

Cours de M2 2017
Méthodes de problèmes inverses et
applications en dynamique des populations

M. Doumic¹ P. Moireau ²

15 décembre 2017

1. Inria, UPMC Univ Paris 06, Lab. J.L. Lions UMR CNRS 7598, Sorbonne Universités
2. Inria – LMS, Ecole Polytechnique, CNRS – Université Paris-Saclay

Table des matières

1 Quelques problèmes inverses en dynamique des populations	13
1.1 Très brève introduction à la dynamique des populations	13
1.1.1 Quelques exemples de populations, et de mesures expérimentales.	14
1.1.2 Comment obtenir une description mathématique de la population ?	18
1.1.3 Quelques équations fondamentales de la dynamique des populations	19
1.1.4 Comportement en temps long des équations linéaires	21
1.1.5 Estimation en dynamique des populations	22
1.2 Problèmes inverses pour la polymérisation des protéines	23
1.2.1 Modèle discret	23
1.2.2 Du modèle discret au modèle continu	25
1.2.3 Estimation en polymérisation des protéines	28
1.3 Exercices	31
2 Introduction à la théorie des problèmes inverses linéaires	33
2.1 Cadre fonctionnel	33
2.2 Pseudo-inverse de Moore-Penrose	35
2.3 Régularisation d'un problème mal posé	39
3 Décomposition d'un opérateur compact en valeurs singulières	45
3.1 Décomposition en valeurs singulières : Définition et premières propriétés . .	45
3.2 Suites régularisantes avec la décomposition en valeurs singulières	48
3.3 Espaces sources et optimalité avec la décomposition en valeurs singulières	50
3.4 Implémentation numérique	54
3.4.1 Cas où l'on connaît déjà la décomposition en valeurs singulières . . .	54
3.4.2 décomposition en valeurs singulières en dimension finie	55
3.4.3 Calcul numérique de la décomposition en valeurs singulières	55
3.4.4 Application en traitement des images	56

4 La régularisation de Tikhonov	57
4.1 Généralités sur la régularisation de Tikhonov	57
4.1.1 Résultats directs	60
4.1.2 Analyse de la méthode de Tikhonov grâce à la décomposition en valeurs singulières	64
4.2 Exemple de l'inversion de l'opérateur intégral	68
4.2.1 Méthode de Tikhonov classique	68
4.2.2 Méthode de Tikhonov généralisée	72
4.3 Résolution algorithmique	74
4.3.1 Exemple de l'opérateur intégral	74
4.3.2 Descente de gradient	74
4.3.3 Méthode de Landweber / descente de gradient à pas constant	76
5 L'assimilation de données	79
5.1 Retour en dimension finie	79
5.1.1 Reformulation du cas statique linéaire	79
5.1.2 Vers la multiplication des observations	82
5.1.3 Prise en compte d'une dynamique discrète	85
5.1.4 Prise en compte d'une dynamique continue	90
5.1.5 Prise en compte d'une erreur de modèle	92
5.1.6 Reconstruction de la condition initiale	99
5.1.7 Analyse de stabilité	100
5.2 Vers les systèmes de dimension infinie	100
5.2.1 Définition du modèle	100
5.2.2 L'approche variationnelle	102
5.2.3 L'approche par filtrage optimal	103
5.2.4 Exemples d'application	104
5.3 Les observateurs de type Luenberger	110
6 Estimation de la densité en statistique	115
6.1 Un point de vue déterministe : suites régularisantes	115
6.1.1 Rappel sur les espaces de Sobolev	116
6.1.2 Régularisation par convolution avec une suite régularisante	117
6.1.3 Inégalités	117
6.2 Estimateurs à noyau d'une densité	119
7 Retour sur les applications en dynamique des populations	123
7.1 La division cellulaire	123
7.1.1 Interprétation des données	123
7.1.2 Estimation sur le problème en âge : un problème jouet instructif . .	124

TABLE DES MATIÈRES

5

7.1.3 Estimation sur le problème en taille	126
7.2 La polymérisation des protéines	129

Introduction

L'objectif de ce cours est d'introduire la notion de problème inverse et les principales méthodes et difficultés de ce domaine, et de les voir à l'oeuvre dans des applications à la biologie cellulaire contemporaine. Il s'agit autant de donner une culture générale du domaine qu'un début de méthode pour aborder des applications réelles.

Dans ce chapitre, on commence donc par une présentation générale du domaine, puis par exposer les deux problèmes appliqués qui serviront d'exemples archétypiques tout au long du cours :

1. la division cellulaire, qui conduira à introduire plus généralement les équations de population structurées, les liens entre modélisation analytique et probabiliste et la question de l'estimation de la densité en statistique.
2. La polymérisation des protéines, qui conduira à introduire les méthodes d'assimilation de donnés et permettra de faire le lien entre les différentes méthodes d'estimation.

Autant que possible, la théorie générale sera illustrée par l'une ou l'autre de ces deux applications.

Dans chacun de ces deux exemples, on retrouvera la question classique de l'estimation de la dérivée d'une fonction à partir d'une mesure bruitée, question qui à son tour servira d'exemple dans la partie la plus théorique du cours, lors de l'exposé des méthodes de problème inverse dans un cadre fonctionnel général (décomposition en valeurs singulières, régularisation de Tikhonov).

Connaissances requises : analyse fonctionnelle de niveau M1.

Qu'est-ce qu'un problème inverse ?

La terminologie de *problème inverse* est apparue comme un domaine de recherche en soi dans les années 1970, mais le « domaine » lui-même est aussi ancien que les mathématiques : autant que la notion d'inverse d'une fonction, d'un nombre... Vous résolvez vous-même tous les jours, sans le savoir, des problèmes que l'on pourrait qualifier de « problèmes inverses ». C'est sans doute une des raisons pour lesquelles la terminologie est assez

variée : « assimilation de données », « problèmes inverses/ problèmes mal posés », « estimation » accolée de divers qualificatifs (« estimation de paramètres », « estimation non paramétrique », « estimation d'état »), etc.

L'exemple que J. Baumeister et A. Leitaõ citent dans [10] (livre auquel la partie fondamentale de ce cours doit beaucoup, de même qu'au livre de Engl, Hanke et Neubauer [27]) comme le plus ancien problème inverse est le calcul du rayon de la terre par Eratosthène au IIIème siècle av. J.C. Je pense que la notion la plus exacte, plutôt que de considérer le domaine des problèmes inverses comme un tout, est de parler, comme J. Keller [33], de problèmes *inverses l'un de l'autre*. Quel est le problème considéré comme « direct », et celui considéré comme « inverse » ? C'est avant tout une affaire de goût, de « nature », d'habitude, ou d'intuition.

L'exemple à la fois le plus parlant et correspondant à tout un domaine au sein des problèmes inverses est l'exemple d'un problème temporel : le problème considéré comme *direct* est : soit un système S. À partir d'une condition initiale connue, de paramètres du problème connus, peut-on déterminer l'évolution temporelle de S ? À partir de cette formulation du problème « direct », il n'y a pas un mais des problèmes inverses :

- Connaissant d'une façon partielle ou complète l'évolution temporelle de S, puis-je retrouver son état initial ? C'est ce qu'on appelle aussi l'*estimation d'état*.
- Connaissant d'une façon partielle ou complète l'évolution temporelle de S, puis-je retrouver les paramètres du système ? C'est ce qu'on appelle aussi l'*estimation des paramètres*.
- Si ce que l'on souhaite estimer (état ou paramètre) consiste en un nombre fini de nombres réels, on parle d'*estimation paramétrique*. Si les paramètres ou la condition initiale sont non pas un nombre fini de grandeurs mais des fonctions, on parle (surtout dans le domaine des statistiques) d'*estimation non paramétrique*.
- Si ce que l'on souhaite c'est guider l'évolution temporelle de S vers un certain état-cible, on entre dans le domaine du *contrôle*, domaine qui en plusieurs aspects rejoint le domaine des problèmes inverses.
- Si à partir de l'évolution sur S on souhaite retrouver non plus seulement les données du problème, mais évaluer le caractère approprié ou non du modèle lui-même, on entre dans le domaine de la *sélection de modèles*, branche à part entière des statistiques.
- Le terme *assimilation de données* est également très employé, à l'origine en météorologie, pour désigner l'ensemble des méthodes permettant de corriger l'état d'une prévision météo en prenant en compte des observations. Cette partie du domaine a développé son propre vocabulaire que nous tâcherons également d'aborder.

Formulé ainsi, à partir de l'exemple temporel pour lequel il y a bien un sens direct et un sens rétrograde universels, on comprend pourquoi ce domaine des problèmes inverses s'est beaucoup développé ces cinquante dernières années. En effet, avec l'accumulation de données d'une part, et les puissances de calcul d'autre part, il devient possible de tester numériquement, à partir d'un jeu de données sur un problème donné, un grand nombre

d'hypothèses, de paramètres, de modèles, et intuitivement on choisira « le meilleur » comme étant celui qui donne par la simulation un résultat le plus proche des données initiales. En disant cela, je viens de vous formuler d'une façon non rigoureuse le principe même d'une des méthodes (si ce n'est LA méthode) les plus utilisées du domaine : les méthodes par minimisation, dont la méthode dite des « moindres carrés » est la plus répandue. Nous y reviendrons plus en détail.

Pour terminer sur le caractère extrêmement vaste du domaine, il existe 4 journaux mathématiques uniquement dédiés à la famille des problèmes inverses - ce qui n'exclut pas, bien sûr, qu'un grand nombre d'articles publiés dans d'autres revues soient eux aussi liés à la résolution ou à des méthodes de problèmes inverses :

- Inverse Problems, créé en 1985,
- Journal of Inverse and Ill-posed Problems, depuis 1993,
- Inverse Problems in Science and Engineering, depuis 1994,
- Inverse Problems and Imaging, depuis 2007.

Problèmes bien posés, problèmes mal posés

Une fois dit que la notion plus exacte est de parler de problèmes inverses l'un de l'autre plutôt que de problèmes inverses de manière générale, on peut se demander pourquoi ce domaine aurait un intérêt à être étudié comme un champ spécifique de recherche. Qu'y a-t-il de spécifique à un problème qualifié de « inverse » par convention, tradition ou habitude, par rapport à un problème « tout court » ? Et somme toute, qu'est-ce qu'un « problème » ? Je dirais : c'est la formulation en termes mathématiques d'une question, formulation qui permet (toujours en mathématiques) son étude et sa résolution, partielle ou complète.

Revenons donc à la définition d'Hadamard (1865-1963) de ce qu'est un « problème bien posé » [30]. Selon lui, lorsqu'un problème modélise une situation *physique*, il doit comporter trois caractéristiques :

- Une solution existe,
- la solution est unique,
- la solution dépend de façon continue des données dans le cadre d'une topologie raisonnable.

On voit bien ce qui préside l'intuition d'Hadamard, fortement liée à notre exemple d'un problème temporel : une vision déterministe du monde, qui considère qu'une situation donnée ne peut conduire qu'à une seule situation dans l'avenir, et qu'un petit changement dans les données ne peut aboutir à un bouleversement complet du résultat. Mais ce qui est naturel (même si non systématique : cf. l' « effet papillon » et la théorie du chaos) pour un problème « direct » lié à la modélisation d'un problème physique ne l'est plus si l'on considère un problème inverse : si l'existence d'une solution semble raisonnable dans un espace bien choisi, ni l'unicité ni la continuité ne semblent acquis. Nous le verrons à de

nombreuses reprises.

Là réside le point commun entre tous les problèmes inverses : il s'agit, du moins dans leur formulation la plus naturelle, de *problèmes mal posés*. On peut donc souvent décomposer leur résolution en deux parties.

1. La régularisation du problème : il s'agit ici de remplacer le problème initial, mal posé, en un problème dit régularisé, bien posé. Ceci fait, il faut également estimer la distance entre les deux problèmes, et à quelle condition la solution du problème régularisé sera effectivement proche de la solution recherchée. Cette partie, générique, peut donner lieu à des méthodes génériques. Nous en verrons un certain nombre d'entre elles.
2. L'étude du problème spécifique considéré, qu'il soit régularisé ou non : en quoi il est mal ou devient bien posé ; comment le résoudre de façon théorique et numérique. Cette partie est spécifique : il y a donc autant de « problèmes inverses » que de « problèmes directs ». Elle est aussi liée profondément à l'étude du problème direct correspondant : les méthodes peuvent être proches pour le problème direct et pour le problème inverse, et l'analyse de deux problèmes inverses l'un de l'autre leur fournit bien souvent un éclairage réciproque.

En décomposant ainsi le problème, le point de vue est typiquement celui des « problèmes inverses ». En assimilation de données, la partie régularisation est davantage comprise comme un *a priori* sur le problème. Les deux points de vue se rejoignent, comme nous le verrons par la suite.

Exercices

1. Un exemple de base : estimer la gravité.

Vous êtes Galilée, aux environs de 1604, et vous venez d'établir la loi de la gravité sur Terre : si $z(t)$ représente la position verticale au cours du temps d'un objet, sa loi est

$$\frac{d^2z}{dt^2} = -g, \quad z(t=0) = z_0, \quad \frac{dz}{dt}(t=0) = 0.$$

Vous souhaitez évaluer la valeur de la constante g . Or qu'observez-vous ? Le déplacement de votre objet au cours du temps, c'est-à-dire $z(t) = -\frac{gt^2}{2} + z_0$.

Proposez plusieurs méthodes pour calculer g . Essayez de comparer ces méthodes selon divers critères :

- la précision
- celle qui demandera le moins de mesures ?
- celle qui demandera le temps d'expérience le plus court ?
- celle qui demandera l'appareillage expérimental le plus simple ?

- celle qui sera la plus *fiable*, vous permettant non seulement de mesurer g mais aussi de vérifier le bien-fondé de votre modèle ?
2. Le calcul du rayon de la Terre par Erathosthène est-il, selon vous, un « problème inverse » ? A quel « problème direct » correspondrait-il ? Expliquez pourquoi.
 3. Cherchez, dans votre vie quotidienne ou votre culture scientifique, un ou des exemples de problèmes « inverses » et formulez les problèmes « directs » correspondants. Pourquoi la dénomination direct/inverse vous semble-t-elle naturelle ? leur résolution vous semble-t-elle importante - plus ou moins que celle du problème « direct » ? Réfléchissez à leur résolution : comment vous y prendriez-vous, quelles difficultés voyez-vous ?

Chapitre 1

Quelques problèmes inverses en dynamique des population

Ce chapitre a pour objet de présenter deux grands types de problèmes inverses en dynamique des populations, auxquels ce cours fera très fréquemment référence. Nous y énonçons les principaux résultats de l'étude des problèmes « direct » qui seront utiles par la suite, et formulons, de manière encore imprécise, les problèmes inverses qui se posent naturellement. La formulation mathématique de ces problèmes se fera rigoureuse par la suite, lorsque le cours introduira les notions nécessaires (chapitre 2).

1.1 Très brève introduction à la dynamique des populations

Soit une population dont les individus évoluent, c'est-à-dire : grandissent, changent, meurent et se reproduisent. On souhaite suivre - et comprendre - le comportement des individus au cours du temps. Les équations dites « de population structurée » ont été écrites dans ce but : elles rendent compte de l'évolution de la population au cours du temps, en fonction d'une variable qu'on appelle « structurante », c'est-à-dire qu'elle est choisie pour être caractéristique de la croissance, de la division et de la mort des individus.

A ce stade, avant même l'écriture des équations (ou bien un processus stochastique, qui permet de décrire individuellement chaque cellule), une question se pose : comment bien choisir la ou les variables structurantes ? Quelle grandeur *intrinsèque* permet de comprendre le développement de la population ? La réponse n'est bien sûr pas la même selon la population que l'on considère, et cette question elle-même peut d'ailleurs être vue comme un problème inverse.

Le premier des deux exemples d'applications en biologie qui va traverser tout ce cours est la division des bactéries. Dans le film en figure 1.1, vous voyez la division cellulaire de la bactérie *E. coli*. Ce sont des données très riches comparées à bien d'autres observations



FIGURE 1.1 – Croissance de bactéries E. Coli [57]

sur les populations puisqu'on peut « filmer » les cellules (avec une certaine précision tout de même : 1 photographie toutes les minutes ou toutes les deux minutes, à laquelle il faut ensuite appliquer un traitement d'image et de reconnaissance de formes), mesurer leur taille, mesurer le temps entre la naissance et la division, mesurer la taille de la mère et la taille de la fille, etc. Cependant toutes ces informations sont descriptives : elles ne donnent pas accès à la machinerie cellulaire, aux causes intrinsèques de la division, aux variables « structurant » la croissance et la division. Une question qui vient assez naturellement face à cela est : qu'est-ce qui *décide* de la division ? Par exemple, est-ce la taille de la cellule, est-ce son âge, ou bien est-ce encore autre chose ?

1.1.1 Quelques exemples de populations, et de mesures expérimentales.

- En biologie cellulaire :
- **les bactéries.** Dans le film en figure 1.1, on voit une colonie de bactéries E. Coli, dans un environnement suffisamment riche en nutriment, de sorte que rien ne freine la division et la croissance des bactéries. Le film provient de l'article [57]. Les questions que l'on se pose : comment se divisent-elles ? Comment la population évolue-t-elle ? Et un individu pris au hasard dans la population ? Peut-on, pour un modèle donné, tester sa validité en le comparant aux données, et peut-on estimer tous les paramètres du modèle ? Peut-on "remonter le temps" : à partir d'une donnée à un temps donné, remonter le temps pour connaître l'état initial ? Deux autres types d'expériences sont possibles.

1. A chaque division, une seule des deux cellules filles est conservée. Cela corres-

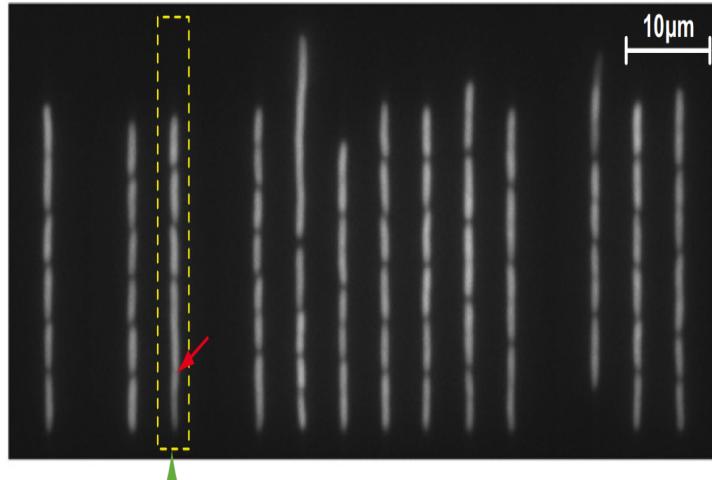


FIGURE 1.2 – Croissance de bactéries E. Coli [57], dans une expérience de microfluidique. On voit ici 13 micro-canaux, dans chacun d'eux une lignée de bactéries dont on ne garde qu'une seule cellule fille, les autres étant poussées vers le bas puis s'en allant.

pond à des expériences de *microfluidique*, protocole mis au point très récemment, d'abord sur les bactéries [62] puis sur les levures [18]. Voir la figure 1.2 pour les bactéries et la figure 1.4 pour les levures.

2. On observe uniquement un échantillon de tailles de la population à un instant donné. Ce type de mesures est beaucoup plus ancien, voir par exemple l'article de 1969 [37], dont on reproduit en figure 1.3 une des courbes. Il est aussi beaucoup plus simple et adaptable à de nombreuses situations expérimentales, où la croissance d'une colonie entière ou d'une lignée cellulaire est inenvisageable (par exemple : observation du plancton in situ, autres types de populations, etc.). Il est donc également très intéressant de comprendre comment tirer parti au mieux de telles données, plus pauvres.
 - **Les levures.** Dans le film en figure 1.4, on voit la division cellulaire de la levure *Saccharomyces cerevisiae*. Le protocole expérimental est différent : Les questions : qu'est-ce qui fait qu'une levure cesse de se diviser ? Quelles sont les caractéristiques de la division ?
 - Les cellules qui se différencient au fur et à mesure des divisions, comme les cellules sanguines, caractérisées par une évolution au fur et à mesure des divisions (depuis les cellules souches multi-potentes jusqu'aux globules).
 - **En chimie des protéines.** L'agrégation de protéines en fibres dites fibres *amyloïdes* est un mécanisme central dans de nombreuses maladies, appelées *maladies amyloïdes*

16 CHAPITRE 1. QUELQUES PROBLÈMES INVERSES EN DYNAMIQUE DES POPULATION

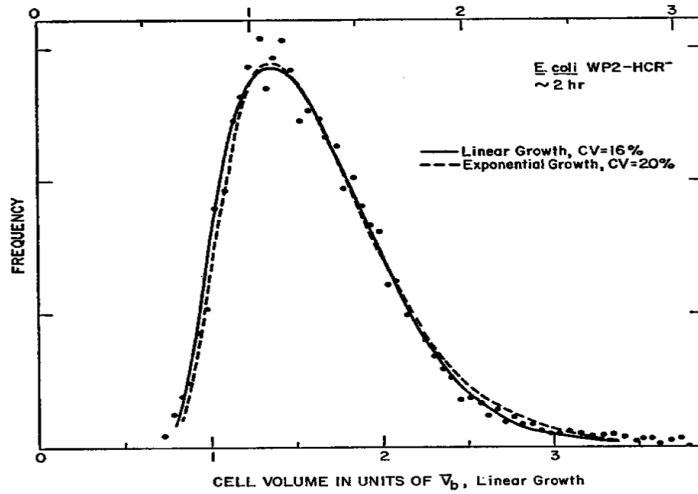


FIGURE 1.3 – Répartition de volumes de bactéries *E. coli* à un instant donné, image provenant de l'article [37] de 1969.

