit :

$$x = \sum_{n=1}^{\infty} (x, e_n)\, e_n, \tag{A.14}$$

avec (c'est l'égalité de Bessel–Parseval) :

$$\|x\|_E^2 = \sum_{n=1}^{\infty} |(x, e_n)|^2 \tag{A.15}$$

Un tel développement est unique, c'est-à-dire que si on a un développement

$$x = \sum_{n=1}^{\infty} x_n e_n$$

avec $\sum_{n=1}^{\infty} |x_n|^2 < \infty$, alors $x_n = (x, e_n)$. Notons toutefois qu'une base hilbertienne *n'est pas* une base algébrique, puisque ce développement n'est pas une combinaison linéaire finie.

On sait construire explicitement des bases hilbertiennes pour certains espaces L^2. Bien évidemment, une base orthogonale d'un espace vectoriel de dimension finie est une base hilbertienne.

Exemple A.6.
Les deux suites de fonctions

$$\left(\sqrt{\frac{2}{\pi}}\sin nx\right)_{n\geq 1}, \text{ ou } \left(\sqrt{\frac{2}{\pi}}\cos nx\right)_{n\geq 0}$$

sont des bases hilbertiennes de $L^2(0,\pi)$. Dans ce cas, le développement obtenu en (A.14) s'identifie à un développement en série de Fourier (après prolongement par imparité eet périodicité).

A.2 Opérateurs linéaires dans les espaces de Hilbert

L'analyse fonctionelle fait interagir la topologie et l'algèbre linéaire. Ainsi, sur une espace de Hilbert, il sera naturel d'étudier les applications qui respectent à la fois la structure d'espace vectoreil (les applications linéaires) et la structure hilbertienne (les applications continues).

A.2.1 propriétés générales

Définition A.5. Un opérateur (linéaire, continu) A d'un espace de Hilbert E dasn un espace de Hilbert F est une application linéaire continue de E dans F, c'est-à-dire qui vérifie :

- $\forall u \in E, Au \in F$;
- $\forall (u,v) \in E \times E, \forall (\alpha,\beta) \in \mathbf{R}^2, A(\alpha u + \beta v) = \alpha Au + \beta Av$;
- $\exists M > 0, \forall u \in E, \|Au\|_F \leq M \|u\|_E$.

Le plus petit nombre M qui vérifie le 3eme point ci-dessus s'apelle la *norme* de l'opérateur A :

$$\|A\| = \sup_{u \in E} \frac{\|Au\|_F}{\|u\|_E}. \tag{A.16}$$

Il s'agit de la même notion que la norme matricielle utilisée en algèbre linéaire (voir par exemple [28] ou [46]).

Rappelons les deux espaces fondamentaux associés à un opérateur linéaire :

- Le noyau de A est le sous-espace de E : $\mathrm{Ker}A = \{u \in E, Au = 0\}$;
- L'image de A est le sous-espace de F : $\mathrm{Im}A = \{v \in F, \exists u \in E, Au = v\}$.

Remarquons que si $\mathrm{Ker}A$ est touours fermé, en tant qu'image reciproque du sous espace fermé 0 de F, alors que $\Im A$ peut ne pas être fermé (voir la discussion sur ce point au théorème 4.1).

Un opérateur est dit de *rang fini* si et seulement si son image est un sous-espace vectoriel de dimension finie de F.

Le théorème suivant est l'un des résultats fondamentaux de la théorie des opérateurs linéaires.

Théorème A.2 (de l'application ouverte)**.** *Soit A un opérateur linéaire de E dans F. L'image par A d'un ouvert de E est un ouvert de F.*

En particulier l'inverse d'un opérateur linéaire continu et bijectif est continu.

Dans le cas où l'espace d'arrivée F est le corps des scalaires, on parle de *forme linéaire*. L'espace vectoriel des formes linéaires continues s'apelle l'espace dual de E, et se note E'. Dans le cas d'un espace de Hilbert, le dual s'identifie de façon canonique à l'espace lui-même.