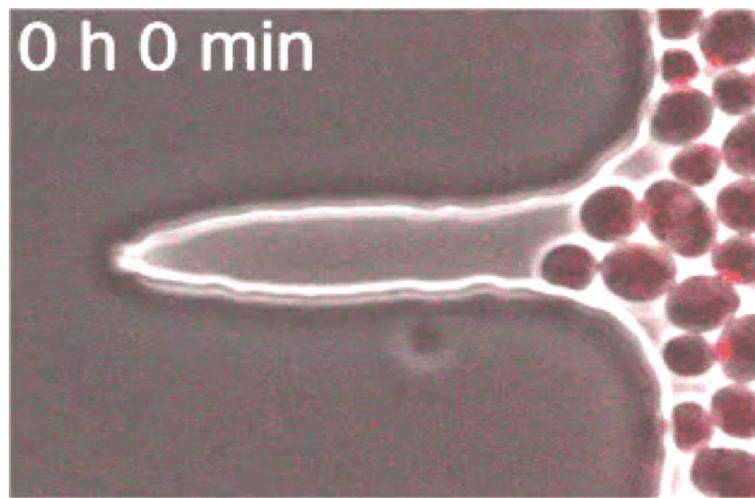


FIGURE 1.4 – Division cellulaire de levures, environnement microfluidique. On observe une des deux cellules filles après chaque division, prise au hasard. [18]

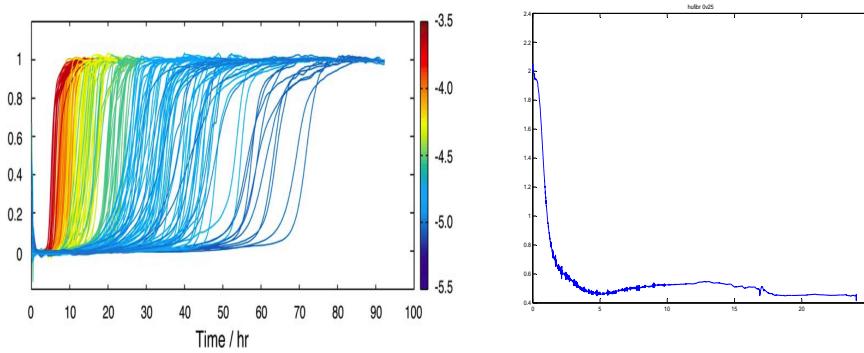


FIGURE 1.5 – mesure de l'évolution temporelle 1/ à gauche : de la masse totale polymérisée [64], dans une expérience de polymérisation spontanée de $\beta 2$ microglobuline, 2/ à droite : du moment d'ordre deux par SLS (H. Rezaei et J. Torrent), dans une expérience de dépolymérisation/fragmentation de la protéine PrP.

en raison de cette propriété. Parmi elles, on peut citer la maladie d'Alzheimer (agrégation de la protéine $A\beta$ et de la protéine τ), les maladies à prion (agrégation de la protéine PrP), la maladie de Huntington (polyglutamine), de Parkinson etc. Dans ces maladies, on observe qu'une protéine, normalement à l'état de monomère, acquiert la capacité pathologique de s'agréger l'une à l'autre, formant ces longs polymères appelés fibres amyloïdes.

Les polymères ne sont évidemment pas à proprement parler des "individus", en ce sens que ce ne sont pas des organismes vivants, mais on peut toutefois adopter des formalismes très proches, du fait qu'ils peuvent se fragmenter (comme les cellules se divisent) ou s'allonger par ajout de monomères (comme les cellules grandissent). D'autres réactions sont cependant également possibles.

Les données sont en général beaucoup plus difficiles à obtenir et plus bruitées en raison de la taille des protéines, environ mille fois plus faible. On distingue plusieurs types de données concernant les expériences *in vitro* (*in vivo* les données sont encore plus parcellaires).

1. Le plus fréquemment, les biologistes mesurent, pour une expérience donnée, l'évolution temporelle d'une quantité dépendant de tous les polymères : par exemple la masse totale polymérisée (c'est la quantité de monomères inclus dans des polymères) (expérience par Thioflavine T), ou encore le »moment d'ordre deux « (voir ... ci-dessous pour une définition), mesuré par dispersion de lumière statique (Static Light Scattering, abrégé en SLS). Voir la figure 1.5.
2. Une répartition des tailles des polymères à un instant donné (équivalent pour

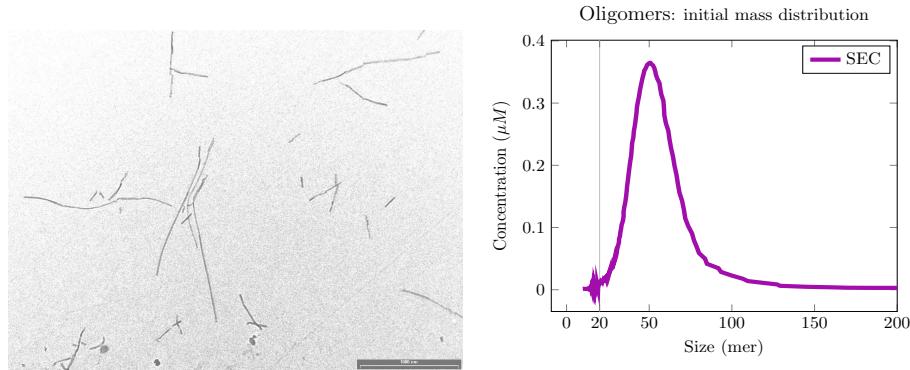


FIGURE 1.6 – mesure de tailles de polymères : 1/ à gauche : par chromatographie d'exclusion de taille pour des oligomères de PrP (équipe d'H. Rezaei), 2/ à droite : par photographie d'un échantillon de fibres de PrP par microscopie électronique (S. Prigent [50]).

les polymères d'une mesure telle qu'en figure 1.3) peut également être obtenue, pour les petits polymères, appelés *oligomères*, par technique de chromatographie d'exclusion de taille, ou pour les fibres les plus grandes par microscopie électronique. Voir la figure 1.6.

Les problèmes inverses les plus typiques : trouver et quantifier les réactions chimiques les plus importantes.

- en théorie de l'évolution : mutations au cours du temps. Données : beaucoup plus parcellaires... Souvent trop pour pouvoir réellement faire des problèmes inverses. C'est là une autre caractéristique des problèmes inverses : ils se nourrissent de données ! Lorsqu'elles sont trop lacunaires, beaucoup de modèles différents sont possibles, on peut trouver plus sage de s'en tenir à des modèles relativement simples qui peuvent suffire à donner une bonne idée qualitative des phénomènes sans se perdre dans des détails invérifiables. C'est alors davantage l'étude des problèmes directs qui est importante.
- etc.

1.1.2 Comment obtenir une description mathématique de la population ?

Cette étape de modélisation est essentielle et plus subtile qu'il n'y paraît. Dans le cadre d'un problème inverse appliqué, schématisons un peu les différentes étapes.

Tout d'abord, il y a une (pré) formulation du problème. Dans les deux exemples du

cours, ce sera :

- pour les bactéries : comment décrire fidèlement la division cellulaire, avec un modèle « suffisamment simple » ? Les données nous permettent-elles de retrouver les lois de la division cellulaire ?
- pour la polymérisation des protéines : quelles sont les réactions dominantes ? Peut-on retrouver, à partir de données sur l'évolution d'un système, son état initial ?

Ces formulations sont imprécises. C'est naturel quand on s'attaque à un problème : elles s'affineront progressivement. Une fois ces questions-cadres posées, **et** une fois une première idée sur les données expérimentales accessibles, nous pouvons choisir un cadre mathématique qui nous semble adéquat.

Prenons l'exemple de la division cellulaire : nous voulons modéliser ou bien la population, ou bien un individu de cette population. A quoi avons-nous accès ?

Nous avons deux grands cadres mathématiques possibles :

- processus probabiliste : il décrira le comportement de chaque cellule,
- équation de population : elle décrira **ou bien** la population dans son ensemble **ou bien** le comportement *moyen* d'une cellule.

Il y a des liens très étroits entre ces deux cadres mathématiques. Nous en dirons un peu plus par la suite, sans pour autant aller au fond du problème ce qui dépasserait largement ce cours.

1.1.3 Quelques équations fondamentales de la dynamique des populations

Voici quelques exemples de variable structurante et d'équations fondamentales de la dynamique des populations. Les mêmes équations - ou des variantes - peuvent bien sûr se retrouver dans de nombreuses applications, parmi lesquelles la division cellulaire occupe une grande place. Pour un exposé détaillé sur de tels modèles, deux livres de référence sont [44, 48].

Pour obtenir ces équations, deux voies sont possibles. La première est celle d'un bilan de masse, dans le même esprit que pour les équations de la physique des fluides. La seconde est celle des processus stochastiques cités plus haut : l'équation peut être obtenue comme satisfait en espérance par la mesure empirique liée au processus. Je détaille ici la méthode du bilan de masse, et je renvoie à [25] pour un article récent exposant comment on peut l'obtenir, dans un cas particulier, à partir d'un processus stochastique.

– L'âge.

Les premières « équations structurées en âge » ont été écrites au début du XXème siècle par Sharpe et Lotka [54], Mc Kendrick [43], Kermack [34, 35]. En 1911, dans un cadre discret, Sharpe et Lotka démontrèrent que lorsque la naissance et la mort sont indépendants du temps, la répartition des âges tend vers un profil stationnaire

indépendant de la condition initiale. La preuve, dans le cas matriciel, repose sur le théorème de Perron-Frobenius.

Dans une de ses variantes, une telle équation peut être écrite comme suit :

$$\begin{cases} \frac{\partial}{\partial t}n(t, a) + \frac{\partial}{\partial a}n(t, a) = -\mu(a)n(t, a), \\ n(t, a = 0) = \int_0^{\infty} B(a)n(t, a)da, \quad n(t = 0, a) = n^{in}(a), \end{cases} \quad (1.1)$$

où $n(t, a)$ représente la concentration de cellules d'âge a à l'instant t , $B(a)$ le taux de naissance et $\mu(a)$ le taux de mortalité. Le cas où une cellule se divise en deux cellules filles peut être vu comme un cas particulier, en remplaçant $\mu(a)$ par $\mu(a) + B(a)$ (car les cellules qui se divisent quittent l'âge a comme si elles mouraient) et, dans la condition au bord, en remplaçant $B(a)$ par $2B(a)$ (car il y a doublement d'une cellule qui se divise).

Si l'on se place dans le contexte de l'expérience de la figure 1.2, comme on ne garde qu'une seule des deux filles à chaque division, il faut uniquement remplacer $\mu(a)$ par $\mu(a) + B(a)$. On obtient alors une équation dite *conservative*.

Vous avez tous déjà vu des pyramides des âges : il s'agit de voir la répartition d'une population en fonction de son âge. Pour autant, cela signifie-t-il que l'âge est la seule « variable structurante » ? Clairement non : par exemple les guerres peuvent expliquer un creux dans la pyramide. Ici cependant, les équations modélisent le fait que **l'âge suffit à caractériser l'évolution de la population**. Si ce n'était pas le cas, il faudrait, au minimum, considérer un taux de division $B(t, a)$ et un taux de mortalité $\mu(t, a)$ qui dépendent également du temps. Cependant de tels procédés sont un peu artificiels et peu informatif, car cette dépendance par rapport au temps peut recouvrir tout et n'importe quoi.

Une autre hypothèse sous-jacente à ce modèle est l'absence de mémoire (ou propriété de Markov pour les processus) : la cellule fille, à sa naissance, a oublié le comportement de la cellule mère.

- La taille.

Les modèles structurés en taille ont été introduits dans la deuxième moitié du XXème siècle, dans plusieurs travaux presque simultanés : Bell et Anderson [12, 13], Sinko et Streifer [55, 56].

En toute généralité, l'équation s'écrit

$$\begin{cases} \frac{\partial}{\partial t}n(t, x) + \frac{\partial}{\partial x}(g(x)n(t, x)) = -\mu(x)n(t, x) - B(x)n(t, x) + 2 \int_x^{\infty} k(y, x)B(y)n(t, y)dy, \\ g(x = 0)n(t, x = 0) = 0, \quad n(t = 0, x) = n^0(x). \end{cases} \quad (1.2)$$

Dans cette équation, $n(t, x)$ désigne la concentration de cellules de taille x à l'instant t , qui peut mourir avec un taux $\mu(x)$, ou se diviser avec un taux $B(x)$, donnant

naissance à deux individus de taille y et $x - y$ avec un taux de probabilité $k(y, dx)$ satisfaisant

$$\int_0^y k(y, dx) = 1, \quad \int_0^y xk(y, dx) = \frac{1}{2}. \quad (1.3)$$

On doit au moins supposer que pour tout y , $k(y, \cdot)$ est une mesure de probabilité à support sur $[0, y]$ et que pour toute fonction continue ψ , l'application $f_\psi : y \rightarrow \psi(x)k(y, dx)$ est Lebesgue-mesurable.

Il y a plusieurs variantes de cette équation qu'on peut aussi généraliser à de la fragmentation non binaire. Ce qui est appelé ici « taille » peut aussi bien représenter une quelconque grandeur physique telle que la longueur, le volume, ou encore un contenu en protéine ou en parasite. Pour cette raison, la nomenclature n'est pas encore unifiée et demeure un peu confuse. En effet le terme de « taille », « âge » ou de « contenu en protéine » fait référence à l'application biologique davantage qu'à la structure mathématique du modèle. Ainsi par exemple certains articles parlent de modèle en taille, *e.g.* [1, 28, 6], alors que je les appellerais plutôt modèles en âge¹, car dans ces modèles la taille à la naissance est la même pour toutes les cellules et pourrait donc être fixée à zéro sans perdre en généralité - exactement comme dans un modèle en âge. De cette façon, la structure en âge apparaît comme un cas particulier de l'équation de croissance-fragmentation où $k(x, y) = \frac{1}{2}(\delta_{x=0} + \delta_{x=y})$.

Afin d'éviter de telles confusions, je préfère parler de ce modèle comme de l'équation de croissance-fragmentation ou de transport-fragmentation (en effet, la croissance peut aussi faire référence à de la coagulation).

- maturité [52], label fluorescent [7], ADN [9], satiété [44] etc.
- un modèle qui connaît un intérêt très récent : le modèle incrémental [58]. La variable structurante n'est plus l'âge mais l'accroissement de taille entre la naissance et la division. Un autre modèle classique est le modèle structuré en âge et en taille [55].

Les équations ci-dessus sont des équations linéaires. On peut aussi en écrire des versions non linéaires, afin de prendre en compte l'interaction avec le milieu ou l'inhibition. Cela se fait par exemple en choisissant des paramètres qui dépendent à la fois de la variable structurante et de la population- par exemple, avec un taux de croissance qui diminue quand la population augmente, etc. Nous résumons brièvement ci-dessous le comportement en temps long dans le cas linéaire uniquement.

1.1.4 Comportement en temps long des équations linéaires

Pour un exposé à la fois détaillé et abordable, je renvoie au livre de Benoît Perthame [48], ch. 3 et 4. Je me borne à donner ici quelques intuitions sur les résultats.

1. L'équation structurée en âge pourrait aussi être généralisée avec une vitesse de vieillissement $g(a)$ pas nécessairement égale à 1.

De manière formelle, les équations peuvent s'écrire sous la forme $\frac{dn}{dt} = An(t)$, où l'on note $n(t) = n(t, \cdot)$ et (en prenant $\mu \equiv 0$)

$$An = -\frac{\partial}{\partial x}(g(x)n) - Bn + 2 \int_x^t nfty B(y)k(y, x)n(t, y)dy.$$

Dans le cas de la dimension finie, donc si on avait $n(t) \in \mathbb{R}^k$ et $A \in \mathcal{M}_k(\mathbb{R})$ une matrice, et si A est une matrice irréductible à coefficients positifs, le théorème de Perron-Frobenius assure que le rayon spectral de A est une valeur propre simple $\lambda > 0$ associée à un vecteur propre strictement positif $N \in \mathbb{R}_+^k$. De cela, on déduit facilement le comportement en temps long de l'équation $\frac{dn}{dt} = An(t)$, $n(0) = n^0$: on a

$$n(t)e^{-\lambda t} \rightarrow < n^0, N > N,$$

et la vitesse de convergence est donnée par la distance entre λ et les valeurs propres suivantes, distance appelée *trou spectral*.

Dans le cas des opérateurs sur les espaces de Banach de dimension infinie, le théorème de Krein-Rutman généralise le théorème de Perron-Frobenius à des opérateurs compacts. Sous certaines hypothèses sur les coefficients $g(x)$ et $B(x)$, hypothèses que nous ne précisons pas pour le moment, on peut donc montrer qu'en temps grand on a

$$n(t, x)e^{-\lambda t} \xrightarrow{\text{espace de convergence adéquat...}} N(x),$$

où $(\lambda, N(x))$ est l'unique solution du *problème aux valeurs propres*

$$\begin{aligned} \lambda N(x) + \frac{\partial}{\partial x}(g(x)N(x)) &= -B(x)N(x) + 2 \int_x^\infty k(y, x)B(y)N(y)dy, \\ g(x=0)N(x=0) = 0, \quad \int_0^\infty N(x)dx &= 1, \quad \lambda > 0, \quad N(x) \geq 0. \end{aligned} \tag{1.4}$$

1.1.5 Estimation en dynamique des populations

Maintenant que nous disposons d'un certain nombre de modèles, nous pouvons revenir à la question initiale : qu'est-ce qui détermine la division cellulaire ? Pour commencer à y répondre, il faut comparer le ou les modèles aux données. Donc

1. Choisir un modèle ou une famille de modèles.
2. Pour ces modèles : qu'est-ce qu'on connaît ? Qu'est-ce qu'on ne connaît pas ?
3. Ce qu'on ne connaît pas : comment l'estimer ?

Le comportement en temps long de l'équation nous sera très utile.

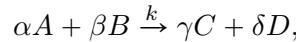
1.2 Problèmes inverses pour la polymérisation des protéines

Nous avons décrit plus haut (partie 1.1.1, et figures 1.5 et 1.6) ce qu’était la polymérisation des protéines, son importance dans les maladies amyloïdes, ainsi que le type de mesures expérimentales qui était possible. Nous avons vu que les questions typiques qui se posent sont : comment estimer des paramètres réactionnels ? Comment estimer une répartition de tailles de polymères à partir de données moyennées sur l’ensemble des tailles ?

Pour pouvoir exprimer ce problème inverse en termes mathématiques, la première étape est la modélisation.

1.2.1 Modèle discret

Pour modéliser la polymérisation des protéines, on commence par lister toutes les réactions qui peuvent se produire. Pour chaque réaction, on écrit la loi d'action de masse telle que découverte empiriquement par Guldberg et Waage en 1863 : on symbolise la réaction telle que α molécules de A réagit avec β molécules de B pour donner γ molécules de C et δ molécules de D , à un taux de réaction k , par le schéma réactionnel



et on écrit alors les équations sur les concentrations, notées $[A]$, $[B]$, $[C]$ et $[D]$:

$$\frac{d[A]}{dt} = -\alpha[A]^\alpha[B]^\beta, \quad \frac{d[B]}{dt} = -\beta[A]^\alpha[B]^\beta, \quad \frac{d[C]}{dt} = +\gamma[A]^\alpha[B]^\beta, \quad \frac{d[D]}{dt} = +\delta[A]^\alpha[B]^\beta.$$

On dit que la dynamique est d’ordre $\alpha + \beta$. Nous nous limitons ici à une seule espèce de protéines, et notons $u_i(t)$ la concentration de polymères contenant i monomères.

- dégradation : chaque polymère peut être dégradé avec un taux de dégradation k_m^i
- nucléation : les monomères peuvent spontanément s’agrérer, à un taux k_{pol}^N , pour donner naissance à un polymère de taille i_0 , appelé *nucleus*. C’est le plus petit polymère stable et ce nucleus peut se dissocier en monomères à un taux k_{frag}^N . Voir [47] pour plus de détails. C’est donc une cinétique d’ordre i_0 :



- Les polymères de taille $i \geq i_0$ peuvent polymériser ou dépolymériser par gain ou perte d’un monomère, avec un taux respectivement k_{pol}^i et k_{dep}^i .



- Les polymères peuvent s'agréger l'un à l'autre par coagulation, ou se casser en polymères plus petits par fragmentation. On néglige ici la fragmentation en plus de deux morceaux, considérant qu'elle est beaucoup plus rare. Les taux sont notés $k_{col}^{i,j}$ pour la coagulation de deux polymères de taille i et j , et $k_{frag}^{j,i}$ pour la fragmentation d'un polymère de taille j en deux polymères de tailles i et $j-i$.



On aurait pu garder la même notation pour la fragmentation et la dépolymérisation en notant $k_{frag}^{1,i} = k_{frag}^{i-1,i} = \frac{1}{2}k_{dep}^i$; et de même pour la coagulation et la polymérisation. On préfère cependant les distinguer car il s'agit de réactions de nature différente, dont les ordres de grandeur peuvent être très différents.

On définit $K_{frag}^j = \sum_{i=2}^{j-2} k_{frag}^{j,i}$. Cela représente le taux total auquel un polymère de taille j peut se casser. Par symétrie on a $k_{frag}^{j,i} = k_{frag}^{j,j-i}$, comme on avait pour la fragmentation continue $k(y, x) = k(y, y-x)$, et aussi $k_{col}^{i,j} = k_{col}^{j,i}$.

Le modèle ci-dessous est l'exacte transcription déterministe de toutes ces réactions, en suivant la loi d'action de masse. Pour pouvoir se placer dans ce cadre, on a fait l'hypothèse que le nombre de molécules de chaque espèce est très grand : si ce n'est plus le cas, il faudra prendre en compte un modèle stochastique.

$$\begin{aligned} \frac{du_1}{dt} &= -i_0 k_{pol}^N (u_1)^{i_0} + i_0 k_{frag}^N u_{i_0} - k_m^1 u_1 \\ &\quad - u_1 \sum_{i \geq i_0} k_{pol}^i u_i + \sum_{i=i_0+1}^{\infty} k_{dep}^j u_j + 2 \sum_{i=2}^{i_0-1} \sum_{j=i_0}^{\infty} i k_{frag}^{i,j} u_j, \end{aligned} \quad (1.8)$$

$$\begin{aligned} \frac{du_{i_0}}{dt} &= k_{pol}^N (u_1)^{i_0} - k_{frag}^N u_{i_0} - k_{pol}^{i_0} u_{i_0} u_1 + k_{dep}^{i_0+1} u_{i_0+1} - k_m^{i_0} u_{i_0} \\ &\quad + 2 \sum_{j=i_0+2}^{\infty} k_{frag}^{i_0,j} u_j - K_{frag}^{i_0} u_{i_0} - \sum_{j \geq i_0} k_{col}^{i_0,j} u_{i_0} u_j, \end{aligned} \quad (1.9)$$

$$\begin{aligned} \frac{du_i}{dt} &= u_1 (k_{pol}^{i-1} u_{i-1} - k_{pol}^i u_i) - (k_{dep}^i u_i - k_{dep}^{i+1} u_{i+1}) - k_m^i u_i \\ &\quad + 2 \sum_{j=i+2}^{\infty} k_{frag}^{i,j} u_j - K_{frag}^i u_i + \frac{1}{2} \sum_{i_0 \leq j \leq i-2} k_{col}^{j,i-j} u_j u_{i-j} - \sum_{j \geq i_0} k_{col}^{i,j} u_i u_j. \end{aligned} \quad (1.10)$$

Ce modèle ou des variantes de ce modèle a été très étudié, ou bien dans le cas général en mathématiques, voir [4, 24] et les références citées par ces travaux, ou, en biologie, en lui appliquant des hypothèses simplificatrices très fortes, voir [47, 36, 64, 17].

Une propriété fondamentale est, comme pour les équations structurées, le bilan de masse : en sommant toutes les équations, l'équation u_i étant multipliée par le poids i , on obtient

$$\frac{d}{dt} \left(u_1 + \sum_{i=i_0}^{\infty} i u_i \right) = -k_m^{(1)} u_1 - \sum_{i \geq i_0} i k_m^i u_i. \quad (1.11)$$

Cela signifie que la masse totale ne change que par production (pas de terme dans cette version du modèle) ou dégradation : la polymérisation, la fragmentation ou la coagulation ne l'influencent pas.

De même on peut définir $P(t) = \sum_{i \geq i_0} u_i$ le nombre total de polymères, et on obtient en sommant les équations

$$\begin{aligned} \frac{dP}{dt} &= k_{pol}^N (u_1)^{i_0} - k_{off}^N u_{i_0} - \sum_{i \geq i_0} k_m^i u_i \\ &\quad + \sum_{i=i_0}^{\infty} K_{frag}^i u_i - \frac{1}{2} \sum_{j \geq i_0} \sum_{i \geq i_0} k_{col}^{i,j} u_i u_j. \end{aligned} \quad (1.12)$$

Cela signifie que la polymérisation et la dépolymérisation ne changent pas le nombre total de polymères : ces termes sont absents de l'équation ci-dessus.

Ces deux propriétés sont fondamentales, et tout modèle approché doit les sauvegarder.

Ce modèle cadre est très utile en particulier quand la taille moyenne des polymères, notée i_M , reste relativement petite. En revanche, quand elle devient grande (elle peut valoir jusqu'à 10^6 dans le cas des fibres amyloïdes), et lorsqu'il s'agit de faire des simulations intensives pour pouvoir estimer les paramètres, la complexité devient trop importante. Il est alors utile, tant pour l'analyse théorique que pour l'analyse appliquée, de se tourner vers une version continue de ce modèle. En effet, on peut remarquer sa ressemblance avec le modèle précédent de croissance fragmentation, et interpréter, par exemple, la polymérisation comme une croissance discrète, et la fragmentation comme une fragmentation discrète [63].

1.2.2 Du modèle discret au modèle continu

Nous nous bornons ici au modèle où l'on ne considère que la polymérisation et la dépolymérisation, et prenons $i_0 = 2$, qui sont considérées comme des réactions principales (*primary pathway*) en biologie, la fragmentation ou la coalescence étant considérées davantage comme des réactions secondaires (*secondary pathway*) [47]. Le modèle cadre devient alors

$$\frac{du_1}{dt} = -u_1 \sum_{i \geq i_0} k_{pol}^i u_i + \sum_{i=i_0+1}^{\infty} k_{dep}^j u_j \quad (1.13)$$

$$\frac{du_i}{dt} = u_1 (k_{pol}^{i-1} u_{i-1} - k_{pol}^i u_i) - (k_{dep}^i u_i - k_{dep}^{i+1} u_{i+1}) \quad (1.14)$$

Ce modèle est appelé *système de Becker-Döring* [11], il a été introduit par Becker et Döring en 1935. Sa limite vers un système continu, qui se trouve être le système introduit par Lifshitz et Slyozov [40] en 1961, a été étudiée en 2002 dans l'article [22], dont nous nous bornons ici à donner une brève introduction. D'autres limites ont été prouvées pour des équations faisant intervenir d'autres termes du modèle-cadre (1.8)–(1.10), par exemple [38] pour la coagulation-fragmentation ou [24] pour la croissance-fragmentation.

Des conditions sur les coefficients (k_{pol}^i , $k_{dep}^i = \mathcal{O}(\sqrt{i})$), ainsi qu'une réécriture de l'équation faisant apparaître les termes de flux $J_i = k_{pol}^i u_1 u_i - k_{dep}^{i+1} u_{i+1}$, permettent d'obtenir des résultats d'existence et d'unicité d'une solution [5] pour une condition initiale de masse totale finie.

Dans le cas où la taille moyenne des polymères, notée i_M , est grande, on voudrait pouvoir remplacer la taille discrète i par une variable continue $x > 0$. Pour cela, on commence par adimensionner l'équation, dans le but de faire apparaître d'une part des quantités sans dimension, supposées être d'ordre de grandeur 1, et d'autre part des valeurs caractéristiques.

On introduit donc les quantités caractéristiques suivantes.

- \mathcal{U}_1 la valeur caractéristique de la concentration en monomères u_1 ,
- \mathcal{U} la valeur caractéristique de la concentration en polymères u_i ,
- T la valeur caractéristique du temps t ,
- \mathcal{K}_{pol} et \mathcal{K}_{dep} les valeurs caractéristiques de la polymérisation et de la dépolymérisation k_{pol}^i et k_{dep}^i ,

et on définit les quantités adimensionnées

$$\begin{aligned}\bar{u}_1(t) &= \frac{u_1(t)}{\mathcal{U}_1}, \quad \bar{u}_i(t) = \frac{u_i(t)}{\mathcal{U}} \quad \text{for } i > i_0, \\ \bar{k}_{pol}^i &= \frac{k_{pol}^i}{\mathcal{K}_{pol}}, \quad \bar{k}_{dep}^i = \frac{k_{dep}^i}{\mathcal{K}_{dep}}.\end{aligned}$$

On obtient les équations sans dimension suivantes

$$\frac{d\bar{u}_1}{dt} = -\mathcal{K}_{pol}\mathcal{U}\bar{u}_1 \sum_{i \geq 2} \bar{k}_{pol}^i \bar{u}_i + \frac{\mathcal{K}_{dep}\mathcal{U}}{\mathcal{U}_1} \sum_{j=3}^{\infty} \bar{k}_{dep}^j \bar{u}_j, \quad (1.15)$$

$$\frac{d\bar{u}_i}{dt} = \mathcal{U}_1 \mathcal{K}_{pol} \bar{u}_1^* (\bar{k}_{pol}^{i-1} \bar{u}_{i-1} - \bar{k}_{pol}^i \bar{u}_i) + \mathcal{K}_{dep} (\bar{k}_{dep}^{i+1} \bar{u}_{i+1} - \bar{k}_{dep}^i \bar{u}_i) \quad (1.16)$$

La conservation de la masse s'écrit

$$\frac{d}{dt} \left(\bar{u}_1 + \gamma \sum_{i=2}^{\infty} \bar{u}_i \right) = 0,$$

avec $\gamma = \frac{\mathcal{U}}{\mathcal{U}_1}$. Si l'on considère une taille moyenne i_M , on aura $\sum i u_i \approx i_M^2 \mathcal{U}$, donc pour que les termes u_1 et $\sum i u_i$ soient d'ordres de grandeur comparables cela conduit à $\mathcal{U}_1 = i_M^2 \mathcal{U}$, soit encore $\gamma = \frac{1}{i_M^2}$.

D'un autre côté, si l'on introduit $x_i = \varepsilon i$ pour $\varepsilon = \frac{1}{i_M}$, de sorte que x_i soit en moyenne de l'ordre de un. On voit alors qu'on a

$$\gamma \sum i \bar{u}_i = \sum (\varepsilon i) \bar{u}_i \varepsilon,$$

ce que l'on peut interpréter comme une somme de Riemann pour la fonction en escalier

$$u^\varepsilon(t, x) := \sum_{i=2}^{\infty} \bar{u}_i(t) \mathbb{1}_{[x_i, x_{i+1}]}(x),$$

où $\mathbb{1}_A$ désigne la fonction indicatrice de l'ensemble A .

De même, dans l'équation (1.16), si l'on choisit

$$\mathcal{U}_1 \mathcal{K}_{pol} = \mathcal{K}_{dep} = \frac{1}{\varepsilon},$$

et si l'on définit, comme pour u_i , des fonctions paramètres

$$a_\varepsilon(x) := \sum_{i=2}^{\infty} \bar{k}_{pol}^i \mathbb{1}_{[x_i, x_{i+1}]}(x), \quad b_\varepsilon(x) := \sum_{i=2}^{\infty} \bar{k}_{dep}^i \mathbb{1}_{[x_i, x_{i+1}]}(x),$$

on reconnaît un schéma aux différences finies pour u . Ces remarques conduisent à deviner heuristiquement le système limite (au premier ordre : au-delà il faudrait faire un développement de Taylor à l'ordre deux...) : c'est le système de Lifshitz-Slyozov [40]

$$\frac{\partial u(t, x)}{\partial t} = -u_1 \frac{\partial}{\partial x} (a(x) u(t, x)) + \frac{\partial}{\partial x} (b(x) u(t, x)), \quad (1.17)$$

$$\frac{d}{dt} \left(u_1(t) + \int_0^\infty x u(t, x) dx \right) = 0. \quad (1.18)$$

Le résultat suivant a été établi dans [22].

Théorème 1 (Théorème 2.3. de [22]). *Supposons que les coefficients $k_{pol}^i \geq 0$ et $k_{dep}^i \geq 0$ vérifient*

$$k_{pol}^i, k_{dep}^i \leq K, \quad |k_{pol}^{i+1} - k_{pol}^i| \leq K/i, \quad |k_{dep}^{i+1} - k_{dep}^i| \leq K/i$$

pour une certaine constante $K > 0$. Considérons une suite $\varepsilon_n \rightarrow 0$. Alors il existe une sous-suite, toujours notée ε_n , et deux fonctions $a, b \in W_{loc}^{1,\infty}((0, \infty)) \cap L^\infty((0, \infty))$ telles que, pour tout $0 < r < R$, on ait

$$\lim_{\varepsilon_n \rightarrow 0} \sup_{i \in (r/\varepsilon_n, R/\varepsilon_n)} (|k_{pol}^i - a(\varepsilon_n i)| + |k_{dep}^i - b(\varepsilon_n i)|) = 0.$$

Supposons de plus qu'il existe quatre constantes $0 < s \leq 1$, $M_0 < \infty$, $\rho < \infty$, $M_s < \infty$ telles que pour tout $\varepsilon > 0$

$$\varepsilon \sum_{i=2}^{\infty} u_i^{0,\varepsilon}, \quad u_1^{0,\varepsilon} + \varepsilon^2 \sum_{i=2}^{\infty} i u_i^{0,\varepsilon} = \rho, \quad \varepsilon \sum_{i=2}^{\infty} (\varepsilon i)^{1+s} u_i^{0,\varepsilon} \leq M_s.$$

Alors la suite ε_n peut être choisie de telle sorte que

$$\begin{cases} u^\varepsilon \rightarrow u, xu^\varepsilon \rightarrow xu \text{ dans } C^0([0, T]; \mathcal{M}^1(0, \infty) - \text{weak} - *), \\ u_1^\varepsilon \rightarrow u \text{ uniformément dans } C^0([0, T]), \end{cases}$$

où (u, u_1) est solution au sens faible du système (1.17)(1.18).

Nous ne détaillons pas davantage ici la preuve de ce résultat, dont il s'agit ici uniquement de donner un aperçu. Le point essentiel est qu'elle repose sur des **estimations sur les moments**, qui permettent d'obtenir de la compacité faible sur la suite u^ε . On rappelle qu'on appelle *moment d'ordre k* de la fonction $u(t, \cdot)$ la fonction $\mu_k(t)$ définie formellement par

$$\mu_k(t) := \int_0^\infty x^k u(t, x) dx.$$

On remarque aussi que la convergence obtenue est une convergence faible, et ne garantit aucune vitesse de convergence du modèle de Becker-Döring vers le modèle de Lifshitz-Slyozov. De fait, en temps grand, les deux modèles ont des comportements différents. De nombreux problèmes mathématiques restent ouverts.

On peut étendre au cas général, vu dans le paragraphe 1.2.1, ce formalisme continu : cela a été fait dans l'article [49].

1.2.3 Estimation en polymérisation des protéines

Nous avons maintenant deux modèles-cadres, l'un discret et l'autre continu, qui nous mettent en mesure de formuler mathématiquement les problèmes inverses que nous souhaitons étudier. Dans ce premier chapitre, nous regardons uniquement le problème dit de l'*estimation d'état* : comment, à partir de mesures (partielles et bruitées) sur l'évolution temporelle des concentrations de polymères, peut-on estimer l'état initial du système ? Les mesures concernées sont celles illustrées par la figure 1.5.

Formulation générale

Commençons par une formulation très générale, qui ne présuppose pas l'un ou l'autre modèle, ni telle ou telle réaction, ce qui la rend facilement adaptable.

1. Formalisation du problème direct

Soit \mathcal{Y} l'espace d'état, et $u : [0, T] \rightarrow \mathcal{Y}$ la solution d'un problème d'évolution qu'on écrit sous la forme

$$\frac{du}{dt} = Au, \quad u(0) = u_0. \quad (1.19)$$

L'opérateur A peut a priori dépendre du temps et/ou dépendre de u . S'il ne dépend pas de u , le problème est linéaire. Si A est le générateur d'un semi-groupe fortement continu, on peut noter $u(t) = \mathbb{T}_{|t}u_0$ où $\mathbb{T}_{|t}$ est le semi-groupe généré par A .² On emploie encore $\mathbb{T}_{|t}$ pour désigner l'opérateur qui à u_0 associe l'unique solution de (1.19), en admettant ici que ce problème est bien posé.

2. Formalisation de la mesure expérimentale

On appelle *opérateur d'observation* et on note $C : \mathcal{Y} \rightarrow \mathcal{Z}$ un opérateur allant de l'espace d'état \mathcal{Y} dans ce qu'on appelle *l'espace d'observation* \mathcal{Z} , qui est l'espace où vivent les mesures (temporelles) effectuées.

3. Formalisation du problème de l'estimation d'état

Dans la communauté de l'assimilation de données, on pose le problème de l'estimation d'état comme

estimer u_0 à travers des mesures z générées pendant le temps $[0, T]$, connaissant le modèle de dynamique A et le modèle d'opérateur d'observation C .

Dans la communauté des problèmes inverses, on notera $\Psi_{|t} : \mathcal{Y} \rightarrow L^2([0, T])$ l'opérateur $u_0 \rightarrow C\mathbb{T}_{|t}u_0$, et on posera le problème inverse comme :

inverser $\Psi_{|t}$ pour reconstruire u_0 , à partir d'une mesure bruitée de $z(t)$ pour $t \in [0, T]$.

Ces deux problèmes sont bien sûr identiques, mais les approches diffèrent et conduisent à développer des méthodes distinctes. L'assimilation de données exploite le caractère temporel du problème, ce que ne fait pas la théorie des problèmes inverses généraux.

Application à la polymérisation des protéines

On reprend le cadre ci-dessus pour l'appliquer à notre problème.

1. Formalisation du problème direct

Dans le cas des modèles vus ci-dessus, l'espace d'état \mathcal{Y} sera $\ell^2(\mathbb{N})$ dans le cas discret et $\mathcal{Y} = L^2(\mathbb{R}_+)$ dans le cas continu - on pourra aussi ne considérer que des espaces tronqués, ce qui sera nécessairement le cas numériquement, et considérer \mathbb{R}_+^N et $L^2([0, x_{max}])$ respectivement.

2. Si vous ne connaissez pas la théorie des semi-groupes ne vous attardez pas à cette formulation pour le moment...

30 CHAPITRE 1. QUELQUES PROBLÈMES INVERSES EN DYNAMIQUE DES POPULATION

L'opérateur A , qui trouve son origine dans les équations (1.8)–(1.10), dépendra d'une part des réactions considérées et d'autre part du cadre - discret ou continu - choisi. Notons que

- plutôt que de prendre le modèle-cadre avec toutes ses équations, il convient d'adapter le modèle au contexte expérimental et à la question biologique soulevée, pour ne garder qu'un modèle suffisamment réduit. Il convient même de prendre pour commencer un modèle le plus réduit possible, et de ne le complexifier que si c'est nécessaire.
- Le cadre, discret ou continu, devra faire l'objet d'un questionnement suivi et au minimum d'une vérification *a posteriori* (analyse de convergence), pour s'assurer de la validité du modèle approché utilisé.

2. Formalisation de la mesure expérimentale

Dans le cas des mesures de la figure 1.5, on voit qu'une mesure par Thioflavine T consiste à mesurer la masse polymérisée, qui s'écrit mathématiquement comme étant le moment d'ordre 1 :

$$\text{cas discret : } \mu_1[u](t) = \sum_{i=i_0}^{\infty} i u_i(t), \quad \text{cas continu : } \mu_1[u](t) = \int_0^{\infty} x u(t, x) dx.$$

Dans le cas de la mesure par SLS, c'est le moment d'ordre 2 qui est mesuré :

$$\text{cas discret : } \mu_2[u](t) = \sum_{i=1}^{\infty} i^2 u_i(t), \quad \text{cas continu : } \mu_2[u](t) = u_1(t) + \int_0^{\infty} x^2 u(t, x) dx.$$

Cela conduit à définir comme espace d'observation $\mathcal{Z} = \mathbb{R}_+$ et comme opérateur d'observation $C : u \rightarrow \mu_k[u]$.

3. Formalisation du problème de l'estimation d'état

Le problème inverse s'écrit : estimer la donnée initiale $u_0 \in L^2([0, x_{max}])$ ou $\ell^2(\mathbb{N})$, à partir d'une mesure bruitée de $z(t) = \mu_k[u](t)$ pour $t \in [0, T]$, où u est solution du système d'équations choisi donné par A .

Un problème spécifique : système dépolymérisant

Précisons encore les choses dans un cas extrêmement simple, mais correspondant pourtant à une première modélisation réaliste (figure 1.5 droite) : le cas où de toutes les réactions listées précédemment, on ne conserve que la dépolymérisation, qui plus est avec un taux de dépolymérisation constant. Nous plaçant dans le cadre du modèle continu, l'équation (1.17) devient, pour un $b > 0$ constant :

$$\frac{\partial u(t, x)}{\partial t} - b \frac{\partial}{\partial x}(u(t, x)) = 0, \quad 0 \leq x \leq x_{max}, \quad 0 \leq t \leq T.$$

On peut alors calculer explicitement l'opérateur $\Psi_{|t}$:

$$\Psi(u_0)(t) = \int_0^{x_{max}} x^k u(t, x) dx = \int_0^{x_{max}} x^k u_0(x + bt) dx = \int_{bt}^{x_{max}} (x - bt)^k u_0(x) dx.$$

On voit que

$$Im(\Psi) := \{z \in H^{n+1}([0, T]), \quad z(T) = \dots = z^{(k)}(T) = 0\},$$

et on a

$$\frac{d^{k+1}}{dt^{k+1}}(\Psi u_0) = (-b)^{k+1} k! u_0(bt).$$

Notre problème est devenu le problème de l'estimation de la dérivée $k + 1$ ième d'une fonction à partir de la mesure bruitée de cette fonction. Ou encore, celui de l'inversion de l'opérateur intégral itéré $k + 1$ fois. Il ne s'agit pas d'un problème bien posé : en effet, une petite erreur sur la norme L^2 de la mesure peut conduire à une très grosse erreur sur sa dérivée. C'est un des problèmes les plus emblématiques de la théorie des problèmes inverses, comme nous le verrons dans les chapitres suivants. Ce sera aussi l'exemple repris à travers l'ensemble de ce cours.

1.3 Exercices

1. Equation structurée en âge : en admettant que toutes les quantités nécessaires sont intégrables/dérivables, intégrez l'équation (1.1) contre le poids 1 et contre le poids a . Que remarquez-vous ? Comment l'interpréter ? Si vous supposez $B(a) = B$ constant et $\mu(a) = \mu$ constant, que pouvez-vous en déduire ? Résolvez l'équation dans ce cas.
2. Equation structurée en taille : en admettant que toutes les quantités nécessaires sont intégrables/dérivables, intégrez l'équation (1.2) contre le poids 1 et contre le poids x . Que remarquez-vous ? (NB : vous aurez besoin d'utiliser les propriétés (1.3) sur k). Comment l'interpréter ? Ecrivez l'équation dans le cas $k(y, x) = \frac{1}{2}(\delta_0 + \delta_{x=y})$ et dans le cas $k(y, x) = \delta_{x=\frac{y}{2}}$ (NB : attention aux intégrales de Dirac en deux dimensions). Comment interprétez-vous ces équations ?
3. Equations structurées : écrivez les formes faibles des équations. En fonction de l'espace sur lequel vous voulez prendre la solution, quelle régularité devez-vous supposer pour la fonction test ? Sur les fonctions paramètres ? Déduez-en les formulations des équations adjointes.
4. Equations structurées : essayer d'écrire un modèle structuré en taille et en accroissement de taille depuis la naissance.

32 CHAPITRE 1. QUELQUES PROBLÈMES INVERSES EN DYNAMIQUE DES POPULATIONS

5. Problèmes inverses : reprenez chaque type de mesures donné en exemple ci-dessus, et listez : 1/ ce à quoi on a un accès immédiat par la mesure ; 2/ les modèles qui peuvent s'appliquer ; 3/ les hypothèses nécessaires pour chaque modèle ; 4/ les inconnues qui restent ; 5/ les questions qui se posent. Parmi ces dernières, lesquelles vous semblent les plus importantes ? 6/ Où y a-t-il du bruit ou des erreurs possibles ?
6. Estimation de la dérivée : soit $f \in L^2(\mathbb{R})$. Donner un exemple de suite de fonctions $f_n \in H^1(\mathbb{R})$ telle que $f_n \xrightarrow{L^2(\mathbb{R})} f$ et $\|f'_n\|_{L^2(\mathbb{R})} \rightarrow +\infty$.
7. **Exercice possible lors de l'étude de chaque chapitre suivant :** appliquez la méthode ou la théorie exposée à un des problèmes inverses que vous aurez listé à la question précédente.

Chapitre 2

Introduction à la théorie des problèmes inverses linéaires

Dans les deux chapitres précédents, nous avons déjà entrevu les difficultés inhérentes aux problèmes inverses. La théorie générale va permettre de préciser ces difficultés, de les quantifier, et d'y répondre autant qu'il est possible.

2.1 Cadre fonctionnel

Dans ce cours, en ce qui concerne les méthodes générales, nous nous plaçons dans le cadre des **opérateurs linéaires sur les espaces de Hilbert**. Beaucoup peut être généralisé aux espaces de Banach et/ou aux opérateurs non linéaires, sous certaines conditions ; cependant par souci de clarté, nous nous limiterons à ce cadre théorique, qui permet d'exposer toutes les idées maîtresses dans un cadre simple.

Soit \mathcal{Y} et \mathcal{Z} deux espaces de Hilbert, dont on note les produits scalaires respectifs $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ et $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$, qui définissent des normes $\|\cdot\|_{\mathcal{Y}}$ et $\|\cdot\|_{\mathcal{Z}}$.

Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire continu et $\Psi^* : \mathcal{Z} \longrightarrow \mathcal{Y}$ son adjoint.¹ On appellera ici *problème direct* l'opérateur Ψ , et *problème inverse* la question : étant donnée $z \in \mathcal{Z}$, trouver $y \in \mathcal{Y}$ « satisfaisant » tel que $\Psi y = z$. Ainsi, ce problème « inverse » mérite bien son nom puisqu'il est lié à l'inversion de l'opérateur Ψ .

Plus précisément, on modélise *l'erreur de mesure* en notant z_ε une mesure bruitée, telle que $\|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$. Formellement, il s'agit donc de trouver $y_\varepsilon \in \mathcal{Y}$ « satisfaisant », *i.e.* vérifiant

1. On rappelle que l'adjoint Ψ^* d'un opérateur Ψ est défini grâce au théorème de représentation de Riesz, par

$$\forall (y, z) \in \mathcal{Y} \times \mathcal{Z}, \quad \langle \Psi y, z \rangle_{\mathcal{Z}} = \langle y, \Psi^* z \rangle_{\mathcal{Y}}$$

1. $\|\Psi y_\varepsilon - z_\varepsilon\|_{\mathcal{Z}} \leq C\varepsilon$, pour une constante $C > 0$ donnée,
2. $\|y_\varepsilon - y\|_{\mathcal{Y}} \rightarrow_{\varepsilon \rightarrow 0} 0$.