Théorème A.3 (de Riesz)**.** *Soit L une forme linéaire continue sur E. Il existe un unique vecteur $x_0 \in E$ tel que*

$$L(x) = (x_0, x),\ \forall x \in E. \tag{A.17}$$

A.2.2 Adjoint d'un opérateur

Théorème A.4. *Soit A un opérateur linéaire continu de E dans F. Il existe un unique opérateur de F dans E, noté A^*, tel que :*

$$\forall u \in E, \forall v \in F,\ (Au, v) = (u, A^* v). \tag{A.18}$$

*Cet opérateur est appelé l'*adjoint *de A. Il vérifie de plus : $(A^*)^* = A$ et $\|A^*\| = \|A\|$.*

Preuve. Fixons tout d'abord $v \in F$. L'application

$$\begin{array}{c} E \mapsto \mathbf{R} \\ u \to (Au, v) \end{array}$$

est linéaire et continue, puisque $|(Au,v)| \leq \|A\| \, \|u\|_E \, \|v\|_F$. D'après le théorème de Riesz, il existe un élément unique de E, noté pour l'instant v^* tel que

$$(Au,v) = (u,v^*),\ \forall u \in E.$$

Il reste à voir que l'application $v \to v^*$ est bien définie, linéaire et continue. On pourra alors noter $A^*v = v^*$, et A^* aura bien les propriétés annoncées.

Le fait que notre application est bien définie résulte de l'unicité dans le théorème de Riesz. Il en est d'ailleurs de même pour la linéarité (en utilisant la linéarité du produit scalaire). On note au passage que la définition est symétrique en A et A^*, de sorte que $(A^*)^* = A$.

Enfin, la continuité résulte de l'inégalité de Cauchy-Schwarz :

$$\|A^*v\|_E^2 = (A^*v, A^*v) = (AA^*v, v) \leq \|A\| \, \|A^*v\|_E \, \|v\|_F,$$

qui montre de plus que $\|A^*\| \leq \|A\|$. Comme nous venons de voir que A et A^* jouent le même rôle, on a finalement $\|A^*\| = \|A\|$.

□

Remarque A.1. En dimension finie, en identifiant l'application linéaire A a sa matrice dans des bases orthogonales de $\mathbf{R}^n$ et $\mathbf{R}^p$, on voit que la matrice de l'opérateur adjoint n'est autre que la matrice transposée de A (prendre $u = e_i$, $i = 1,\ldots,n$ et $v = f_j$, $j = 1,\ldots,p$ des éléments des bases considérées).

La proposition suivante rassemble quelques propriétés simples de l'adjoint.

Proposition A.4. *Soient A et B sont deux opérateurs linéaires, α et β deux scalaires.*
- *Linéarité :* $(\alpha A + \beta B)^* = \alpha A^* + \beta B^*$.
- *Composition :* $(AB)^* = B^*A^*$.

Il existe des relations remarquables entre le noyau et l'image d'un opérateur et ceux de son adjoint.

Proposition A.5. *On a les relations suivantes (ou $\overline{X}$ indique l'adhérence de l'ensemble X) :*
- $\mathrm{Ker}A^* = (\mathrm{Im}A)^\perp$;
- $(\mathrm{Ker}A)^\perp = \overline{\mathrm{Im}A^*}$.

Définition A.6. Un opérateur dans E est dit auto-adjoint si et seulement si :

$$\forall (u,v) \in E \times E,\ (Au,v) = (u,Av) \tag{A.19}$$

Remarque A.2. En dimension finie, les opérateurs auto-adjoints sont ceux qui ont une matrice symétrique.

A.2.3 Opérateurs compacts

Définition A.7. Soit $A \in \mathcal{L}(E,F)$. On dit que A est un opérateur compact si et seulement si l'image de toute partie bornée de E est relativement compacte dans F.

Remarque A.3. Cette condition veut dire que si $B \subset E$ est borné, $\overline{A(B)}$ (l'adhérence de $A(B)$) est compacte dans F.