La première condition signifie que par rapport à l'erreur de mesure effectuée, on fait « aussi bien que possible » : avoir une distance inférieure à ε serait dénuée d'intérêt, puisque le « vrai » y , dont on sait seulement qu'il appartient à $B_{\mathcal{Z}}(z_\varepsilon, \varepsilon)$, n'en sera pas mieux approché. La quantité $\|\Psi y_\varepsilon - z_\varepsilon\|_{\mathcal{Z}}$ est appelée le résidu ou *discrepancy*.

La deuxième condition, qui sera précisée davantage ultérieurement, signifie que si l'erreur de mesure tend vers zéro alors la solution approchée y_ε converge bien vers « la » solution exacte y . En effet, comme nous allons le voir, l'ensemble des solutions y_ε qui vérifient $\|\Psi y_\varepsilon - z_\varepsilon\|_{\mathcal{Z}} \leq C\varepsilon$ n'est pas nécessairement borné.

Notre problème inverse est donc bien posé au sens d'Hadamard *ssi* Ψ est bijectif : en ce cas, la solution est $y := \Psi^{-1}z$ existe, est unique, et dépend continument de z . Du point de vue de la théorie général, nous considérerons le problème comme résolu : l'étude de l'opérateur Ψ^{-1} peut être très ardue, elle est spécifique. Grâce à la continuité de Ψ^{-1} , si l'on mesure z_ε au lieu de z , l'unique solution $y_\varepsilon = \Psi^{-1}z_\varepsilon$ convient puisque $\|y - y_\varepsilon\| = O(\varepsilon)$: l'erreur de mesure ne sera pas amplifiée par l'inversion. D'un point de vue numérique (l'opérateur Ψ étant approché par une matrice), il ne reste « que » à résoudre numériquement l'inversion de Ψ , ou du moins un système d'équations linéaires. Cela peut se faire par un grand nombre de méthodes d'algèbre linéaire : décomposition LU, élimination de Gauss-Jordan, méthode des co-facteurs, inversion par blocs, décomposition en éléments propres, etc.

Si l'opérateur Ψ n'est pas surjectif, le problème est double. Tout d'abord, l'existence d'une solution n'est pas garantie. Ensuite, même si la mesure z appartient à l'image de Ψ notée $Im(\Psi) \subsetneq \mathcal{Z}$, et qu'on puisse donc trouver y tel que $\Psi y = z$, l'inverse Ψ^{-1} n'est pas continu, et une petite erreur de mesure sur z peut conduire à une grande erreur sur y , voire à l'absence de solution, puisque l'on peut trouver ε aussi petit qu'on veut et $z_\varepsilon \in \mathcal{Z}$ tel que $\|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$ et $z_\varepsilon \notin Im(\Psi)$. En particulier, on pourra trouver y_ε avec $\|y_\varepsilon\|_{\mathcal{Y}} \rightarrow \infty$ avec $\|\Psi y_\varepsilon - z_\varepsilon\| \leq \varepsilon^2$.

Résoudre le problème inverse ne pourra alors se faire qu'en ajoutant *un a priori / une condition source* concernant y : est-il borné, satisfait-il une condition supplémentaire ? On notera cet a priori $y \in K \subset \mathcal{Y}$, et par la suite nous préciserons des ensembles \mathcal{K} convenables.

On note $z_\varepsilon \in \mathcal{Z}$ une mesure de $z = \Psi y$ satisfaisant $\|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$, $\mathcal{K} \subset \mathcal{Y}$ l'ensemble des conditions admissibles, et le problème inverse que nous souhaitons résoudre s'écrit donc

$$\text{Trouver } y_\varepsilon \in \mathcal{K} \subset \mathcal{Y}, \quad \|\Psi y_\varepsilon - z_\varepsilon\| \leq \varepsilon. \tag{2.1}$$

2. par l'absurde : soit $z \in \overline{Im(\Psi)} \setminus Im(\Psi)$ et $z_n \rightarrow z$, $z_n \in Im(\Psi)$, tous de norme égale à 1. Soit y_n tel que $\Psi y_n = z_n$. Si (y_n) est bornée, elle est faible-* compacte donc en extrayant une sous-suite on aurait $y_n \xrightarrow{*} y$, donc $\Psi y = z$ et cela contredit le fait que $z \notin Im(\Psi)$.

Pour choisir \mathcal{K} , un choix assez général est le suivant : soit $B : \mathcal{Y}_B \longrightarrow \mathcal{Z}_B$ linéaire continue avec \mathcal{Y}_B dense dans \mathcal{Y} et \mathcal{Z}_B un espace de Hilbert. Pour une constante $\delta > 0$ donnée, on définit

$$\mathcal{K} := \{y \in \mathcal{Y}_B, \quad \|By\|_{\mathcal{Z}_B} \leq \delta\} \quad (2.2)$$

Cela conduit à définir les solutions de notre problème inverse comme appartenant à l'espace

$$\{y \in \mathcal{Y}_B, \quad \|By\|_{\mathcal{Z}_B} \leq \delta \quad \|\Psi y - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon\}.$$

Classiquement, on choisit $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}$ et $B = Id$: cela impose une borne aux solutions. Un autre choix standard, pour $\mathcal{Y} = L^2$, est de prendre $\mathcal{Y}_B = \mathcal{Z}_B = H^s$ avec $s > 0$ et $B = Id$: cela impose une régularité à la solution x , ainsi qu'une borne sur sa norme. Ce choix est aussi lié au problème inverse considéré et à l'*echelle de Hilbert* correspondante - nous en reparlerons. Il reste à choisir, de façon constructive, une telle solution, et à étudier son comportement lorsque le bruit ε tend vers zéro. D'autres questions se posent :

- Quel lien entre δ et ε ?
- Comment choisir l'a priori sur y ? Y a-t-il moyen de se passer de cet a priori ?
- Que faire si l'on ignore le niveau de bruit ε ?

Toutes ces questions ont des réponses mathématiques qui vont se préciser au fur et à mesure du cours.

2.2 Pseudo-inverse de Moore-Penrose

Définition 1. Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire borné.

1. $y^{LS} \in \mathcal{Y}$ est dite solution au sens des moindres carrés (least-squares solution) de l'équation $\Psi y = z$ si

$$\|\Psi y^{LS} - z\|_{\mathcal{Z}} = \inf_{y \in \mathcal{Y}} \|\Psi y - z\|_{\mathcal{Z}}. \quad (2.3)$$

2. $y^{BA} \in \mathcal{Y}$ est dite meilleure solution approchée (best-approximate solution) si y est une solution au sens des moindres carrés de $\Psi y = z$ et

$$\|y^{BA}\|_{\mathcal{Y}} = \inf_{y^{LS} \text{ least squares sol. of } \Psi y = z} \|y^{LS}\|_{\mathcal{Y}}. \quad (2.4)$$

Le problème est tout d'abord de savoir si ces définitions ont une ou des solutions. Cela est lié à la définition du pseudo-inverse de Ψ .

Définition 2. Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire borné. Soit $\tilde{\Psi} = \Psi|_{Ker(\Psi)^\perp} : Ker(\Psi)^\perp \longrightarrow Im(\Psi)$, et soit $\tilde{\Psi}^{-1} : Im(\Psi) \longrightarrow Ker(\Psi)^\perp$ son inverse. Le pseudo-inverse de Moore-Penrose, noté Ψ^\dagger , est l'unique extension linéaire de $\tilde{\Psi}^{-1}$ à $D(\Psi^\dagger) := Im(\Psi) + Im(\Psi)^\perp$ vérifiant $Ker(\Psi^\dagger) = Im(\Psi)^\perp$.

$\tilde{\Psi}^{-1}$ est bien défini (parce que comme $\text{Ker}(\Psi)$ est fermé on a $\mathcal{Y} = \text{Ker}(\Psi) + \text{Ker}(\Psi)^\perp$ et la décomposition est unique - revoir si besoin la théorie des projections sur des espaces orthogonaux...), donc on définit Ψ^\dagger de façon unique pour $z \in D(\Psi^\dagger)$ grâce à son unique décomposition $z = z_1 + z_2$, $z_1 \in \text{Im}(\Psi)$ et $z_2 \in \text{Im}(\Psi)^\perp$, par $\Psi^\dagger z = \tilde{\Psi}^{-1} z_1$.

Proposition 1. Soit $P_{\text{Ker}(\Psi)}$ et $P_{\overline{\text{Im}(\Psi)}}$ les projecteurs orthogonaux respectivement sur $\text{Ker}(\Psi) \subset \mathcal{Y}$ (fermé) et $\overline{\text{Im}(\Psi)} \subset \mathcal{Z}$ (fermé). Alors $\text{Im}(\Psi^\dagger) = \text{Ker}(\Psi)^\perp$ et on a les quatre équations suivantes :

$$\Psi \Psi^\dagger \Psi = \Psi, \quad (2.5)$$

$$\Psi^\dagger \Psi \Psi^\dagger = \Psi^\dagger, \quad (2.6)$$

$$\Psi^\dagger \Psi = \text{Id} - P_{\text{Ker}(\Psi)}, \quad (2.7)$$

$$\Psi \Psi^\dagger = P_{\overline{\text{Im}(\Psi)} \cap D(\Psi^\dagger)}. \quad (2.8)$$

Une solution de l'équation (2.5) est dite inverse interne, de (2.6) inverse externe.

Proposition 2. Ψ^\dagger a un graphe fermé $\text{gr}(\Psi^\dagger)$. De plus, Ψ^\dagger est borné (donc continu) ssi $\text{Im}(\Psi)$ est fermé.

Preuve. Montrer que le graphe de Ψ^\dagger est fermé n'implique pas directement que Ψ^\dagger est continu, car le théorème du graphe fermé ne s'applique que pour une application définie sur un espace complet - ce qu'en général n'est pas $D(\Psi^\dagger)$, plus précisément, ce que $D(\Psi^\dagger)$ est ssi $\text{Im}(\Psi)$ est fermé, auquel cas $D(\Psi^\dagger) = \mathcal{Z}$. On montre que

$$\text{gr}(\Psi^\dagger) = \mathcal{U} := \text{Im}(\Psi)^\perp \times \{0\} + \mathcal{Z} \times \text{Ker}(\Psi)^\perp \cap \{(\Psi y, y), y \in \mathcal{Y}\}.$$

Cette égalité permet de conclure car il s'agit alors de la somme orthogonale d'espaces fermés dans $\mathcal{Z} \times \mathcal{Y}$. Tout d'abord, soit $(z, y) \in \text{gr}(\Psi^\dagger)$. On décompose $z = z_1 + z_2$ avec $z_1 \in \text{Im}(\Psi)$ et $z_2 \in \text{Im}(\Psi)^\perp$, donc $y = \tilde{\Psi}^{-1} z_1 \in \text{Ker}(\Psi)^\perp$. On peut donc écrire $(z, y) = (z_2, 0) + (\Psi y, y)$ avec $y \in \text{Ker}(\Psi)^\perp$ et on a $\text{gr}(\Psi^\dagger) \subset \mathcal{U}$. Réciprocurement, si on a $(z, y) = (z_2, 0) + (\Psi y, y) \in \mathcal{U}$, les mêmes équations montrent aisément que $(z, y) \in \text{gr}(\Psi^\dagger)$.

On en conclut que si $\text{Im}(\Psi)$ est fermé, alors $D(\Psi^\dagger) = \mathcal{Z}$ et comme $\text{gr}(\Psi^\dagger)$ est fermé on en déduit, par le théorème du graphe fermé, que Ψ^\dagger est continu. La réciproque est aussi évidente car l'image réciproque d'un fermé par une application continue est fermé, et $\text{Im}(\Psi) = (\Psi^\dagger)^{-1}(\mathcal{Y})$. ■

Cette construction fondamentale permet d'établir un lien entre les solutions au sens des moindres carrés et le pseudo-inverse de Moore-Penrose.

Théorème 2. Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire continu borné, Ψ^\dagger son pseudo-inverse et $z \in D(\Psi^\dagger)$. L'équation $\Psi y = z$ a une unique meilleure solution approchée qui est $y^\dagger := \Psi^\dagger z$. Les solutions au sens des moindres carrés sont l'ensemble $y^\dagger + \text{Ker}(\Psi)$.

Preuve. On décompose $z = z_1 + z_2$ avec $z_1 \in Im(\Psi)$ et $z_2 \in Im(\Psi)^\perp$. Soit $y^\dagger \in Ker(\Psi)^\perp$ tel que $\Psi y = z_1$, i.e. $y^\dagger = \Psi^\dagger z = \Psi^\dagger z_1$. Cela entraîne que $\|\Psi y^\dagger - z\|_{\mathcal{Z}} = \|z_2\|_{\mathcal{Z}}$.

Soit $\tilde{y} \in \mathcal{Y}$, $\tilde{y} = y_1 + y_2$ avec $y_1 \in Ker(\Psi)^\perp$ et $y_2 \in Ker(\Psi)$; on a

$$\|\Psi \tilde{y} - z\|_{\mathcal{Z}}^2 = \|\Psi y_1 - z_1 - z_2\|_{\mathcal{Z}}^2 = \|\Psi y_1 - z_1\|_{\mathcal{Z}}^2 + \|z_2\|_{\mathcal{Z}}^2 \geq \|z_2\|_{\mathcal{Z}}^2 = \|\Psi y^\dagger - z\|_{\mathcal{Z}}^2,$$

ce qui montre que y^\dagger est bien solution au sens des moindres carrés, et \tilde{y} est une autre solution au sens des moindres carrés ssi $\Psi y_1 = z_1$, i.e. $y_1 = y^\dagger$: l'ensemble des solutions au sens des moindres carrés est bien donné par $y^\dagger + Ker(\Psi)$, ensemble sur lequel l'élément de norme minimal est unique et est y^\dagger . ■

Une autre caractérisation, en lien avec toute la suite de notre étude, est donnée par l'équation suivante.

Théorème 3. Soit $z \in D(\Psi^\dagger)$. $y \in \mathcal{Y}$ est une solution au sens des moindres carrés ssi y est solution de l'équation dite normale

$$\Psi^* \Psi y = \Psi^* z. \quad (2.9)$$

Preuve. Soit y une solution au sens des moindres carrés. Alors d'après le théorème 2 $y = \tilde{\Psi}^{-1}z + y_2$, avec $y_2 \in Ker(\Psi)$. Donc $\Psi y = z_1$ avec $z = z_1 + z_2$, $z_1 \in Im(\Psi)$ et $z_2 \in Im(\Psi)^\perp$. Il reste à montrer que $\Psi^* z = \Psi^* z_1$, i.e. que $z_2 \in Ker(\Psi^*)$: on a, pour tout $\tilde{z} \in \mathcal{Y}$,

$$\langle \Psi^* z, \tilde{z} \rangle_{\mathcal{Y}} = \langle z, \Psi \tilde{z} \rangle_{\mathcal{Z}} = \langle z_1, \Psi \tilde{z} \rangle_{\mathcal{Z}} = \langle \Psi^* z_1, \tilde{z} \rangle_{\mathcal{Y}},$$

ce qui achève la preuve que y est solution de $\Psi^* \Psi y = \Psi^* z$. Réciproquement, soit y solution de $\Psi^* \Psi y = \Psi^* z$, alors $\Psi y - z \in Ker(\Psi^*)$, et on a de façon évidente que $Ker(\Psi^*) = Im(\Psi)^\perp$, donc $\Psi y - z \in Im(\Psi)^\perp$, donc par l'égalité (2.7) on a

$$\Psi^\dagger(\Psi y) = (Id - P_{Ker(\Psi)})(y) = \Psi^\dagger z,$$

et donc $y = \Psi^\dagger z + P_{Ker(\Psi)}(y)$, et le théorème 2 permet de conclure que y est une solution au sens des moindres carrés. ■

Ces résultats positifs permettent de montrer le lien entre inverse au sens des moindres carrés et inverse de Moore-Penrose, et qu'en ce sens donc l'inverse de Moore-Penrose est la meilleure estimation que l'on puisse imaginer.

Cependant,

- on peut aussi montrer que si $z \notin D(\Psi^\dagger)$ alors **il n'existe aucune solution au sens des moindres carrés**: il suffit de prendre $z \in \overline{Im(\Psi)}$, $z \notin Im(\Psi)$ et une suite $z_n \in Im(\Psi)$, $z_n \rightarrow z$. La suite $y_n = \Psi^\dagger z_n$ ne peut pas converger sans contredire le fait que $z \notin Im(\Psi)$, et pourtant $\|\Psi y_n - z\| \rightarrow 0$ donc une solution au sens des moindres carrés devrait vérifier $\|\Psi y - z\|_{\mathcal{Z}} = 0$, donc $\Psi y = z$, ce qui contredirait encore une fois $z \notin Im(\Psi)$.
- La définition de la *meilleure solution approchée* ne résout pas le problème de non-continuité puisque Ψ^\dagger n'est pas continu dans le cas général.

Nous allons donc dans le paragraphe suivant aborder le concept de régularisation.

Illustration : estimation de la dérivée d'une fonction

C'est l'exemple paradigmique qui sera repris tout au long de ce cours, que l'on formalisera davantage, et sur lequel on appliquera successivement diverses méthodes, afin de mieux en apprécier les caractéristiques, les points communs et les divergences. On a déjà vu, à travers les applications à la division cellulaire ou à la polymérisation des protéines, comment ce problème pouvait apparaître naturellement dans de nombreux problèmes inverses.

Dans le cadre fonctionnel ci-dessus, on peut préciser les choses : on prend $\mathcal{Y} = \mathcal{Z} = L^2([0, 1], dx)$, muni du produit scalaire dans L^2 , et pour $y \in \mathcal{Y}$ on définit

$$\Psi y : [0, 1] \rightarrow \mathbb{R}, \quad t \mapsto \int_0^t y(s) ds. \quad (2.10)$$

Le problème inverse est : étant donné $z \in \mathcal{Z}$, trouver $y \in \mathcal{Y}$ tel que

$$\forall t \in [0, 1], \quad \int_0^t y(s) ds = z(t).$$

On vérifie aisément que Ψ est linéaire, continue dans L^2 :

$$\begin{aligned} \left\| \left(\int_0^t y(s) ds \right) \right\|_{\mathcal{Z}}^2 &= \int_0^1 \left(\int_0^t y(s) ds \right)^2 dt \leq \int_0^1 t \int_0^t y(s)^2 ds dt \\ &= \int_0^1 y(s)^2 \int_s^1 t dt ds = \int_0^1 y(s)^2 \left(\frac{1}{2} - \frac{s^2}{2} \right) ds \leq \frac{1}{2} \|y\|_{\mathcal{Y}}^2, \end{aligned}$$

donc $\|\Psi\|_{\mathcal{Y} \rightarrow \mathcal{Z}} \leq \frac{1}{\sqrt{2}}$. **Exercice : quelle est la norme de Ψ ?**

$Im(\Psi)$ est dense dans L^2 par densité de $C_0^1([0, 1]) \subset Im(\Psi)$. Mais $Im(\Psi) \subsetneq \mathcal{Z}$ puisque la dérivée d'une fonction L^2 au sens des distributions n'est pas dans L^2 mais dans H^{-1} .

Supposons maintenant que l'on prenne $\mathcal{Z} = Im(\Psi)$ **Exercice : comment caractériser $Im(\Psi)$?** muni de la norme H^1 : cette fois, le problème inverse est bien posé. **Tout est une question de norme : il n'existe pas de problème intrinsèquement bien ou mal posé.**

En revanche, ce qui est caractéristique des problèmes inverses est que l'espace dans lequel on doit travailler n'est pas imposé uniquement par des raisons de commodité mathématique mais d'abord par des considérations expérimentales. Ici la question qui nous guide sera donc : **à quel espace naturel appartient la mesure z_ε ? Dans quelle norme est-elle ε -proche de la « vraie » valeur $z = \Psi y$?**

En fonction de la réponse, le problème considéré sera plus ou moins « mal posé ».

2.3 Régularisation d'un problème mal posé

Par la suite on supposera souvent Ψ linéaire compact, injectif, non surjectif, à image dense dans \mathcal{Z} . En effet, à la suite de l'étude ci-dessus, on fait les deux remarques suivantes.

- Si l'opérateur Ψ n'est pas injectif, l'unicité d'une solution ne sera plus vérifiée : il y aura un *déficit d'information*, qui pour être levé nécessitera d'enrichir le problème en ajoutant une autre caractérisation. On peut se restreindre par exemple à l'espace quotient $\mathcal{Y}/\text{Ker}(\Psi)$, $\text{Ker}(\Psi)$ étant le noyau de Ψ . On se limite donc pour l'instant au cas d'un opérateur Ψ injectif, considérant que l'étude consistant à étudier le noyau de Ψ fait partie des études spécifiques à chaque problème inverse, et qu'une fois connu $\text{Ker}(\Psi)$ il n'y a pas grand'chose d'autre à faire que de se placer sur l'espace-quotient ou d'enrichir le problème en ajoutant un a priori.
- si l'image de Ψ n'est pas dense, de même que pour le cas où l'opérateur n'est pas injectif, on peut décomposer \mathcal{Z} en

$$\mathcal{Z} = \overline{\text{Im}(\Psi)} + \overline{\text{Im}(\Psi)}^\perp,$$

et on décompose une mesure z_ε en sa projection sur chacun de ces sous-espaces. Pour la projection de la mesure sur $\overline{\text{Im}(\Psi)}^\perp$, on ne pourra rien faire...

Définition 3. Soit \mathcal{Y}, \mathcal{Z} deux espaces de Hilbert et $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire borné, injectif, à image dense. On appelle famille régularisante pour Ψ , une famille d'opérateurs (pas nécessairement linéaires) continus $R_\alpha : \mathcal{Z} \longrightarrow \mathcal{Y}$, définis pour $\alpha > 0$, tels que

$$\forall y \in \mathcal{Y}, \quad \lim_{\alpha \rightarrow 0} R_\alpha \Psi y = y. \quad (2.11)$$

Notons tout d'abord que l'opérateur inverse $\Psi^{-1} : \text{Im}(\Psi) \longrightarrow \mathcal{Y}$ est non borné³, donc il existe par exemple une suite $(z_n = \Psi y_n)_{n \in \mathbb{N}}$ telle que $\|y_n\|_{\mathcal{Y}} = 1$ et $\|z_n\|_{\mathcal{Z}} \rightarrow 0$.

Exemple : pour notre opérateur d'intégration vu ci-dessus et $\mathcal{Y} = \mathcal{Z} = L^2$, on peut prendre la suite $y_n = \cos(2\pi nx)$, $z_n = \frac{1}{2\pi n} \sin(2\pi nx)$.

Cela implique nécessairement que la famille (R_α) est non bornée : $\sup_\alpha \|R_\alpha\| = \infty$. En effet, prenant une telle suite $(y_n, \Psi y_n)$ et lui appliquant la relation (2.11), on aboutirait à la contradiction $y_n \rightarrow 0$ si (R_α) était bornée).

Revenons maintenant à notre « problème inverse » : mesurant z_ε tel que $\|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$, en supposant que $z \in \text{Im}(\Psi)$ avec $\Psi y = z$, de bons candidats pour approcher y sont donnés par la famille $y_{\varepsilon, \alpha}$ définie par

$$y_{\varepsilon, \alpha} := R_\alpha z_\varepsilon, \quad y_\alpha := R_\alpha z. \quad (2.12)$$

3. En effet, si Ψ^{-1} était borné il serait continu, et donc on pourrait le prolonger par continuité, ce qui entraînerait que $\text{Im}(\Psi)$ serait fermé.

Le but est de trouver le meilleur candidat possible pour estimer y . On peut décomposer l'erreur d'estimation comme suit :

$$\begin{aligned}\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}} &\leq \|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{Y}} + \|y_\alpha - y\|_{\mathcal{Y}} \\ &\leq \varepsilon \|R_\alpha\| + \|R_\alpha \Psi y - y\|_{\mathcal{Y}}.\end{aligned}\quad (2.13)$$

(la deuxième inégalité n'est vraie que si R_α est linéaire) Le premier terme est petit à cause de ε , mais quand $\alpha \rightarrow 0$, il explose, au contraire du second terme qui tend vers 0 quand $\alpha \rightarrow 0$ en vertu de (2.11). Il faut donc choisir α tendant vers 0 quand $\varepsilon \rightarrow 0$, mais pas trop vite cependant, de façon à garantir simultanément $\varepsilon \|R_\alpha\| \rightarrow 0$ et $\|R_\alpha \Psi y - y\|_{\mathcal{Y}} \rightarrow 0$. **Ce type d'équilibre entre deux termes et d'effets contradictoires se retrouve partout dans les problèmes inverses.** Si on régularise trop, le second terme ne tendra pas assez vite vers 0. Au contraire, si on ne régularise pas assez, le premier terme explose. Quand nous verrons ce type de découpage en statistiques, il y a une analogie avec le découpage « biais » (le second terme : on a en effet *biaisé* le problème en considérant la régularisation) / « variance » (le premier : celui-ci est lié à l'erreur de mesure expérimentale).

On en vient à la définition suivante d'une stratégie de régularisation.

Définition 4. Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire borné de \mathcal{Y} dans \mathcal{Z} , \mathcal{Y} et \mathcal{Z} deux espaces de Hilbert. On appelle méthode de régularisation (convergente) une paire (R_α, α) où R_α est une famille régularisante comme définie par la définition 3 et $\alpha(\varepsilon, z_\varepsilon)$ est une règle définissant une suite de paramètres α telle que pour tout $z \in \text{Im}(\Psi)$, $\Psi^\dagger z = y$, on ait

$$\limsup_{\varepsilon \rightarrow 0, z_\varepsilon \in \mathcal{Z}, \|z_\varepsilon - z\|_{\mathcal{Z}} \leq \varepsilon} \|R_{\alpha(\varepsilon, z_\varepsilon)} z_\varepsilon - y\|_{\mathcal{Y}} = 0. \quad (2.14)$$

Si cette règle ne dépend que de ε et non de z_ε , on parle de régularisation a priori, sinon de régularisation a posteriori.

Du découpage (2.13) on déduit facilement la proposition suivante.

Proposition 3. Soit R_α une famille régularisante telle que définie par la définition 3. Il existe une méthode de régularisation a priori convergente.

Preuve. Nous montrons qu'une telle suite existe de façon non constructive. Soit $z \in \text{Im}(\Psi)$ et $y \in \mathcal{Y}$ tel que $\Psi y = z$. Il faut prouver qu'il existe une fonction $\alpha(\varepsilon)$ telle que pour tout $\eta > 0$, il existe $\varepsilon > 0$ assez petit pour que $\|z_\varepsilon - z\| \leq \varepsilon$ implique $\|R_{\alpha(\varepsilon)} z_\varepsilon - y\| \leq \eta$. Soit $\eta > 0$. On commence par le terme de « biais » dans (2.13) : soit la fonction $\tilde{\alpha}(\eta)$ telle que

$$\|R_{\tilde{\alpha}(\eta)} z - y\|_{\mathcal{Y}} \leq \frac{\eta}{2},$$

ce qui est possible par définition de R_α . Le premier terme est alors inférieur à

$$\|R_{\tilde{\alpha}(\eta)} z_\varepsilon - R_{\tilde{\alpha}(\eta)} z\|_{\mathcal{Y}},$$

et comme $R_{\tilde{\alpha}(\eta)}$ est continue, il existe $g(\eta)$ assez petit tel que $\|z_{g(\eta)} - z\| \leq g(\eta)$ implique

$$\|R_{\tilde{\alpha}(\eta)}z_{g(\eta)} - R_{\tilde{\alpha}(\eta)}z\|_{\mathcal{Y}} \leq \frac{\eta}{2}.$$

On peut alors choisir une telle fonction g croissante et tendant vers 0 quand η tend vers 0, ce qui permet de définir son inverse g^{-1} , et de choisir comme règle $\alpha := \tilde{\alpha}(g^{-1}(\varepsilon))$. ■

La question qui se pose maintenant est comment choisir une telle règle de façon constructive et optimale ? Avant de passer à des résultats positifs, il faut être conscient de leurs limites, et en particulier des deux faits suivants :

- si l'on ne connaît pas le niveau de bruit, il n'y a aucun moyen d'avoir une méthode convergente.
- Si l'on n'impose aucun a priori sur l'espace où se situe y , la convergence peut être arbitrairement lente.

Cela est prouvé dans les deux résultats suivants, dont la preuve est élémentaire.

Proposition 4 (théorème 3.3. de [27], Bakushinskii [3]). *Soit $\Psi : \mathcal{Y} \rightarrow \mathcal{Z}$ un opérateur linéaire borné. S'il existe une régularisation R_α où le choix de α ne dépende que de z_ε et non de ε , alors Ψ^\dagger est borné.*

Preuve. soit une telle méthode de régularisation et $z \in Im(\Psi)$ avec $\Psi y = z$. D'après (2.14) on a

$$\limsup_{\varepsilon \rightarrow 0, z_\varepsilon \in \mathcal{Z}, \|z_\varepsilon - z\|_{\mathcal{Z}} \leq \varepsilon} \|R_{\alpha(z_\varepsilon)}z_\varepsilon - y\|_{\mathcal{Y}} = 0.$$

Donc en particulier, prenant $z_\varepsilon = z$ et faisant tendre ε vers 0, on a $R_{\alpha(z)}z = \Psi^\dagger z$ pour tout $z \in D(\Psi^\dagger)$. Si on prend une suite $z_n \in Im(\Psi)$ avec $z_n \rightarrow z \in Im(\Psi)$, on en déduit que $\Psi^\dagger z_n = R_{\alpha(z_n)}z_n \rightarrow \Psi^\dagger z$, ce qui signifie que Ψ^\dagger est continu sur $Im(\Psi)$ donc y est borné, donc on peut le prolonger par continuité à $\overline{Im(\Psi)}$ donc à \mathcal{Z} entier, donc (proposition 2) Ψ^\dagger est continu borné sur \mathcal{Z} . (NB : pour montrer que Ψ^\dagger continu sur $Im(\Psi)$ implique Ψ^\dagger borné sur $Im(\Psi)$: on peut par exemple raisonner par l'absurde, prendre une suite $\tilde{z}_n \in Im(\Psi)$, $\|\tilde{z}_n\|_{\mathcal{Z}} = 1$ et $\|\Psi^\dagger \tilde{z}_n\|_{\mathcal{Z}} \geq n$, donc par linéarité on peut aussi prendre $z_n = \frac{\tilde{z}_n}{n}$, ce qui veut dire que $z_n \rightarrow 0 \in Im(\Psi)$, donc par continuité $\Psi^\dagger z_n \rightarrow 0$, ce qui contredit le fait que $\|\Psi^\dagger z_n\|_{\mathcal{Z}} \geq 1$.) ■

Ceci montre l'impossibilité d'une méthode dite *error free* ; nous verrons cependant que des méthodes empiriques, comme la méthode de la courbe en L (*L-curve*, voir [10]) existent. Le résultat suivant montre, quant à lui, qu'aucune vitesse de convergence ne peut être obtenue en l'absence d'a priori supplémentaire sur la condition initiale.

Proposition 5 (Prop. 3.11. de [27], voir aussi [53]). *Soit Ψ un opérateur linéaire borné injectif continu à image dense non fermée et (R_α, α) une méthode de régularisation. Il*

n'existe pas de fonction $f : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ croissante avec $\lim_{\varepsilon \rightarrow 0} f(\varepsilon) = 0$ telle que

$$\forall z \in Im(\Psi), \|z\|_{\mathcal{Z}} \leq 1, \|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon, \Rightarrow \|R_{\alpha(\varepsilon, z_\varepsilon)} z_\varepsilon - \Psi^\dagger z\|_{\mathcal{Y}} \leq f(\varepsilon).$$

Preuve. On raisonne par l'absurde : supposons qu'il existe une telle fonction f . On va montrer qu'alors Ψ^{-1} est continue - ce qui est faux comme déjà vu. Soit $z \in Im(\Psi)$, $\|z\|_{\mathcal{Z}} = 1$, et $z_k \rightarrow z$ avec $z_k \in Im(\Psi)$. On note $y_k = \Psi^{-1} z_k$. Soit $\eta > 0$ et $\varepsilon = f^{-1}(\eta)$. Pour k assez grand, on a $\|z_k - z\|_{\mathcal{Z}} \leq \varepsilon$, donc $\tilde{y}_k := R_{\alpha(\varepsilon, z_k)} z_k$ vérifie $\|\tilde{y}_k - y\|_{\mathcal{Y}} \leq \eta$. On a alors

$$\|y_k - y\|_{\mathcal{Y}} \leq \|y_k - \tilde{y}_k\|_{\mathcal{Y}} + \|\tilde{y}_k - y\|_{\mathcal{Y}} \leq 2\eta,$$

en appliquant également la proposition à z_k . Donc $y_k \rightarrow y$, donc Ψ^{-1} est séquentiellement continue donc continue. ■

Ce résultat montre que pour pouvoir obtenir des vitesses de convergence de nos méthodes, il faut **préciser des informations sur y** : cette importance des termes sources montre qu'il y a un équilibre à faire entre la confiance qu'on a dans la mesure (symbolisée par ε) et la confiance qu'on a dans la source, symbolisée par le second terme dans (2.13). Si, dans (2.2), on avait pris comme condition source $B = Id_{\mathcal{Y}}$, on peut réécrire le découpage (2.13) de sorte qu'il fasse apparaître cette condition : on aura

$$\|y_{\varepsilon, \alpha} - y\|_{\mathcal{Y}} \leq \varepsilon \|R_\alpha\| + C \|R_\alpha \Psi - Id\|_{\mathcal{Y} \rightarrow \mathcal{Y}}.$$

De manière plus générale, formellement :

$$\|y_{\varepsilon, \alpha} - y\|_{\mathcal{Y}} \leq \varepsilon \|R_\alpha\| + C \|(R_\alpha \Psi - Id)B^{-1}\|_{\mathcal{Z} \rightarrow \mathcal{Y}}.$$

On voit ainsi apparaître encore une fois le rôle décisif joué d'une part par la valeur de l'a priori δ , qui contre-balance ε , et d'autre part l'importance de la norme choisie pour cet a priori, qui déterminera la vitesse de convergence vers 0 du second terme, lié à la régularisation. Nous y reviendrons.

Exemple : pour notre opérateur d'intégration Ψ , une méthode bien connue est celle des **suites régularisantes (mollifier)**. Soit $\rho \in \mathcal{C}_0^\infty(\mathbb{R})$, avec $\int \rho(x)dx = 1$, on définit $\rho_\alpha(x) := \frac{1}{\alpha} \rho(\frac{x}{\alpha})$, et pour $z_\varepsilon \in \mathcal{Z} = L^2([0, 1])$ on définit $R_\alpha := \rho'_\alpha *$ et donc

$$y_{\varepsilon, \alpha} := \rho_\alpha * z'_\varepsilon = \rho'_\alpha * z_\varepsilon.$$

Exercice : montrer que R_α est bien une famille régularisante au sens de la définition 3.

Résolution : soit $y \in L^2([0, 1])$. On a

$$R_\alpha \Psi y = \rho'_\alpha * \left(t \rightarrow \int_0^t y(s)ds \right) = \rho_\alpha * y,$$

et on sait que $\|\rho_\alpha * y - y\|_{L^2} \rightarrow_{\alpha \rightarrow 0} 0$. Cela peut se vérifier aussi très simplement par transformée de Fourier puis en découplant l'intégrale. En revanche, si l'on ne suppose pas plus de régularité sur x , on n'aura pas de taux de convergence : pour cela, il faut par exemple supposer $y \in H^1([0, 1])$.

Définition 5. Soit $\mathcal{K} \subset \mathcal{Y}$. On note

$$\Omega(\varepsilon, \mathcal{K}) := \sup_{y \in \mathcal{K}, \|\Psi y\|_{\mathcal{Z}} \leq \varepsilon} \|y\|_{\mathcal{Y}} \quad (2.15)$$

et pour une méthode de régularisation R , on note

$$\Delta(\varepsilon, \mathcal{K}, R) := \sup_{y \in \mathcal{K}, z_\varepsilon \in \mathcal{Z}, \|\Psi y - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon} \|Rz_\varepsilon - y\|_{\mathcal{Y}} \quad (2.16)$$

Une méthode R_0 sera dite optimale sur \mathcal{K} si $\Delta(\varepsilon, \mathcal{K}, R_0) = \inf_R \Delta(\varepsilon, \mathcal{K}, R)$ (parmi une classe de méthodes).

On a une borne inférieure sur toute méthode donnée par $\Omega(\varepsilon, \mathcal{K})$. Prenons en effet $z_\varepsilon = 0$: comme $R0 = 0$, il est évident que

$$\Omega(\varepsilon, \mathcal{K}) \leq \Delta(\varepsilon, \mathcal{K}, R_0).$$

Pour ce qui est des méthodes dans toute leur généralité, nous nous arrêtons ici. Nous verrons ensuite la *décomposition en valeurs singulières* (Singular Value Decomposition ou SVD), outil fondamental et très efficace pour l'analyse des opérateurs compacts donc très utile en pratique (puisque la discrétisation numérique oblige à se placer en dimension finie, dans un cadre matriciel). Ce n'est pas la méthode la plus efficace en terme de temps de calcul, en revanche...

Chapitre 3

Décomposition d'un opérateur compact en valeurs singulières

La décomposition en valeurs singulières (ou SVD : singular value decomposition) peut être vue comme une généralisation de la décomposition spectrale pour des opérateurs compacts non auto-adjoints, ou, dans le cadre matriciel, pour des matrices rectangulaires non diagonalisables.

C'est une méthode très utilisée par exemple en analyse d'image, et qui a de nombreux liens avec les autres méthodes que nous verrons par la suite, par exemple la minimisation par moindres carrés. Historiquement, elle fut tout d'abord développée en géométrie différentielle, et les premiers travaux remontent à la fin du XIX^e siècle (pour les matrices). Séparément, il y eut des travaux pour les opérateurs intégraux : Ehrhard Schmidt en 1907, Emile Picard en 1910. Pour une revue historique, se référer à une conférence de G.W. Stewart de 1992, *On the Early History of the Singular Value Decomposition*, disponible sur internet (archives wikipedia).

Je me place ici dans le cadre d'un opérateur linéaire compact $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$, \mathcal{Y} et \mathcal{Z} deux espaces de Hilbert. Dans la pratique de l'analyse numérique, il s'agira toujours de calcul matriciel (\mathcal{Y} et \mathcal{Z} de dimension finie), mais le cadre des opérateurs permet de voir les résultats de façon intrinsèque, affranchie de la description dans une base donnée.

De plus, on suppose que $Im(\Psi)$ est dense dans \mathcal{Z} .

3.1 Décomposition en valeurs singulières : Définition et premières propriétés

Théorème 4 (Décomposition en valeurs singulières). *Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur compact injectif et \mathcal{Y} de dimension infinie. Alors il existe des suites $(e^j)_{j \in \mathbb{N}}$, $(f^j)_{j \in \mathbb{N}}$, et $(\sigma^j)_{j \in \mathbb{N}}$, appelé système singulier, tel que*

1. $e^j \in \mathcal{Y}, f^j \in \mathcal{Z}, \forall j \in \mathbb{N},$
2. $0 < \sigma_{j+1} \leq \sigma_j, \forall j \in \mathbb{N}, \lim_{j \rightarrow \infty} \sigma_j = 0,$
3. $\langle e^j, e^k \rangle_{\mathcal{Y}} = \delta_{j=k}, \langle f^j, f^k \rangle_{\mathcal{Z}} = \delta_{j=k} \quad \forall j, k \in \mathbb{N}, j \neq k,$
4. $\Psi e^j = \sigma_j f^j, \Psi^* f^j = \sigma_j e^j, \forall j \in \mathbb{N},$
- 5.

$$\Psi y = \sum_{j=0}^{\infty} \sigma_j \langle y, e^j \rangle_{\mathcal{Y}} f^j, \quad \Psi^* z = \sum_{j=0}^{\infty} \sigma_j \langle z, f^j \rangle_{\mathcal{Z}} e^j, \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}.$$

Preuve. La décomposition en valeurs singulières est une forme de généralisation de la décomposition dans une base orthonormée de vecteurs propres pour un opérateur auto-adjoint compact.

Rappel : soit $K : \mathcal{Y} \longrightarrow \mathcal{Y}$ un opérateur linéaire auto-adjoint compact. Alors l'espace $\overline{Im(K)}$ admet une base orthonormale (e_n) formée par les vecteurs propres de K associés aux valeurs propres non nulles, et on a

$$Ky = \sum_{n=0}^{\infty} \lambda_n \langle y, e_n \rangle_{\mathcal{Y}} e_n,$$

où la suite des valeurs propres non nulles (λ_n) tend vers 0.

On applique la décomposition spectrale à l'opérateur auto-adjoint compact et positif $K = \Psi^* \Psi$, dont on note les vecteurs propres orthonormés (e^j) associés à des valeurs propres non nulles $\sigma_j^2 > 0$, car

$$\langle \Psi^* \Psi e^j, e^j \rangle_{\mathcal{Y}} = \sigma_j^2 = \langle \Psi e^j, \Psi e^j \rangle_{\mathcal{Z}} > 0.$$

On définit alors $f^j = \frac{1}{\sigma_j} \Psi e^j$, et on vérifie sans peine toutes les propriétés requises. ■

Corollaire 1 (Critère de Picard). *Soit $\Psi : \mathcal{Y} \longrightarrow \mathcal{Z}$ un opérateur linéaire compact injectif à image dense et $(e^j, f^j, \sigma_j)_{j \in \mathbb{N}}$ son système de valeurs singulières. Alors $z \in D(\Psi^\dagger)$ ssi*

$$\sum_{j=0}^{\infty} \sigma_j^{-2} \langle z, f^j \rangle_{\mathcal{Z}}^2 < \infty, \tag{3.1}$$

auquel cas on a

$$\Psi^\dagger z = \sum_{j=0}^{\infty} \sigma_j^{-1} \langle z, f^j \rangle_{\mathcal{Z}} e^j. \tag{3.2}$$

Preuve. Si l'on a (3.1), alors la définition de y par (3.2) est possible, et on vérifie de façon immédiate que $\Psi y = z$. Réciprocurement, si y est tel que $\Psi y = z$, alors en écrivant $y = \sum \langle y, e^j \rangle e^j$, on a $\langle \Psi y, f^j \rangle = \langle y, e^j \rangle \sigma_j = \langle z, f^j \rangle$, donc $\langle y, e^j \rangle = \frac{1}{\sigma_j} \langle z, f^j \rangle$, et comme $\sum \langle y, e^j \rangle^2 < \infty$ on obtient (3.1). ■

Exercice : trouver le système singulier pour l'opérateur intégral Ψ défini par (2.10), pour $\mathcal{Y} = L^2$ et $\mathcal{Z} = L^2$ puis pour $\mathcal{Y} = L^2$ et $\mathcal{Z} = H^s$ avec $s \in \mathbb{R}$.

Résolution : il faut d'abord prouver que Ψ est compact, puis définir Ψ^* , enfin chercher les solutions de l'équation $\Psi^* \Psi y = \lambda y$.

Prouver que Ψ est compact peut se faire par exemple par le plongement compact de H^1 dans L^2 (inégalités de Sobolev). Autres méthodes : alternative de Fredholm. Voir [20, 21] comme livre de référence en analyse fonctionnelle pour les EDP.

Définition de Ψ^* : soit $y, z \in L^2([0, 1])$. On a

$$\langle \Psi y, z \rangle_{\mathcal{Z}} = \int_0^1 z(t) \int_0^t y(s) ds dt = \int_0^1 y(s) \int_s^1 z(t) dt ds = \langle y, \Psi^* z \rangle_{\mathcal{Y}}$$

$$\text{donc } \Psi^* z(s) = \int_s^1 z(t) dt. \text{ Donc } \Psi^* \Psi y(s) = \int_s^1 \int_0^t y(\sigma) d\sigma dt.$$

Solutions de l'équation $\Psi^* \Psi y = \lambda y$ avec $\lambda \neq 0$: en dérivant, on obtient tout d'abord

$$\lambda y'(t) = - \int_0^t y(\sigma) d\sigma, \quad y(1) = 0,$$

puis en dérivant une seconde fois

$$\lambda y''(t) = -y(t), \quad y'(0) = 0, \quad y(1) = 0.$$

Cela entraîne classiquement $y(t) = A \cos(\sqrt{\lambda^{-1}} t) + B \sin(\sqrt{\lambda^{-1}} t)$, et comme $y'(0) = 0$ on a $B = 0$, et comme $y(1) = 0$ on a $\sqrt{\lambda^{-1}} = \frac{1}{\sigma_j} = \frac{\pi}{2} + j\pi$ avec $j \in \mathbb{N}$.

$$\text{Finalement on a } \sigma_j = \frac{2}{\pi(2j+1)}, \quad e^j(t) = \sqrt{2} \cos(\sigma_j^{-1} t), \quad f^j(t) = \sqrt{2} \sin(\sigma_j^{-1} t).$$

La vitesse de décroissance vers zéro de la suite des valeurs singulières σ_j caractérise le degré auquel un problème est mal posé - en référence, toujours, à deux espaces donnés \mathcal{Y} et \mathcal{Z} : en effet, si $z = f^j$, la solution du problème inverse $\Psi y = z$ est donnée par $y = \frac{e^j}{\sigma_j}$ qui tend vers l'infini quand j tend vers l'infini, alors que la norme de f^j reste égale à 1. Plus cette convergence est rapide, plus forte est la non-continuité de la dépendance par rapport aux données mesurées. On dit que le degré de problème mal posé (*degree of ill-posedness*)

est n si $\sigma_j = O(j^{-n})$. Si la décroissance de σ_j est en $o(j^{-n})$ pour tout n , on parle de problème sévèrement mal posé (*severely ill-posed problems*).