Exemple A.7.
- Tout opérateur de rang fini est compact.
- L'injection canonique de $H_0^1(0,1) \to L^2(0,1)$ est comapcte (plus généralement, on peut remplacer $]0,1[$ par un ouvert borné régulier de $\mathbf{R}^n$).

– Une classe importante d'opérateurs compacts est fournie par certains opérateurs intégraux, étudiés au chapitre 3.

Commençons par quelques propriétés de base des opérateurs compacts.

Proposition A.6. *Soit E, F, G trois espaces de Hilbert.*

i) *L'ensemble des opérateurs compacts de E dans F est un sous-espace vectoriel de $\mathcal{L}(E,F)$.*

ii) *Si $A_1 \in \mathcal{L}(E,F)$ est compact et $A_2 \in \mathcal{L}(F,G)$, alors $A_2A_1 \in \mathcal{L}(E,G)$ est compact.*

iii) *Si $A_1 \in \mathcal{L}(E,F)$ et $A_2 \in \mathcal{L}(F,G)$ est compact, alors $A_2A_1 \in \mathcal{L}(E,G)$ est compact.*

iv) *Si $A \in \mathcal{L}(E,F)$ est compact, $A^* \in \mathcal{L}(F,E)$ est aussi compact.*

v) *Soit $(A_n)_{n\in\mathbf{N}}$ une suite d'opérateurs compacts de E dans F. Si A_n converge vers A dans $\mathcal{L}(E,f)$, c'est-à-dire si*

$$\|A_n - A\| = \sup_{u\neq 0} \frac{\|A_n u - Au\|_F}{\|u\|_E} \underset{n\to\infty}{\longrightarrow} 0,$$

alors A est compact.

En d'autres termes, les opérateurs compacts forment un sous-espace vectoriel fermé de $\mathcal{L}(E,f)$.

Le théorème suivant fournit (dans le cas des espaces de Hilbert) une caracterisation à la fois utile et plus proche de l'intuition.

Théorème A.5. *Un opérateur de E dans F est compact si et seulement si il est limite d'une suite d'opérateurs de rang fini.*

Ce théorème signifie que les opérateurs compacts sont ceux qui ń
ressemblent ż le plus aux opérateurs de dimension finie usuels. Signalons que ce résultat n'est plus valable si E et F sont des espaces de Banach. Il existe par contre une différence, qui sera fondamentale pour l'étude des problèmes mal posés.

Proposition A.7. *Si E n'est* pas *de dimension finie, alors l'identité $E \to E$ n'est jamais compacte.*

Remarque A.4. Nous ne donnons pas la démonstration (voir par exemple[12]), qui est une conséquence de ce que la boule unité d'un espace vectoriel normé de dimension finie n'est pas compacte.

Corollaire A.3. *Soit A un opérateur compact de E dans F, où E et F sont deux espaces de Hilbert qui ne sont pas de dimension finie. Alors A n'est jamais inversible dans $\mathcal{L}(E,F)$.*

Preuve. Si A est inversible, son inverse A^{-1} vérifie :

$$AA^{-1} = I.$$

Comme A est compact, et que l'identité ne peut l'être d'après la proposition A.7, nous avons une contradiction. □

Remarque A.5. Dans le corollaire précédent, l'inverse (algébrique) de A peut ou non exister (A peut ou non être injectif), mais s'il existe, il ne sera pas continu. Ceci est lié au caractère non fermé de l'image de A, et est développé au chapitre 6.

Pour conclure, citons sans démonstration une version ń
abstraite ż de l'alternative de Fredholm. Ce résultat concerne les équations du type

$$(I-A)u = f \tag{A.20}$$

où A es tun opérateur compact dans E.