Exercice : soit l'application $\Psi = Id$, avec les espaces $\mathcal{Z} = L^2([0, 1])$ et $\mathcal{Y} = \{f \in H^1([0, 1]), f(0) = 0\}$ muni du produit scalaire

$$\langle y_1, y_2 \rangle_{\mathcal{Y}} = \int_0^1 y'_1(s) y'_2(s) ds,$$

qui définit bien une norme sur \mathcal{Y} par les inégalités de Poincaré. Montrer que $\Psi : \mathcal{Y} \rightarrow \mathcal{Z}$ est compact. Quel est Ψ^* ? Quelle est la décomposition en valeurs singulières de Ψ ? Comment l'interpréter?

$\Psi = Id : \mathcal{Y} \rightarrow \mathcal{Y}$ est-elle compacte?

Pour aller plus loin : Reprendre la démonstration précédente pour Ψ l'opérateur intégral défini par (2.10), $\mathcal{Y} = L^2([0, 1])$ et $\mathcal{Z} = H^s([0, 1])$ avec $s < \frac{1}{2}$. Pourquoi faut-il supposer $s < \frac{1}{2}$? Faire les calculs complets pour $s = -\frac{1}{2}$.

3.2 Suites régularisantes avec la décomposition en valeurs singulières

Par la formule du critère de Picard, on voit que les erreurs de mesure faites selon les directions f^j avec j grand, i.e. σ_j petit, sont amplifiées, et ce d'autant plus que la décroissance des σ_j vers zéro est rapide. On voit ainsi naturellement apparaître une idée fondamentale : celle de **filtrer** les contributions selon certaines directions. La façon la plus simple est de tronquer : définir $y_{\varepsilon, \alpha}$ par

$$y_{\varepsilon, \alpha} = R_{\alpha} z_{\varepsilon} = \sum_{j=0}^{N_{\alpha}} \sigma_j^{-1} \langle z_{\varepsilon}, f^j \rangle_{\mathcal{Z}} e^j, \quad (3.3)$$

et choisir N_{α} de façon appropriée : en particulier, $N_{\alpha} \rightarrow \infty$ quand $\varepsilon \rightarrow 0$ de sorte que $y_{\varepsilon} \rightarrow y$. Plus généralement, on peut définir

$$R_{\alpha} z := \sum_{j=0}^{\infty} r(\alpha, \sigma_j) \sigma_j^{-1} \langle z, f^j \rangle_{\mathcal{Z}} e^j, \quad \forall y \in \mathcal{Z}. \quad (3.4)$$

On a le théorème suivant.

Théorème 5. Soit $\Psi : \mathcal{Y} \rightarrow \mathcal{Z}$ un opérateur compact injectif de système singulier $(e^j, f^j, \sigma_j)_{j \in \mathbb{N}}$ et soit $r : (0, 1] \times (0, \sigma_0] \rightarrow [-1, 1]$ une fonction de filtre continue vérifiant les conditions suivantes

1. $\forall \alpha \in (0, 1), \exists c(\alpha) > 0$ tel que $\frac{|r(\alpha, \sigma)|}{\sigma} \leq c(\alpha)$, $\forall \sigma \in (0, \sigma_0]$,
2. $\lim_{\alpha \rightarrow 0} r(\alpha, \sigma) = 1 \forall \sigma \in (0, \sigma_0]$.

Alors, la famille R_α définie par (3.4) est une famille régularisante. De plus, pour $y_{\varepsilon, \alpha}$ défini par (3.3), on a l'inégalité :

$$\|y_{\varepsilon, \alpha} - y\|_{\mathcal{Y}}^2 \leq \varepsilon c(\alpha) + \sup_{\sigma} |r(\alpha, \sigma) - 1| \|y\|_{\mathcal{Y}}. \quad (3.5)$$

Preuve. La première condition nous assure de la bonne définition de R_α , borné, sur \mathcal{Z} , puisque $\|R_\alpha\|_{\mathcal{Z} \rightarrow \mathcal{Y}} \leq c(\alpha)$. Il reste à montrer (2.11). Soit $y \in \mathcal{Y}$ et $\varepsilon > 0$; on veut montrer que pour α assez petit, on a $\|R_\alpha \Psi y - y\|_{\mathcal{Y}} \leq \varepsilon$. On a par définition

$$R_\alpha \Psi y = \sum_{j=0}^{\infty} r(\alpha, \sigma_j) \sigma_j^{-1} \langle \Psi y, f^j \rangle_{\mathcal{Z}} e^j = \sum_{j=0}^{\infty} r(\alpha, \sigma_j) \langle y, e^j \rangle_{\mathcal{Y}} e^j,$$

donc

$$\|R_\alpha \Psi y - y\|_{\mathcal{Y}} = \sum_{j=0}^{\infty} \left(r(\alpha, \sigma_j) - 1 \right)^2 \langle y, e^j \rangle_{\mathcal{Y}}^2 \leq \sum_{j=0}^N \left(r(\alpha, \sigma_j) - 1 \right)^2 \langle y, e^j \rangle_{\mathcal{Y}}^2 + 4 \sum_{j=N+1}^{\infty} \langle y, e^j \rangle_{\mathcal{Y}}^2.$$

Soit tout d'abord N assez grand pour avoir $\sum_{j \geq N+1} \langle y, e^j \rangle_{\mathcal{Y}}^2 \leq \frac{\varepsilon^2}{8}$. Par continuité de r et grâce à la seconde condition, on choisit alors α_0 assez petit pour que

$$\forall \alpha \leq \alpha_0, \forall \sigma \in \{\sigma_N, \dots, \sigma_0\}, \quad \left(r(\alpha, \sigma) - 1 \right)^2 \leq \frac{\varepsilon^2}{2\|x\|_{\mathcal{Y}}^2},$$

par convergence de $r(\alpha, \sigma)$ vers 1 pour tout $\sigma \in \{\sigma_N, \dots, \sigma_0\}$. Ceci entraîne finalement que $\|R_\alpha \Psi y - y\|_{\mathcal{Y}} \leq \varepsilon$.

R_α est donc bien une famille régularisante. Soit maintenant $z_\varepsilon \in \mathcal{Z}$ tel que $\|z - z_\varepsilon\| \leq \varepsilon$. On définit $y_{\varepsilon, \alpha} = R_\alpha z_\varepsilon$ et on a, d'après (2.13)

$$\begin{aligned} \|y_{\varepsilon, \alpha} - y\|_{\mathcal{Y}}^2 &\leq \varepsilon c(\alpha) + \sqrt{\sum_{j=0}^{\infty} \left(r(\alpha, \sigma_j) - 1 \right)^2 \langle y, e^j \rangle_{\mathcal{Y}}^2} \\ &\leq \varepsilon c(\alpha) + \sup_{\sigma} |r(\alpha, \sigma) - 1| \|y\|_{\mathcal{Y}}. \end{aligned}$$

■

3.3 Espaces sources et optimalité avec la décomposition en valeurs singulières

A partir de la décomposition en valeurs singulières, il y a une façon très simple de définir les puissances de l'opérateur y compris pour des valeurs non entières : pour $s > 0$ on définit, pour tout $y \in \mathcal{Y}$:

$$(\Psi^* \Psi)^s y := \sum_{j=0}^{\infty} \sigma_j^{2s} \langle y, e^j \rangle_{\mathcal{Y}} e^j. \quad (3.6)$$

On voit que cette définition est moins restrictive que $y \in \mathcal{Y}$, et peut être appliquée à des suites d'éléments a_j dont on définit formellement l'élément $y = \sum a_j e^j$, avec $a_j \notin \ell^2$ mais $a_j \sigma_j^{2s} \in \ell^2$. Cela définit donc un espace qu'on note \mathcal{Y}^{-s} plus grand que \mathcal{Y} .

De même, on peut définir des espaces *plus petits* que \mathcal{Y} de la même façon mais pour $s < 0$. On peut ainsi définir une suite d'espaces sources \mathcal{Y}^s , sur lesquels on peut définir une norme hilbertienne :

$$\mathcal{Y}^s = \text{Im}((\Psi^* \Psi)^s) = \left\{ y \in \mathcal{Y}, \quad \sum_{j=0}^{\infty} \frac{\langle y, e^j \rangle_{\mathcal{Y}}^2}{\sigma_j^{4s}} < +\infty \right\}, \quad \|y\|_{\mathcal{Y}^s}^2 := \sum_{j=0}^{\infty} \frac{\langle y, e^j \rangle_{\mathcal{Y}}^2}{\sigma_j^{4s}}. \quad (3.7)$$

On peut définir une norme de la même façon pour des espaces \mathcal{Y}^s avec $s < 0$. Remarquez également que l'on peut identifier l'espace $\overline{\text{Im}(\Psi)}$ et l'espace $\mathcal{Y}^{-\frac{1}{2}}$. En effet, soit $z \in \overline{\text{Im}(\Psi)}$, on n'a pas $z = \Psi y$ mais on peut définir des scalaires $y_j = \frac{\langle z, f^j \rangle_{\mathcal{Z}}}{\sigma_j}$, cela définit un élément $y = \sum_{j=0}^{\infty} y_j e^j$ auquel on peut appliquer la norme

$$\|y\|_{\mathcal{Y}^{-\frac{1}{2}}}^2 := \sum_{j=0}^{\infty} \sigma_j^2 \langle y, e^j \rangle_{\mathcal{Y}}^2 = \sum_{j=0}^{\infty} \langle z, f^j \rangle_{\mathcal{Z}}^2.$$

Cela montre que $y \in \mathcal{Y}^{-\frac{1}{2}}$. Réciproquement, considérons un élément $y \in \mathcal{Y}^{-\frac{1}{2}}$. On peut alors considérer la suite d'éléments tronqués $z_n = \sum_{j=0}^n \sigma_j \langle y, e^j \rangle f^j \in \text{Im}(\Psi)$. On voit que $z_n \xrightarrow{\mathcal{Z}} z$ donc $z \in \overline{\text{Im}(\Psi)}$. Cela montre que $\Psi : \mathcal{Y}^{-\frac{1}{2}} \rightarrow \overline{\text{Im}(\Psi)} (= \mathcal{Z})$ est une transformation unitaire entre les deux espaces.

Pour notre exemple où Ψ est un opérateur d'intégration, il y a bien sûr un lien entre les espaces de Sobolev H^s et les espaces \mathcal{Y}^s , que nous développerons plus tard si nous en avons le temps (« en gros », dans notre exemple, \mathcal{Y}^s est un espace « de type » $H^{\frac{s}{2}}$.) De manière plus générale, on voit que l'on a une suite d'espaces imbriqués $\mathcal{Y}^{s_2} \subset \mathcal{Y}^{s_1} \subset \mathcal{Y}$ pour $0 < s_1 < s_2$. Cela nous permet de définir des espaces \mathcal{K} naturels sur lesquels estimer

l'optimalité d'une méthode, telle que définie par la définition 5 : ce sont les boules des espaces \mathcal{Y}^s , que nous notons $B_s(x_0, \delta)$ pour la boule de \mathcal{Y}^s de centre x_0 et de rayon δ au sens de la norme de \mathcal{Y}^s . Sur ces boules, nous allons tâcher de caractériser l'optimalité des méthodes de régularisation.

Par rapport au chapitre précédent, où l'on parlait de condition source, l'a priori $y \in B_s(y_0, \delta)$ se réécrit $\|B(y - y_0)\|_{\mathcal{Y}_B} \leq \delta$ avec $B = Id$ et $\mathcal{Y}_B = \mathcal{Y}^s$ muni de sa norme.

On avait une borne inférieure pour cette optimalité grâce à la définition de $\Omega(\varepsilon, \mathcal{K})$, avec $\Omega(\varepsilon, \mathcal{K}) \leq \Delta(\varepsilon, \mathcal{K}, R_0)$. On commence donc par un résultat sur Ω .

Proposition 6. Pour tout $\delta > 0$, $s > 0$, on a, pour tout $y \in \mathcal{Y}^s$:

$$\|y\|_{\mathcal{Y}} \leq \|\Psi y\|_{\mathcal{Z}}^{\frac{2s}{2s+1}} \|y\|_{\mathcal{Y}^s}^{\frac{1}{2s+1}}. \quad (3.8)$$

Cela entraîne l'inégalité suivante pour $\Omega(\varepsilon, B_s(0, \delta))$ défini par (2.15) :

$$\Omega(\varepsilon, B_s(0, \delta)) \leq \varepsilon^{\frac{2s}{2s+1}} \delta^{\frac{1}{2s+1}}, \quad (3.9)$$

et de plus il existe une suite $\varepsilon_k \rightarrow 0$ telle que

$$\Omega(\varepsilon_k, B_s(0, \delta)) = \varepsilon_k^{\frac{2s}{2s+1}} \delta^{\frac{1}{2s+1}}. \quad (3.10)$$

On voit que la grandeur qui entre en jeu est $\varepsilon^{2s} \delta$: l'a priori et la norme de y^\dagger se révèlent aussi importants que l'a priori et la norme du niveau de bruit. La différence principale est que **contrairement au niveau de bruit, comme on le verra par la suite, l'on peut s'en passer a priori, et bâtir une méthode convergente (et même optimale) a posteriori.**

Preuve. Rappelons la définition de $\Omega(\varepsilon, B_s(0, \delta))$ donnée par (2.15) :

$$\Omega(\varepsilon, B_s(0, \delta)) := \sup_{\|y\|_{\mathcal{Y}^s} \leq \delta, \|\Psi y\|_{\mathcal{Z}} \leq \varepsilon} \|y\|_{\mathcal{Y}}.$$

On voit donc comment l'inégalité (3.8) entraîne de façon immédiate (3.9).

On cherche donc une estimation de $\|y\|_{\mathcal{Y}}$ en fonction de $\|y\|_{\mathcal{Y}^s}$ et de $\|\Psi y\|_{\mathcal{Z}} = \|y\|_{\mathcal{Y}^{-\frac{1}{2}}}$. Comme on a $\mathcal{Y}^s \subset \mathcal{Y} \subset \mathcal{Y}^{-\frac{1}{2}}$, on reconnaît là une inégalité d'interpolation : on souhaite obtenir une borne supérieure sur un espace de régularité intermédiaire entre celle de deux autres espaces.

Rappelons l'inégalité de Hölder sur \mathbb{R}^N : pour $a \in \ell^p(\mathbb{R})$ et $b \in \ell^{p'}(\mathbb{R})$ avec $\frac{1}{p} + \frac{1}{p'} = 1$, $p, p' > 1$ on a

$$\sum_{j=0}^{\infty} |a_j b_j| \leq \left(\sum_{j=0}^{\infty} |a_j|^p \right)^{\frac{1}{p}} \left(\sum_{j=0}^{\infty} |b_j|^{p'} \right)^{\frac{1}{p'}}.$$

On souhaite ici une inégalité qui fasse apparaître à gauche $\sum_{j=0}^{\infty} y_j^2$ et à droite $\sum_{j=0}^{\infty} \sigma_j^{-4s} y_j^2$ et $\sum_{j=0}^{\infty} \sigma_j^2 y_j^2$. Pour utiliser l'inégalité de Hölder, écrivons $y_j^2 = a_j b_j$ avec $a_j = \sigma_j^{2\alpha} y_j^{2\beta}$ et $b_j = \sigma_j^{-2\alpha} y_j^{2-2\beta}$ ce qui est vrai pour tout $\alpha, \beta \in \mathbb{R}$. On veut donc trouver $p, p' > 1$ tels que d'une part $a_j^p = \sigma_j^2 y_j^2$ et d'autre part $b_j^{p'} = \sigma_j^{-4s} y_j^2$. Identifiant les termes, cela conduit aux relations suivantes sur α, β et p, p' :

$$2\alpha p = 2, \quad 2\beta p = 2, \quad 4s = 2\alpha p', \quad p'(2 - 2\beta) = 2.$$

Cela conduit à choisir $\alpha = \beta = \frac{1}{p}$, et $\frac{p'}{p} = 2s$ soit $p'(1 - \frac{1}{p'}) = 2s$ soit $p' = 2s + 1$. L'inégalité de Hölder conduit alors à

$$\|y\|_{\mathcal{Y}}^2 = \sum_{j=0}^{\infty} y_j^2 \leq \left(\sum_{j=0}^{\infty} \sigma_j^2 y_j^2 \right)^{\frac{2s}{2s+1}} \left(\sum_{j=0}^{\infty} \sigma_j^{-4s} y_j^2 \right)^{\frac{1}{2s+1}} = \|\Psi y\|_{\mathcal{Z}}^{\frac{4s}{2s+1}} \|y\|_{\mathcal{Y}^s}^{\frac{2}{2s+1}},$$

qui est l'inégalité (3.8). En passant au sup dans la définition de $\Omega(\varepsilon, B_s(0, \delta))$ on obtient (3.9). Pour l'égalité, on se sert des vecteurs de la base singulière : prenons $y = b_k e^k$, on a $\Psi e^k = \sigma_k f^k$ donc $\|\Psi y\|_{\mathcal{Z}} = \sigma_k b_k = \varepsilon$ pour $b_k = \frac{\varepsilon}{\sigma_k}$, et $\|y\|_{\mathcal{Y}} = b_k = \frac{\varepsilon}{\sigma_k}$, $\|y\|_{\mathcal{Y}^s} = b_k \sigma_k^{-2s} = \varepsilon \sigma_k^{-(2s+1)} = \delta$ pour $\varepsilon = \varepsilon_k := \sigma_k^{2s+1} \delta \rightarrow_{k \rightarrow \infty} 0$. Pour cette valeur de ε on a donc

$$\Omega(\varepsilon_k, B_s(0, \delta)) \geq b_k = \frac{\varepsilon_k}{\sigma_k} = \varepsilon_k^{1-\frac{1}{2s+1}} \delta^{\frac{1}{2s+1}} = \varepsilon_k^{\frac{2s}{2s+1}} \delta^{\frac{1}{2s+1}},$$

et l'inégalité contraire vue à l'étape précédente entraîne l'égalité recherchée. \blacksquare

On a donc obtenu une borne inférieure pour $\Delta(\varepsilon, B_s(0, \delta), R)$ qui donne une vitesse de convergence maximale pour toute méthode R . La question est maintenant de trouver, si cela est possible, une méthode qui réalise cette convergence, auquel cas il s'agira bien d'une méthode optimale (ou d'ordre optimale, si ce n'est qu'à une constante près). Regardons la preuve ci-dessus : nous avons utilisé une certaine valeur d'indice pour un vecteur singulier. Tout naturellement, une première méthode qui vient à l'esprit est donc de tronquer la suite des valeurs singulières à ce niveau : choisissons

$$r(\alpha, \sigma) = \mathbb{1}_{\sigma \geq \alpha}, \tag{3.11}$$

et voyons quelle valeur de α conviendra le mieux. On a alors

$$y_{\varepsilon, \alpha} - y = y_{\varepsilon, \alpha} - y_\alpha + y_\alpha - y = \sum_{j, \sigma_j \geq \alpha} \sigma_j^{-1} \langle z_\varepsilon - z, f^j \rangle_{\mathcal{Z}} e^j + \sum_{j, \sigma_j < \alpha} \sigma_j^{-1} \langle z, f^j \rangle e^j.$$

On voit que l'erreur de mesure est dans le premier terme, où l'explosion est limitée car la somme est finie. Le deuxième terme est régulier, et comporte les plus grands indices : la décroissance dépendra de la régularité de l'a priori, puisque plus la régularité (dans la suite d'espaces \mathcal{Y}^s) est grande, plus la décroissance est rapide. Evaluons le premier terme (« variance »)

$$\|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{Y}}^2 = \sum_{j, \sigma_j \geq \alpha} \sigma_j^{-2} \langle z_\varepsilon - z, f^j \rangle_{\mathcal{Z}}^2 \leq \frac{1}{\alpha^2} \sum_{j, \sigma_j \geq \alpha} \langle z_\varepsilon - z, f^j \rangle_{\mathcal{Z}}^2 \leq \frac{\varepsilon^2}{\alpha^2}.$$

Sous l'hypothèse que nous avons faite sur le bruit, cette estimation sur la « variance » est optimale : pour s'en rendre compte il suffit de prendre la suite $\alpha_k = \sigma_k$ et comme mesure bruitée $z_\varepsilon = z + \varepsilon f^k$.

Evaluons le second terme à l'aide de l'a priori $y \in \mathcal{Y}^s$ avec $s > 0$ et notant $y_j := \langle y, e^j \rangle_{\mathcal{Y}}$:

$$\|y_\alpha - y\|_{\mathcal{Y}}^2 = \sum_{j, \sigma_j < \alpha} y_j^2 = \sum_{j, \sigma_j < \alpha} \sigma_j^{-4s+4s} y_j^2 \leq \alpha^{4s} \sum_{j, \sigma_j < \alpha} \sigma_j^{-4s} y_j^2 \leq \alpha^{4s} \|y\|_{\mathcal{Y}^s}^2.$$

Revenons à notre inégalité de départ :

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}}^2 \leq \frac{\varepsilon^2}{\alpha^2} + \delta^2 \alpha^{4s}.$$

Considéré comme une fonction de α , le membre de droite a un minimum unique en $\alpha = (\frac{\varepsilon}{\sqrt{2s}\delta})^{\frac{1}{2s+1}}$ ¹, et on obtient

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}} \leq \sqrt{2} \delta^{\frac{1}{2s+1}} \varepsilon^{\frac{2s}{2s+1}}.$$

On a une méthode non pas *optimale* mais *d'ordre optimal* : nous avons perdu dans nos estimations une constante $\sqrt{2}$. Cela nous suffit et nous mène à la définition suivante.

Définition 6. Une méthode est dite optimale si $\Delta(\varepsilon, B_s(0, \delta), R_\alpha) = \varepsilon^{\frac{2s}{2s+1}} \delta^{\frac{1}{2s+1}}$, et d'ordre optimal si $\Delta(\varepsilon, B_s(0, \delta), R_\alpha) = O\left(\varepsilon^{\frac{2s}{2s+1}} \delta^{\frac{1}{2s+1}}\right)$.

Exemple : avec notre exemple de l'opérateur de dérivation, les espaces \mathcal{Y}^s relatifs à $\mathcal{Y} = \mathcal{Z} = L^2$ sont des sous-espaces des espaces $H^{\frac{s}{2}}([0, 1])$. On voit que l'on attend une convergence au mieux en $O(\varepsilon^{\frac{1}{2}})$ si $x \in H^1$, en $O(\varepsilon^{\frac{2}{3}})$ si $x \in H^2$, en $O(\varepsilon^{\frac{s}{s+1}})$ si $x \in H^s$. Cela sera confirmé par la suite à l'aide d'estimations directes (en tout point similaires à celles-ci).

1. Plus simplement on peut prendre $\alpha = (\frac{\varepsilon}{\delta})^{\frac{1}{2s+1}}$, valeur qui permet d'égaliser les deux membres du terme de droite : on sait que cette valeur sera du même ordre de grandeur que le minimum.

3.4 Implémentation numérique

La décomposition en valeurs singulières a un premier avantage : elle rend l'analyse très explicite, et une fois connue, elle fournit des méthodes d'ordre optimal explicites. Cependant elle a un inconvénient majeur : elle est coûteuse à implémenter numériquement dans le cas général.

3.4.1 Cas où l'on connaît déjà la décomposition en valeurs singulières

Prenons le cas de l'opérateur d'intégration ci-dessus, avec tout d'abord $\mathcal{Y} = \mathcal{Z} = L^2$, et une donnée $z_\varepsilon = (z_{i+\frac{1}{2}})$ donnée sur une grille uniforme de $[0, 1]$: les étapes sont les suivantes.

- Calculer $z_\varepsilon^j = \langle z_\varepsilon, f^j \rangle_{L^2}$: par une méthode classique d'approximation de l'intégrale. La page wikipédia http://fr.wikipedia.org/wiki/Calcul_numerique_d'une_intégrale est très bien faite. On se heurte ici à une vraie difficulté : étant donnée notre hypothèse sur le bruit, on a seulement $z_\varepsilon \in L^2$. Or les méthodes d'intégration reposent toujours sur le fait que la fonction sous-jacente soit régulière - sans quoi une évaluation ponctuelle est dénuée de sens. Si l'on prend par exemple la formule de quadrature d'ordre m on aura une convergence en $O(h^m)$ si z_ε est C^m ... ce qu'il n'est pas. Il est donc a priori inutile de prendre une méthode trop précise, et surtout il faudrait avant de faire le calcul de l'intégrale régulariser la fonction z_ε ... Mais alors notre décomposition en valeurs singulières n'a plus d'intérêt : c'est une régularisation qui va se superposer à une autre régularisation ! Bref : prenons pour commencer la méthode des rectangles

$$\langle z_\varepsilon, f^j \rangle_{L^2} = \sum_{k=1}^{N+1} z_\varepsilon^k f^j(x_k) \Delta x_k$$

- tronquer pour un $J = O((\frac{\delta}{\varepsilon})^{\frac{1}{2s+1}})$.
- Choisir $y_{\varepsilon,\alpha} := \sum_{j=0}^J \frac{z_\varepsilon^j}{\sigma_j} e^j$. Ici l'inconvénient de la faible vitesse de convergence semble disparaître : en effet, on verra « expérimentalement » si le résidu / *discrepancy* est inférieur ou non à ε . On peut ainsi choisir J non pas *a priori* par la formule précédente, mais *a posteriori*, comme le plus petit J tel que l'erreur soit plus petite que ε ; ou encore arrêter quand la diminution de l'erreur devient trop faible. Le résidu / *discrepancy* s'écrit maintenant (pour la méthode des rectangles)

$$\|z_\varepsilon - z_{\varepsilon,\alpha}\|_{\ell^2}^2 = \sum_{i=1}^{N+1} \left(\sum_{j=0}^J \left(\sum_{k=1}^{N+1} z_\varepsilon^k f^j(x_k) \Delta x_k \right) f^j(x_i) - z_\varepsilon^i \right)^2 \Delta x$$

Exercice : le faire en matlab ou scilab, avec comme fonction test sur $[0, 1]$ par exemple $f(x) = (x - \frac{1}{2})^2$, $f(x) = 1$, $f(x) = \mathbb{1}_{\frac{1}{4} \leq x \leq \frac{3}{4}}$, $f(x) = 2x\mathbb{1}_{x \leq \frac{1}{2}} + 1\mathbb{1}_{x > \frac{1}{2}}$ et $f(x) = e^{-\frac{1}{\sigma^2}(x - \frac{1}{2})^2}$ en faisant varier σ , en faisant varier aussi le nombre de points de la grille et le bruit ε . Essayer également de coder deux méthodes d'intégration différente. Penser à la structure des codes, sachant que l'on verra par la suite d'autres méthodes de régularisation. Pour la fois prochaine (et pour ceux qui veulent) : réfléchir déjà à la structure des codes dont on pourra discuter, et des tests à faire.

Les courbes à tracer importantes sont :

- le résidu (*discrepancy*) par rapport au paramètre de régularisation α (ou ici à l'inverse de l'indice de troncature J) : on doit avoir une courbe croissante, pour les différentes méthodes numériques et pour différents niveaux de bruit ; voir quelles échelles sont les meilleures (a priori : log log)
- précision de l'estimation par rapport au niveau de bruit (en log-log scale : ainsi on devrait obtenir des droites) ; ceci pour des méthodes de régularisation a priori et a posteriori, et pour différentes fonctions tests (qui sont plus ou moins régulières). Cela permet de tester numériquement si les résultats sont ou non conformes à la théorie.

Une autre question se pose : comment simuler le bruit ? Dans quel espace aura-t-on alors une estimation de ce bruit ? On peut par exemple utiliser la fonction rand ou randn et coder un bruit additif ou multiplicatif.

3.4.2 décomposition en valeurs singulières en dimension finie

Voyons tout d'abord la version de la décomposition en valeurs singulières pour les matrices.

Théorème 6. Soit M une matrice de taille $m \times n$ à coefficients dans \mathbb{R} ou \mathbb{C} (avec $m \geq n$?) Alors il existe une factorisation de M sous la forme

$$M = U\Sigma V^*, \quad (3.12)$$

où Σ est une matrice de taille $m \times n$ dont tous les coefficients sont nuls sauf les coefficients diagonaux, qui sont positifs ou nuls, U est une matrice unitaire carrée de taille $m \times m$ et V une matrice unitaire carrée de taille $n \times n$. (i.e. $U^*U = UU^* = Id_m$, $V^*V = VV^* = Id_n$). Si on range les valeurs $\Sigma_{i,i}$ par ordre décroissant, Σ est déterminé de façon unique.

La preuve repose par exemple sur le théorème spectral appliqué à M^*M ...

3.4.3 Calcul numérique de la décomposition en valeurs singulières

la commande svd de Matlab fait appel à la procédure DGESVD de la bibliothèque LAPACK. Un des algorithmes les plus utilisés est celui de Golub-Reinsch (1970) : voir

l'article [29]. Il y a aussi des commandes plus « économiques » : svds ne calcule que les premiers éléments (6 par défaut, où K si on le précise).

Aussi, dans l'exemple ci-dessus, si l'on ignorait la décomposition en valeurs singulières de l'opérateur d'intégration, on peut procéder comme suit, si l'on suppose toute fonction discrétisée selon une grille.

- Choisir une méthode d'approximation de l'intégrale.
- Traduire cette méthode sous une forme matricielle.
- Calculer la svd ou du moins les premiers coefficients de la svd correspondante à l'aide de la commande svd ou svds de matlab
- Comparer par rapport à la valeur exacte calculée plus haut.
- Appliquer à une fonction test.

On voit ici apparaître toute une série de calculs qui sont *spécifiques* à chaque problème inverse considéré. On voit aussi à quel point cette méthode est liée à l'étude du problème direct : meilleure l'approximation donnée par le schéma numérique du problème direct (nécessaire pour la discrétisation et donc conditionnant la matrice A), meilleure la résolution du problème inverse.

3.4.4 Application en traitement des images

Le traitement d'images est un domaine à part entière, je ne prétends pas ici épuiser le sujet mais plutôt donner un (très) léger aperçu.

Un modèle pour le floutage des images est le suivant, en deux dimensions :

$$\Psi y(s, t) = \int_{\mathbb{R}^2} k(s, t, s', t') y(s', t') ds' dt',$$

et en supposant que le flou est uniforme et homogène :

$$k(s, t, s', t') = k_0(s - s', t - t'),$$

c'est-à-dire que notre problème : retrouver y en mesurant z , revient à de la *déconvolution* :

$$\Psi y = k_0 \star y.$$

Par exemple, pour un appareil photo, la lumière est uniformément répartie sur l'objectif, qui est un disque : on aura $k_0(s, t) = \mathbb{1}_{s^2 + t^2 \leq r^2}$. Un problème crucial est bien sûr non seulement d'inverser Ψ , mais aussi - d'abord - de déterminer k_0 . Imaginons qu'on connaisse k_0 (étalonnage de la caméra, qui se fait par ex. sur une image connue) alors, l'image étant pixelisée, on peut transformer la matrice image en un vecteur image (en mettant bout à bout les coefficients), on aura pour k_0 une matrice, on peut appliquer la décomposition en valeurs singulières.

Chapitre 4

La régularisation de Tikhonov

4.1 Généralités sur la régularisation de Tikhonov

On a défini ce qu'est une *solution au sens des moindres carrés*. On a également démontré que pour $z \in D(\Psi^\dagger)$, $y \in \mathcal{Y}$ est une solution au sens des moindres carrés *ssi* elle satisfait l'équation normale (2.9)

$$\Psi^* \Psi y = \Psi^* z,$$

ce qui se réécrit, comme on l'a vu ci-dessus, grâce à la décomposition en valeurs singulières (avec Ψ injectif - sinon on doit ajouter $\text{Ker}(\Psi)$) :

$$y_j = \frac{1}{\sigma_j} z_j.$$

La décomposition en valeurs singulières fournit donc une méthode de résolution directe pour la méthode des moindres carrés, dont on a vu que son inconvénient principal est qu'elle ne converge pas. D'où la nécessité de définir des méthodes de régularisation, pour lesquelles, si l'on souhaite obtenir des taux de convergence, il faut aussi définir des a priori. On a alors vu, en fonction de l'a priori et de l'espace source choisi, comment obtenir un tel taux de convergence et on a défini leur optimalité.

Maintenant si l'on en revient au problème des moindres carrés en oubliant pour l'instant la décomposition en valeurs singulières : sachant que 1/ on ne connaît le vrai z qu'à ε près dans l'espace \mathcal{Z} , et 2/ on souhaite que la norme où une norme de y « n'explose pas », on peut imaginer deux autres types de méthodes, symétriques l'une de l'autre, qui « améliorent » « corrigent » la méthode des moindres carrés. Afin de rester dans un cadre fonctionnel déjà vu et sur lequel l'analyse est commode, nous considérerons souvent comme a priori : « $y \in B_s(0, \delta)$. » En effet pour un tel a priori, la décomposition en valeurs singulières nous a permis de connaître la convergence optimale de toute méthode de régularisation.

- La méthode des résidus : on minimise $\|y\|_{\mathcal{Y}^p}$ sous la contrainte $\|\Psi y - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$. On appelle *résidu* ou *discrepancy* la quantité $\|\Psi y - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$. Si l'on a déjà une connaissance a priori de y on minimisera $\|y - y_0\|_{\mathcal{Y}^p}$ avec y_0 l'a priori.
- La méthode des quasi-solutions : on minimise $\|\Psi y - z_\varepsilon\|_{\mathcal{Z}}$ sous la contrainte $\|y\|_{\mathcal{Y}^p} \leq \delta$ (ou $\|y - y_0\|_{\mathcal{Y}^p} \leq \delta$).

Entre ces deux méthodes, la *méthode de Tikhonov* apparaît comme un compromis : au lieu d'imposer l'une des contraintes et de minimiser l'autre, on minimise la somme pondérée des deux

$$J(y) := \frac{1}{\varepsilon^2} \|\Psi y - z_\varepsilon\|_{\mathcal{Z}}^2 + \frac{1}{\delta^2} \|y - y_0\|_{\mathcal{Z}_B}^2.$$

ou bien, dans le cas spécifique de $B = Id$, $\mathcal{Y}_B = \mathcal{Y}^p$, $\mathcal{K} = B_{\mathcal{Y}^p}(0, \delta)$

$$J(y) := \frac{1}{\varepsilon^2} \|\Psi y - z_\varepsilon\|_{\mathcal{Z}}^2 + \frac{1}{\delta^2} \|y - y_0\|_{\mathcal{Y}^p}^2.$$

Ici, on a « intuitivement » divisé chaque quantité par sa « valeur cible », de façon à donner à chacune un poids qui semble équivalent. Une autre façon de voir est que chacune des deux méthodes citées ci-dessus - méthode des résidus et méthode des quasi-solutions - est un problème d'*optimisation sous contrainte*. Une des méthodes phare pour résoudre un problème d'optimisation sous contrainte est la méthode des *multiplicateurs de Lagrange*, qui permet de remplacer un problème d'optimisation sous contrainte par un problème d'optimisation sans contrainte. Nous verrons par la suite (intervention de P. Moireau sur l'assimilation de données) que c'est par ce biais de l'optimisation que l'on voit le mieux le lien entre problèmes inverses et contrôle.

Une autre façon encore de voir les choses (vocabulaire de l'assimilation de données) est de considérer que y_0 correspond à une première mesure, à laquelle on attribue la confiance pondérée par δ , et z_ε une nouvelle mesure de y , faite à travers *l'opérateur d'observation* qui est ici Ψ . Alors minimiser J correspond à une méthode des moindres carrés sur l'opérateur étendu $\bar{\Psi} : \mathcal{Y}^p \rightarrow \mathcal{Z} \times \mathcal{Y}^p$, l'espace $\mathcal{Z} \times \mathcal{Y}^p$ étant muni du produit scalaire produit pondéré (i.e. la somme des produits scalaires multipliés par les coefficients ε^{-2} et δ^{-2} respectivement), avec $\bar{\Psi}(y) := (\Psi y, y)$ et comme mesure $\bar{z}_\varepsilon = (z_\varepsilon, y_0)$. On voit que dans ce formalisme, même si l'opérateur Ψ n'est pas injectif, l'opérateur $\bar{\Psi}$ a été rendu *injectif*. On comprend mieux aussi le terme *assimilation de données* : il s'agit, au fur et à mesure de nouvelles observations, de corriger les précédentes en agrégant l'ensemble des connaissances. C'est pourquoi **la définition de la fonctionnelle $J(y)$ est fondamentale**.

Encore une remarque d'ordre général : lorsque nous avons vu la décomposition en valeurs singulières, toute la théorie utilisait le fait que l'opérateur Ψ était linéaire. Avec la méthode des moindres carrés, ou ses généralisations ci-dessus (quasi-solutions, résidus ou Tikhonov), on voit qu'il est très facile d'étendre les définitions à un opérateur Ψ non linéaire.

Si $B = Id$ et $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}$, on parle de la méthode de Tikhonov classique.

On a choisi ici de noter \mathcal{Y}^p l'espace de la contrainte, car il faut bien avoir à l'esprit qu'il ne s'agit pas du même espace que celui de l'espace-source, pour lequel on garde la notation s .

Jusqu'ici on ne voit pas apparaître de « paramètre de régularisation ». En revanche, on voit le jeu entre ε et δ comme auparavant : il est donc naturel de rechercher la méthode permettant d'optimiser leur rapport. Cela conduit à la méthode de Tikhonov *généralisée* :

Minimiser la quantité

$$J_\alpha(y) := \frac{1}{2} \|\Psi y - z_\varepsilon\|_{\mathcal{Z}}^2 + \frac{\alpha^2}{2} \|By\|_{\mathcal{Z}_B}^2. \quad (4.1)$$

Très utile pour le calcul numérique autant que théorique, et permettant aussi de mettre en évidence le lien entre la régularisation de Tikhonov et la décomposition en valeurs singulières, commençons par un calcul classique liant le minimum ci-dessus et la solution d'une équation.

Théorème 7. *Soit $y_{\varepsilon,\alpha} \in \mathcal{Y}_B$, \mathcal{Y}_B dense dans \mathcal{Y} , et $B : \mathcal{Y}_B \rightarrow \mathcal{Z}_B$ continue au sens des normes de \mathcal{Y} et de \mathcal{Z}_B . On note $B^* : \mathcal{Z}_B \rightarrow \mathcal{Y}$ l'adjoint de B au sens de la norme de \mathcal{Y} . Alors $y_{\varepsilon,\alpha}$ est un minimiseur de J_α sur \mathcal{Y}_B ssi $y_{\varepsilon,\alpha}$ est solution de l'équation*

$$(\Psi^* \Psi + \alpha^2 B^* B) y_{\varepsilon,\alpha} = \Psi^* z_\varepsilon \quad (4.2)$$

soit dans le cas de la méthode de Tikhonov classique, i.e. si $B = Id$ et $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}$:

$$\Psi^* \Psi y_{\varepsilon,\alpha} + \alpha^2 y_{\varepsilon,\alpha} = \Psi^* z_\varepsilon. \quad (4.3)$$

Remarque 1. *Ce théorème donne une équivalence entre les minimiseurs de J_α et les solutions de l'équation (4.2), mais ne donne pas de conditions sur B qui permettent d'assurer l'existence et l'unicité d'un minimiseur. Ce sera le cas par exemple si B est inversible d'inverse continue de \mathcal{Y}_B dans \mathcal{Z}_B (et donc en particulier c'est le cas pour la méthode de Tikhonov classique) : dans ce cas, pour tout $\alpha > 0$, J_α possède un unique minimiseur sur \mathcal{Y}_B . En effet, J_α est convexe et si B est inversible d'inverse continue elle tend vers l'infini quand $\|y\|_{\mathcal{Y}_B}$ tend vers l'infini.*

Preuve. La fonction J_α est différentiable au sens de Fréchet en tout $y \in \mathcal{Y}_B$.

Définition 7 (Rappel sur la dérivée de Fréchet). *Soit $F : \mathcal{Y} \rightarrow \mathcal{Z}$ une application et $D(F)$ son domaine de définition. On dit que F est Fréchet dérivable en un point $y \in Int(D(F))$ s'il existe un opérateur linéaire borné $F'(y) : \mathcal{Y} \rightarrow \mathcal{Z}$ tel que*

$$\lim_{h \in \mathcal{Y}, h \rightarrow 0} \frac{\|F(y + h) - F(y) - F'(y)h\|_{\mathcal{Z}}}{\|h\|_{\mathcal{Y}}} = 0$$

On applique ici la définition sur \mathcal{Y}_B (et non sur \mathcal{Y}) à l'opérateur $J_\alpha(y)$, et l'on calcule classiquement, pour $h \in \mathcal{Y}_B$:

$$J_\alpha(y + h) - J_\alpha(y) = \langle \Psi h, \Psi y - z_\varepsilon \rangle_{\mathcal{Z}} + \alpha^2 \langle Bh, By \rangle_{\mathcal{Z}_B} + \frac{1}{2} \|\Psi h\|_{\mathcal{Z}}^2 + \frac{\alpha^2}{2} \|Bh\|_{\mathcal{Z}_B}^2, \quad (4.4)$$

et donc, B étant borné dans \mathcal{Y}_B :

$$\forall h \in \mathcal{Y}_B, J'_\alpha(y)(h) = \langle \Psi h, \Psi y - z_\varepsilon \rangle_{\mathcal{Z}} + \alpha^2 \langle Bh, By \rangle_{\mathcal{Z}_B}.$$

ce qu'on peut encore écrire, par dualité en notant bien $B^* : \mathcal{Z}_B \longrightarrow \mathcal{Y}$ l'adjoint de B de \mathcal{Z}_B dans \mathcal{Y} (ce qui est différent de l'adjoint de B de \mathcal{Z}_B dans \mathcal{Y}_B)

$$\forall h \in \mathcal{Y}_B, J'_\alpha(y)(h) = \langle \Psi^* \Psi y - \Psi^* z_\varepsilon, h \rangle_{\mathcal{Y}} + \alpha^2 \langle B^* By, h \rangle_{\mathcal{Y}},$$

et par densité de \mathcal{Y}_B dans \mathcal{Y} cela est aussi vrai pour tout $h \in \mathcal{Y}$. En un point minimisant de J_α , sa dérivée s'annule, et on obtient l'équation (4.2). Réciproquement, si $y_{\varepsilon, \alpha}$ vérifie l'équation (4.2), alors en notant $y = y_{\varepsilon, \alpha} + h$, l'identité ci-dessus (4.4) appliquée à $y_{\varepsilon, \alpha}$ et à h devient

$$J_\alpha(y) - J_\alpha(y_{\varepsilon, \alpha}) = \frac{1}{2} \|\Psi h\|_{\mathcal{Z}}^2 + \frac{\alpha^2}{2} \|Bh\|_{\mathcal{Z}_B}^2 = \frac{1}{2} \|\Psi y - \Psi y_{\varepsilon, \alpha}\|_{\mathcal{Z}}^2 + \frac{\alpha^2}{2} \|By - By_{\varepsilon, \alpha}\|_{\mathcal{Z}_B}^2 \geq 0,$$

avec égalité *ssi* $y = y_{\varepsilon, \alpha}$. ■

Cela revient à résoudre un problème inverse approché : la solution approchée est la solution exacte du problème approché. Contrairement au problème inverse exact, ce problème approché est bien posé, comme le montre le théorème 7.

4.1.1 Résultats directs

Il n'y a pas besoin de passer par la décomposition en valeurs singulières pour montrer qu'à partir de la définition de J_α , on peut construire une méthode de régularisation au sens de la définition 4. Ces méthodes directes ont leur intérêt car, sous certaines hypothèses, elles peuvent se généraliser à des problèmes non linéaires.

Remarquons tout d'abord que le théorème de Lax-Milgram est une autre façon de donner une solution unique au problème de minimisation, dès lors que B est inversible (donc en particulier c'est le cas pour la méthode de Tikhonov classique).

Résultat de convergence grâce à l'écriture sous forme de minimisation

Commençons par un résultat général.

Théorème 8. Soit $z_\varepsilon \in \mathcal{Z}$, $y_{\varepsilon,\alpha}$ l'unique solution de (4.2) dans le cas de la méthode de Tikhonov classique, i.e. avec $B = Id$ et $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}$. C'est donc aussi l'unique minimiseur de J_α défini par (4.1) (Ψ injectif). Soit $z \in Im(\Psi)$, $z = \Psi y$ tel que $\|z - z_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon$. Si l'on choisit $\alpha = \alpha(\varepsilon)$ tel que

$$\lim_{\varepsilon \rightarrow 0} \alpha(\varepsilon) = 0, \quad \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon}{\alpha(\varepsilon)} = 0, \quad (4.5)$$

alors $\lim_{\varepsilon \rightarrow 0} y_{\varepsilon,\alpha(\varepsilon)} = y$, donc on a bien défini une méthode de régularisation $R_{\alpha(\varepsilon)} z_\varepsilon = y_{\varepsilon,\alpha}$ convergente.

Une fois de plus, on remarque que sans hypothèse supplémentaire sur y , on n'a pas de vitesse de convergence.

Preuve. On ne peut utiliser que la fonctionnelle J_α , qui ne comporte pas d'estimation directe de $\|y_{\varepsilon,\alpha} - y\|$. On procède donc comme suit.

- On prouve (par J_α) que $y_{\varepsilon,\alpha(\varepsilon)}$ est borné uniformément, donc on peut extraire une sous-suite séquentiellement faiblement convergente $y_n = y_{\varepsilon_n, \alpha(\varepsilon_n)}$ vers \tilde{y} .