Théorème A.6. *Soit A un opérateur compact dans un epace de Hilbert E.*
- *Le noyau* $\mathrm{Ker}\,(I-A)$ *est de dimension finie, et l'image* $\mathrm{Im}\,(I-A)$ *est fermée dans E.*
- *Si $I-A$ est injectif, il est aussi surjectif, et alors l'inverse $(I-A)^{-1}$ est continu. Si $I-A$ n'est pas injectif, l'équation* (A.20) *a une solution si et seulement si $f \in \mathrm{Ker}\,(I-A)^{\perp}$.*

Remarque A.6.
- Ce théorème étend aux opérateurs de la forme ń
  $I+$compactż le résultat analogue bien connu en dimension finie (pour tous les systèmes d'équations linéaires).
- Le second point du théorème veut dire que l'équation (A.20) n'a de solution que si le second membre satisfait des conditions d'orthogonalité au noyau (de dimension finie).

A.3 Décomposition spectrale des opérateurs auto-adjoints compacts

Dans toute cette section, A désigne un opérateur auto-adjoint compact dans E.

Définition A.8. Le *spectre* de A est l'ensemble

$$\sigma(A) = \{\lambda \in \mathbf{C}, A - \lambda I \text{ n'est pas inversible dans } \mathcal{L}(E)\}.$$

Définition A.9. Un nombre $\lambda \in \mathbf{C}$ est une *valeur propre* si et seulement si $A - \lambda I$ n'est pas injectif.

Rappelons que, a priori, si $\lambda \in \sigma(A)$, trois cas peuvent se produire pour l'opérateur $A - \lambda I$:
- il peut ne pas être injectif, et dans ce cas λ est une valeur propre ;
- il peut ne pas être surjectif ;
- il peut être bijectif, mais l'inverse n'est pas continu.

On démontre qu'en fait, si A est commpact et $\lambda \neq 0$, les deux dernières alternatives ne peuvent pas se produire.

Pour simplifier l'énoncé du théorème suivant, nous ferons l'hypothèse que A n'est pas de rang fini (si c'était le cas, les valeurs propres seraient en nombre fini, et 0 pourrait ne pas être valeur propre).

Proposition A.8. *Notons Λ l'ensemble des valeurs propres de A.*

i) $\sigma(A) = \{0\} \cup \Lambda$ *;*

ii) *Toute valeur propre non-nulle est de multiplicité finie ;*

iii) *A a au plus une infinité dénombrable de valeurs propres, dont le seul point d'accumulation possible est 0 ;*

iv) *Les valeurs propres de A sont réelles et des vecteurs propres correspondant à des valeurs propres distinctes sont orthogonaux ;*

v) *L'un des nombres $\pm\|A\|$ est une valeur propre de A.*

Les valeurs propres non-nulles d'un opérateur auto-adjoint compact peuvent donc être rangées en une suite qui tend vers 0.

Nous pouvons maintenant énoncer le résultat le plus important de cette section.

Théorème A.7. *Notons* $(\lambda_n)_{n\in\mathbf{N}}$ *les valeurs propres de A, avec* $\lim_{n\to\infty}\lambda_n = 0$*. Il existe une base hilbertienne* $(e_n)_{n\in\mathbf{N}}$ *de* $(KerA)^{\perp}$ *telle que*

$$\forall x \in E,\ x = x_0 + \sum_{n\in\mathbf{N}} (x,e_n)e_n,\ Ax = \sum_{n\in\mathbf{N}} \lambda_n (x,e_n)e_n, \tag{A.21}$$

où $x_0 \in \mathrm{Ker}A$.

Bien entendu, les séries dans (A.21) sont à prendre au sens de la norme de E. Ici encore, si A est de rang fini, les sommes ci-dessus sont en fait des sommes finies.

Il est clair que tout opérateur donné par une formule comme (A.21) est compact. Une définition ń opératoire ż des opérateurs compacts est : un opérateur donné par l'équation (A.21), où la suite $(\lambda_n)_{n\in\mathbf{N}}$ est une suite réelle qui tend vers 0.

Ce résultat est l'analogue de la diagonalisation des matrices symétriques dans une base orthonormée. Dans la base hilbertienne donnée par le théorème, l'action de A est diagonale.

Bibliographie

[1] K. Aki and P. Richards. *Quantitative Seismology : Theory and Methods*. Freeman, 1980.

[2] V. I. Arnold. *Équations différentielles ordinaires*. Éditions de Moscou, 19xx.

[3] A. Bamberger, G. Chavent, C. Hemon, and P. Lailly. Inversion of normal incidence seismograms. *Geophysics*, 47(5) :757–770, 1982.

[4] A. Bamberger, G. Chavent, and P. Lailly. Une application de la théorie du contrôle à un problème inverse de sismique. *Annales de géophysique*, 33 :183–199, 1977.

[5] A. Bamberger, G. Chavent, and P. Lailly. About the stability of the inverse problem in 1D wave equations – applications to the interpretation of seismic profiles. *J. Appl. Math. Optim.*, 5 :1–47, 1979.

[6] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhäuser-Verlag, Zürich, 1989.

[7] J. Baumeister. *Stable Solution of Inverse Problems*. Vieweg, Braunschweig, 1987.

[8] J. Bear and A. Verruijt. *Modeling Groundwater Flow and Pollution*, volume 2 of *Theory and Applications of Transport in Porous Media*. Kluwer, 1987.

[9] Å. Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[10] N. Bleistein, J. K. Cohen, and J. J. W. Stockwell. *Mathematics of Multidimensional Seismic Imaging, Migration and Inversion*. Number 13 in Interdisciplinary Applied Mathematics. Springer, 2000.

[11] J. F. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation numérique – Aspects théoriques et pratiques*. Number 27 in Mathématiques et Applications. Springer, Berlin, 1997.

[12] H. Brézis. *Analyse fonctionnelle – Théorie et applications*. Masson, Paris, 1983.

[13] J. R. Cannon. *The one-dimensional heat equation*. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1984. With a foreword by Felix E. Browder.

[14] G. Chavent. Problèmes inverses. notes de cours de DEA, Université Paris Dauphine.

[15] G. Chavent. Identification of distributed parameter systems : About the output least squares method, its implementation and identifiability. In *Proc. 5th IFAC Symposium on Identification and System Parameter Estimation*, pages 85–97. Pergamon Press, 1979.

[16] G. Chavent. On the theory and practice of non-linear least squares. *Advances in Water Resources*, 14(2) :55–63, 1991.

[17] F. Clément, G. Chavent, and S. Gòmez. Migration-based traveltime waveform inversion of 2-D simple structures : A synthetic example. *Geophysics*, 66(3) :845–860, 2001.

[18] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer-Verlag, New-York, 1992.

[19] R. Dautray and J. L. Lions, editors. *Analyse mathématique et calcul numérique pour les sciences et les techniques*. Masson, 1982.

[20] G. de Marsily. *Hydrogéologie quantitative*. Masson, 1981.

[21] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.

[22] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, 1996.

[23] T. A. Driscoll. Eigenmodes of isospectral drums. *SIAM Rev.*, 39(1) :1–17, 1997.

[24] H. W. Engl. Regularization methods for the stable solution of inverse problems. *Surveys Math. Indust.*, 3 :71–143, 1993.

[25] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1996.

[26] J.-C. Gilbert. Optimisation : Théorie et algorithmes. Cours à l'ENSTA, 1998. disponible sur l' Internet à `http://www-rocq.inria.fr/ gilbert/ensta/optim.html`.

[27] M. S. Gockenbach, M. J. Petro, and W. W. Symes. C++ classes for linking optimization with complex simulations. *ACM Transactions on Mathematical Software*, 25(2) :191–212, 1999.

[28] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 3ème edition, 1996.

[29] R. Gorenflo and S. Vessella. *Abel Integral Equations. Analysis and Applications*. Number 1461 in Lecture Notes in Mathematics. Springer, 1991.

[30] A. Griewank. Some bounds on the complexity of gradients, jacobians, and hessians. Technical Report MCS-P355-0393, Mathematics and Computer Science Division, Argonne National Laboratory, 1993.

[31] C. W. Groetsch. *Inverse Problems in the Mathematical Sciences*. Vieweg, Wiesbaden, 1993.

[32] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University Press, 1923.

[33] E. Hairer, S. Norsett, and G. Wanner. *Solving Ordinary Differential Equations : 1 : Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer Verlag, 1987.

[34] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations : 2 : Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer Verlag, 1991.

[35] M. Hanke. *Conjugate Gradient Type Methods for Ill-Posed Problems*, volume 327 of *Pitman Research Notes in Mathematics*. Longman, Harlow, 1995.

[36] P. C. Hansen. The discrete Picard condition for discrete ill-posed problems. *BIT*, 30 :658–672, 1990.

[37] P. C. Hansen. *Rank-deficient and discrete ill-posed problems : numerical aspects of linear inversion*. SIAM, Philadelphia, 1998.

[38] G. T. Herman, editor. *Image Reconstruction from Projections : the Fundamentals of Computerized Tomography*. Academic Press, New York, 1980.

[39] V. Isakov. *Inverse Problems for Partial Differential Equations*. Number 127 in Applied Mathematical Sciences. Springer, New-York, 1998.

[40] M. Kac. Can one hear the shape of a drum ? *Amer. Math. Monthly*, 73 :1–23, 1966.

[41] J. B. Keller. Inverse problems. *Amer. Math. Monthly*, 83 :107–118, 1976.

[42] C. T. Kelley. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics. SIAM, Philadelphia, 1999.

[43] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Number 120 in Applied Mathematical Sciences. Springer-Verlag, New-York, 1996.

[44] R. Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences*. Springer, 1989.

[45] L. Landweber. An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.*, 73 :615–624, 1951.

[46] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Masson, Paris, 1986. 2 volumes.

[47] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM, Philadelphie, 1995. Édition originale par Prentice Hall, 1974.

[48] A. K. Louis. Medical imaging, state of the art and future developments. *Inverse Problems*, 59 :277–294, 1992.

[49] B. Lucquin and O. Pironneau. *Introduction au calcul scientifique*. Masson, Paris, 1996.

[50] J. J. Moré and S. J. Wright, editors. *Optimization Software Guide*. Number 14 in Frontiers in Applied Mathematics. SIAM, Philadelphie, 1993.

[51] V. A. Morozov. *Methods for Solving Incorrectly Posed problems*. Springer-Verlag, New-York, 1984.

[52] R. Mosé. Habilitation à diriger des recherches, Université Louis Pasteur de Strasbourg, 1998.

[53] F. Natterer. *The Mathematics of Computerized Tomography*. Wiley, New-York, 1986.

[54] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New-York, 1999.

[55] P.-A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques appliquées pour la maîtrise. Masson, Paris, 1983.

[56] J. M. Restrepo, G. K. Leaf, and A. Griewank. Circumventing storage limitations in variational data assimilation studies. *SIAM J. Sci. Comput.*, 19(5) :1586–1605 (electronic), 1998.

[57] P. Siegel. *Transfert de masse en milieux poreux complexes : modélisation et estimation de paramètres par éléments finis mixtes hybrides*. Thèse de doctorat, Université Louis Pasteur de Strasbourg, 1995.

[58] G. W. Stewart. *Afternotes goes to Graduate School – Lectures on Advanced Numerical Analysis*. SIAM, 1997.

[59] N.-Z. Sun. *Inverse Problems in Groundwater Modeling*. Number 6 in Theory and Application of Transport in Porous Media. Kluwer, 1994.

[60] W. W. Symes. A differential semblance criterion for inversion of multioffset seismic reflection data. *J. Geoph. Res.*, 98 :2061–2073, 1993.

[61] W. W. Symes and M. Kern. Inversion of reflection seismograms by differential semblance analysis : Algorithm structure and synthetic examples. *Geophysical Prospecting*, 42 :565–614, 1994.

[62] V. Torczon. On the convergence of multidimensional search algorithms. *SIAM J. Optim.*, 1 :123–145, 1991.

[63] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[64] P. van Laarhoven and E. Aarts. *Simulated Annealing, Theory and Practice*. Kluwer, Dordrecht, 1987.