Preuve :

$$J_\alpha(y_{\varepsilon,\alpha}) \leq J_\alpha(y) \leq \frac{\varepsilon^2}{2} + \frac{\alpha^2}{2}(\varepsilon) \|y\|_{\mathcal{Y}}^2,$$

donc $\|y_{\varepsilon,\alpha}\|_{\mathcal{Y}}^2 \leq \frac{\varepsilon^2}{\alpha(\varepsilon)^2} + \|y\|_{\mathcal{Y}}^2$, borné uniformément puisque $\varepsilon/\alpha \rightarrow 0$. A Noter que dans la méthode de Tikhonov généralisée, on aura également une borne uniforme sur $\|y_{\varepsilon,\alpha}\|_{\mathcal{Y}}^2$ à condition que B ait un inverse continu. C'est bien le cas pour $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}^p$ avec $p > 0$ et $B = Id$ car alors on a $\|By\|_{\mathcal{Z}_B} = \|y\|_{\mathcal{Y}^p} \geq \|y\|_{\mathcal{Y}}$ donc $\|B^{-1}\|_{\mathcal{Z}_B \rightarrow \mathcal{Y}_B} \leq 1$.

- Comme Ψ est borné, Ψy_n converge faiblement vers $\Psi \tilde{y}$.
- On montre (par J_α) que $\Psi y_n - z_{\varepsilon_n}$ tend vers 0, or comme z_ε tend vers z , cela implique que $\Psi \tilde{y} = z$, donc $\tilde{y} = y$.

Preuve :

$$\|\Psi y_n - z_{\varepsilon_n}\|_{\mathcal{Z}}^2 \leq 2J_{\alpha(\varepsilon_n)}(y_n) \leq 2J_{\alpha(\varepsilon_n)}(y) = \varepsilon^2 + \alpha(\varepsilon_n)^2 \|y\|_{\mathcal{Y}}^2 \rightarrow 0.$$

- Il reste à montrer que la convergence de y_n n'est pas seulement faible, mais forte. Pour cela il suffit de montrer que $\|y_n\| \rightarrow \|y\|$.

Preuve : pour la \limsup , c'est donné par J_α puisqu'on a vu $\|y_{\varepsilon,\alpha(\varepsilon)}\|_{\mathcal{Y}}^2 \leq \frac{\varepsilon^2}{\alpha(\varepsilon)^2} + \|y\|_{\mathcal{Y}}^2$.

Pour la \liminf c'est lié aux propriétés des suites faiblement convergentes (voir par ex. [23] p.81)

■

On remarque aussi que le théorème de Lax-Milgram est une autre façon de donner une solution unique au problème de minimisation, dès lors que B est inversible (donc en particulier c'est le cas pour la méthode de Tikhonov classique).

De plus pour la méthode classique, en notant y_α la solution avec Ψ^*z au second membre, on a :

$$\|y_\alpha - y_{\varepsilon,\alpha}\|_{\mathcal{Y}} \leq C \frac{\varepsilon}{\alpha}.$$

Ce n'est toujours pas suffisant, bien sûr, pour avoir des vitesses de convergence.

Vitesse de convergence pour la méthode de Tikhonov classique

Maintenant essayons des résultats de convergence utilisant des a priori sur y , pour trouver, si possible, des vitesses de convergence... et faisons-le pour l'instant sans la décomposition en valeurs singulières, uniquement en manipulant l'équation.

Dans le cas (nécessaire pour l'étude) où $z \in \text{Im}(\Psi)$, on utilise toujours la décomposition

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}} \leq \|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{Y}} + \|y_\alpha - y\|_{\mathcal{Y}},$$

où l'on a noté $z = \Psi y$, et y_α l'unique solution de 4.2 avec $B = \text{Id}$ et $\mathcal{Y}_B = \mathcal{Z}_B = \mathcal{Y}$:

$$(\Psi^*\Psi + \alpha^2)y_\alpha = \Psi^*z = \Psi^*\Psi y. \quad (4.6)$$

Lemme 1. Soit Ψ linéaire continu injectif à image dense dans \mathcal{Z} . Soit $y_{\varepsilon,\alpha} \in \mathcal{Y}$ l'unique solution de l'équation (4.3) et y_α l'unique solution de (4.6) pour z , $z_\varepsilon \in \mathcal{Z}$. Alors

1. Pour tout $z_\varepsilon \in \mathcal{Z}$, on a $y_{\varepsilon,\alpha} \in \text{Im}(\Psi^*)$ et

$$\|\Psi y_{\varepsilon,\alpha}\|_{\mathcal{Z}} \leq \|z_\varepsilon\|_{\mathcal{Z}}, \quad \|y_{\varepsilon,\alpha}\|_{\mathcal{Y}} \leq \frac{1}{\alpha} \|z_\varepsilon\|_{\mathcal{Z}}.$$

2. Si de plus $z \in \text{Im}(\Psi)$, alors $y_\alpha \in \text{Im}(\Psi^*\Psi)$, et notant $\Psi y = z$ on a

$$\|y_\alpha\|_{\mathcal{Y}} \leq \|y\|_{\mathcal{Y}},$$

3. Si de plus $y \in \text{Im}(\Psi^*)$, alors $y_\alpha \in \text{Im}(\Psi^*\Psi\Psi^*)$ et notant $\Psi^*v = y$ on a

$$\|y_\alpha - y\|_{\mathcal{Y}} \leq \alpha \|v\|_{\mathcal{Z}}.$$

4. Si de plus $y \in \text{Im}(\Psi^*\Psi)$, alors $y_\alpha \in \text{Im}\left((\Psi^*\Psi)^2\right)$ et notant $\Psi^*\Psi w = y$ on a

$$\|y_\alpha - y\|_{\mathcal{Y}} \leq \alpha^2 \|w\|_{\mathcal{Y}}.$$

Remarque 2. Pour tout ceci, on n'a pas eu besoin de la construction explicite de la décomposition en valeurs singulières, ce qui permet de s'affranchir de l'hypothèse de compacité sur Ψ : cela découle de manipulations sur l'équation.

Ces manipulations sont en fait des généralisations de manipulations très courantes sur les EDP : il s'agit, pour obtenir des estimations, de multiplier l'équation par des fonctions test adaptées.

Notons aussi que même sans faire un recours explicite à la SVD ou à un espace de régularité \mathcal{Y}^s , l'idée sous-jacente aux hypothèses successives $z \in \text{Im}(\Psi)$, $y \in \text{Im}(\Psi^*)$ (ou $z \in \text{Im}(\Psi\Psi^*)$), $y \in \text{Im}(\Psi^*\Psi)$ (ou $z \in \text{Im}(\Psi\Psi^*\Psi)$) est bien celle d'une régularité croissante.

Notons enfin qu'avec ce lemme, aucune hypothèse sur z n'est nécessaire pour estimer le terme de « variance » : en effet, appliquant l'inégalité obtenue en (1.) à $z_\varepsilon - z$, on a

$$\|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{Y}} \leq \frac{\varepsilon}{\alpha},$$

alors qu'au contraire le terme de « biais » ne peut être estimé que sous des hypothèses de régularité plus forte, comme $y \in \text{Im}(\Psi^*)$ ou $y \in \text{Im}(\Psi^*\Psi)$, ce qui, si la décomposition en valeurs singulières s'applique, revient respectivement à $y \in \mathcal{Y}^{s=\frac{1}{2}}$ et $y \in \mathcal{Y}^{s=1}$.

Preuve.

- La première inégalité est immédiatement obtenue en faisant le produit scalaire de l'équation (4.2) contre $y_{\varepsilon,\alpha}$ et en utilisant le premier terme du membre de gauche :

$$\langle \Psi y_{\varepsilon,\alpha}, \Psi y_{\varepsilon,\alpha} \rangle_{\mathcal{Z}} + \alpha^2 \langle y_{\varepsilon,\alpha}, y_{\varepsilon,\alpha} \rangle_{\mathcal{Y}} = \langle z_\varepsilon, \Psi y_{\varepsilon,\alpha} \rangle_{\mathcal{Z}}.$$

La deuxième inégalité en découle en utilisant le membre de droite :

$$\alpha^2 \langle y_{\varepsilon,\alpha}, y_{\varepsilon,\alpha} \rangle_{\mathcal{Y}} = \alpha^2 \|y_{\varepsilon,\alpha}\|_{\mathcal{Y}} \leq |\langle z_\varepsilon, \Psi y_{\varepsilon,\alpha} \rangle_{\mathcal{Z}}| \leq \|z_\varepsilon\|_{\mathcal{Z}} \|\Psi y_{\varepsilon,\alpha}\|_{\mathcal{Z}} \leq \|z_\varepsilon\|_{\mathcal{Z}}^2.$$

- Si $z \in \text{Im}(\Psi)$, $\alpha^2 y_\alpha = \Psi^*\Psi(y - y_\alpha)$, donc $y_\alpha \in \text{Im}(\Psi^*\Psi)$: c'est la propriété régularisante de l'équation. On peut donc faire le produit scalaire de l'équation (4.6) avec u_α tel que $\Psi^*\Psi u_\alpha = y_\alpha$: on obtient

$$\langle \Psi^*\Psi y_\alpha, u_\alpha \rangle_{\mathcal{Y}} + \alpha^2 \langle y_\alpha, u_\alpha \rangle_{\mathcal{Y}} = \langle \Psi^*\Psi y, u_\alpha \rangle_{\mathcal{Y}},$$

soit

$$\|y_\alpha\|_{\mathcal{Y}}^2 + \alpha^2 \|\Psi u_\alpha\|_{\mathcal{Z}}^2 = \langle y, y_\alpha \rangle_{\mathcal{Y}} \leq \|y\|_{\mathcal{Y}} \|y_\alpha\|_{\mathcal{Y}},$$

donc $\|y_\alpha\|_{\mathcal{Y}} \leq \|y\|_{\mathcal{Y}}$.

- Si $y \in \text{Im}(\Psi^*)$, notant $\Psi^*v = y$ on a

$$\alpha^2 y_\alpha = \Psi^*\Psi\Psi^*v - (\Psi^*\Psi)^2 u_\alpha,$$

qui implique $y_\alpha \in \text{Im}(\Psi^*\Psi\Psi^*)$. On peut maintenant réécrire (4.6) sous la forme

$$\Psi^*\Psi(y_\alpha - y) + \alpha^2(y_\alpha - y) = \alpha^2 y = \alpha^2 \Psi^*v,$$

où l'on reconnaît (4.6) avec comme second membre $z = \alpha^2 v$. On peut alors appliquer l'inégalité vue en 1 et on obtient

$$\|y_\alpha - y\|_{\mathcal{Y}} \leq \frac{1}{\alpha} \|\alpha^2 v\|_{\mathcal{Z}} = \alpha \|v\|_{\mathcal{Z}},$$

ce qu'on voulait.

4. Le raisonnement est le même : de même que l'inégalité 3 est l'inégalité 1 appliquée à $\alpha^2 v$ remplaçant z_ε , l'inégalité 4 est l'inégalité 2 appliquée à $\alpha^2 w$ remplaçant y . ■

Avec tout cela, il vient très simplement, en supposant $y \in \text{Im}(\Psi^*)$ () :

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}} \leq \frac{\varepsilon}{\alpha} + \alpha \|(\Psi^*)^\dagger y\|_{\mathcal{Y}},$$

où le premier terme est obtenu en appliquant l'innégalité (1) du lemme précédent à $y_{\varepsilon,\alpha} - y$. On a donc une convergence en $O(\varepsilon^{\frac{1}{2}})$ obtenue pour un $\alpha = O(\varepsilon^{\frac{1}{2}})$.

En supposant $y \in \text{Im}(\Psi^* \Psi)$:

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{Y}} \leq \frac{\varepsilon}{\alpha} + \alpha^2 \|(\Psi^* \Psi)^\dagger y\|_{\mathcal{Y}},$$

et on a une convergence en $O(\varepsilon^{\frac{2}{3}})$ obtenue pour un $\alpha = O(\varepsilon^{\frac{1}{3}})$.

S'agit-il de convergence *optimale* sur les espaces de régularité considérés ? Les résultats ci-dessus ne le prouvent pas positivement : pour ce faire, il faudrait exhiber un exemple pour lequel la convergence serait exactement celle-là et pas plus. La décomposition en valeurs singulières donne une réponse explicite à cette question pour les opérateurs compacts.

4.1.2 Analyse de la méthode de Tikhonov grâce à la décomposition en valeurs singulières

Nous considérons ici la méthode de Tikhonov généralisée avec $B = \text{Id} : \mathcal{Y}_B \longrightarrow \mathcal{Z}_B$ et $\mathcal{Z}_B = \mathcal{Y}_B = \mathcal{Y}^p$, muni de la norme

$$\|y\|_{\mathcal{Y}^p}^2 := \sum_{j=0}^{\infty} \frac{\langle y, e_j \rangle^2}{\sigma_j^{4p}}.$$

Pour appliquer le théorème 7, il faut avoir $\mathcal{Y}^p \subset \mathcal{Y}$ et donc $p \geq 0$. Il s'agit alors d'une régularisation plus forte que (ou, si $p = 0$, égale à) la régularisation de la méthode de Tikhonov classique. Nous verrons ci-dessous de quelle manière on peut généraliser à $p > -\frac{1}{4}$. $B : D(B) \subset \mathcal{Y} \longrightarrow \mathcal{Z}_B$ avec l'espace de départ $D(B)$ muni de la norme sur \mathcal{Y} . On a

$$\forall (y_B \in \mathcal{Z}_B, y \in D(B)), \quad \langle y_B, By \rangle_{\mathcal{Z}_B} = \langle B^* y_B, y \rangle_{\mathcal{Y}},$$

Ici B est l'identité et $D(B) = \mathcal{Z}_B = \mathcal{Y}^p$ donc

$$\langle y_B, y \rangle_{\mathcal{Z}_B} = \sum_{j=0}^{\infty} \frac{\langle y_B, e^j \rangle_{\mathcal{Y}} \langle y, e^j \rangle_{\mathcal{Y}} e^j}{\sigma_j^{4p}} = \langle B^* y_B, y \rangle_{\mathcal{Y}} = \sum_{j=0}^{\infty} \langle y_B, e^j \rangle_{\mathcal{Y}} \langle B^* y, e^j \rangle_{\mathcal{Y}},$$

et donc en identifiant

$$B^* y = \sum_{j=0}^{\infty} \frac{\langle y, e^j \rangle_{\mathcal{Y}}}{\sigma_j^{4p}} e^j.$$

L'équation (4.2) s'écrit comme suit dans la décomposition en valeurs singulières :

$$\sum_{j=0}^{\infty} (\sigma_j^2 + \alpha^2 \sigma_j^{-4p}) \langle y, e^j \rangle_{\mathcal{Y}} e^j = \sum_{j=0}^{\infty} \sigma_j \langle z_\varepsilon, f^j \rangle_{\mathcal{Z}} e^j.$$

On peut interpréter l'équation en terme de méthode de régularisation : rappelons la définition 3.4

$$R_\alpha z := \sum_{j=0}^{\infty} r(\alpha, \sigma_j) \sigma_j^{-1} \langle z, f^j \rangle_{\mathcal{Z}} e^j, \quad \forall y \in \mathcal{Z}.$$

La méthode de Tikhonov revient donc à prendre une méthode de régularisation R_α définie par (3.4) avec $r = r_p$ défini par

$$r_p(\alpha, \sigma_j) := \frac{\sigma_j^2}{\sigma_j^2 + \alpha^2 \sigma_j^{-4p}} = \frac{1}{1 + \alpha^2 \sigma_j^{-4p-2}}.$$

Pour des raisons d'homogénéité avec la méthode de troncature, où l'on avait tronqué pour $\sigma \geq \alpha$, on remplace par la suite α par $\tilde{\alpha} := \alpha^{\frac{1}{2p+1}}$ (α et $\tilde{\alpha}$ varient dans le même sens dès que $p > -\frac{1}{2}$). Dans ce cas on a

$$r_p(\tilde{\alpha}, \sigma) = \frac{1}{1 + \tilde{\alpha}^{4p+2} \sigma_j^{-4p-2}}.$$

Pour vérifier s'il s'agit bien d'une méthode de régularisation, vérifions les deux conditions du théorème 5 - et faisons-le sans a priori sur la valeur de $p \in \mathbb{R}$.

1.

$$\frac{r_p(\tilde{\alpha}, \sigma)}{\sigma} = \frac{1}{\sigma(1 + (\tilde{\alpha}^{-1}\sigma)^{-4p-2})} = \frac{1}{\sigma + \tilde{\alpha}^{4p+2}\sigma^{-4p-1}}.$$

Si $p < -\frac{1}{4}$, $-4p-1 > 0$ donc $\lim_{\sigma \rightarrow 0} \frac{r_p(\tilde{\alpha}, \sigma)}{\sigma} = +\infty$: n'ayant pas de borne uniforme en σ de $\frac{r(\tilde{\alpha}, \sigma)}{\sigma}$, on n'a pas une méthode de régularisation. Nous supposons donc par la suite $p \geq -\frac{1}{4}$. Si $p = -\frac{1}{4}$ on a

$$\frac{r_{-\frac{1}{4}}(\tilde{\alpha}, \sigma)}{\sigma} = \frac{1}{\sigma + \tilde{\alpha}} \leq \frac{1}{\tilde{\alpha}} = c_{-\frac{1}{4}}(\tilde{\alpha}),$$

et la première condition du théorème 5 est vérifiée.

Si $p > -\frac{1}{4}$, la fonction $f_p(\sigma) = \sigma(1 + (\tilde{\alpha}^{-1}\sigma)^{-4p-2})$ est convexe minimale en $\bar{\sigma}_p > 0$ tel que

$$f'_p(\bar{\sigma}) = 0 = 1 - (4p+1)(\tilde{\alpha}^{-1}\bar{\sigma})^{-4p-2}, \quad f_p(\bar{\sigma}) = \bar{\sigma}(1 + \frac{1}{4p+1}) = \frac{4p+2}{4p+1}(4p+1)^{\frac{1}{4p+2}}\tilde{\alpha}.$$

Donc

$$\frac{r_p(\tilde{\alpha}, \sigma)}{\sigma} \leq \frac{1}{\min_{\sigma} \sigma(1 + (\tilde{\alpha}^{-1}\sigma)^{-4p-2})} = \frac{1}{4p+2} \frac{(4p+1)^{\frac{4p+1}{4p+2}}}{\tilde{\alpha}} = c_p(\tilde{\alpha}),$$

et on a $c_p(\tilde{\alpha}) = \frac{C_p}{\tilde{\alpha}}$ avec une constante $C_p > 0$ qui ne dépend que de p (et la formule ci-dessus s'étend d'ailleurs au cas $p = -\frac{1}{4}$).

2. Pour $\sigma > 0$ fixé, on a $\lim_{\tilde{\alpha} \rightarrow 0} r_p(\tilde{\alpha}, \sigma) = \lim_{\tilde{\alpha} \rightarrow 0} \frac{1}{1 + (\tilde{\alpha}\sigma^{-1})^{4p+2}} = 1$ dès que $p > -\frac{1}{2}$, ce qui est le cas puisqu'on a imposé $p > -\frac{1}{4}$ à la condition précédente. De plus on a (utile pour les calculs suivants)

$$r(\tilde{\alpha}, \sigma) - 1 = \frac{1}{1 + (\tilde{\alpha}\sigma^{-1})^{4p+2}} - 1 = \frac{\tilde{\alpha}^{4p+2}}{\sigma^{4p+2} + \tilde{\alpha}^{4p+2}}.$$

Cela montre que dès que $p \geq -\frac{1}{4}$, la méthode de Tikhonov généralisée peut être définie et est une méthode de régularisation convergence.

Etudions maintenant son optimalité. Pour obtenir des taux de convergence de la méthode, il faut, comme d'habitude, estimer au mieux le terme de « biais » $\|y_\alpha - y\|_{\mathcal{Y}}$ et le terme de « variance » $\|y_{\varepsilon, \alpha} - y_\alpha\|_{\mathcal{Y}}$ en fonction d'un a priori.

– Estimation de la « variance » :

$$\|y_{\varepsilon, \alpha} - y_\alpha\|_{\mathcal{Y}}^2 = \sum_{j=0}^{\infty} \frac{r_p(\tilde{\alpha}, \sigma_j)^2}{\sigma_j^2} \langle z_\varepsilon - z, f^j \rangle_{\mathcal{Z}}^2 \leq c_p(\tilde{\alpha})^2 \|z_\varepsilon - z\|_{\mathcal{Z}}^2 = \frac{C_p^2 \varepsilon^2}{\tilde{\alpha}^2}.$$

– Estimation du « biais » : il faut estimer, supposant $y \in \mathcal{Y}^s$ avec $s \geq 0$:

$$\begin{aligned} \|y_\alpha - y\|_{\mathcal{Y}}^2 &= \sum_{j=0}^{\infty} (r_p(\tilde{\alpha}, \sigma_j) - 1)^2 \langle y, e^j \rangle_{\mathcal{Y}}^2 = \sum_{j=0}^{\infty} \sigma_j^{4s} \frac{\tilde{\alpha}^{8p+4}}{(\sigma_j^{4p+2} + \tilde{\alpha}^{4p+2})^2} \frac{\langle y, e^j \rangle_{\mathcal{Y}}^2}{\sigma_j^{4s}} \\ &\leq \|y\|_{\mathcal{Y}^s}^2 \max_{\sigma \in (0, \sigma_0]} \frac{\tilde{\alpha}^{8p+4} \sigma^{4s}}{(\sigma^{4p+2} + \tilde{\alpha}^{4p+2})^2} = \|y\|_{\mathcal{Y}^s}^2 \max_{\sigma \in (0, \sigma_0]} g_{\tilde{\alpha}^{4p+2}}(\sigma^2)^2, \end{aligned}$$

avec

$$g_\beta(x) = \beta \frac{x^s}{x^{2p+1} + \beta}, \quad g'_\beta(x) = \beta \frac{x^{s+2p}(s - (2p+1)) + \beta s x^{s-1}}{(x^{2p+1} + \beta)^2}.$$

On distingue selon les cas.

- $s < 2p + 1$: g_β est croissante sur $[0, \bar{x}]$ et décroissante sur $[\bar{x}, \infty)$ avec

$$\bar{x}^{2p+1} = \frac{s\beta}{2p+1-s}, \quad g_\beta(\bar{x}) = \frac{2p+1-s}{2p+1} \left(\frac{\beta s}{2p+1-s} \right)^{\frac{s}{2p+1}}.$$

Cela entraîne, dès que $\tilde{\alpha}$ est assez petit pour que $\bar{x} < \sigma_0^2$:

$$\max_{\sigma \in (0, \sigma_0]} g_{\tilde{\alpha}^{4p+2}}(\sigma^2)^2 = \left(\frac{2p+1-s}{2p+1} \right)^2 \left(\frac{s}{2p+1-s} \right)^{\frac{2s}{2p+1}} \tilde{\alpha}^{4s} = O\left(\tilde{\alpha}^{4s}\right)$$

- $s \geq 2p + 1$: g_β est croissante sur $[0, \infty)$ donc

$$\max_{\sigma \in (0, \sigma_0]} g_{\tilde{\alpha}^{4p+2}}(\sigma^2)^2 = g_{\tilde{\alpha}^{4p+2}}(\sigma_0^2)^2 = \tilde{\alpha}^{8p+4} \frac{\sigma_0^{4s}}{\sigma_0^{4p+2} + \tilde{\alpha}^{4p+2}} = O\left(\tilde{\alpha}^{8p+4}\right).$$

Finalement on a obtenu l'inégalité

$$\|y_\alpha - y\|_{\mathcal{Y}}^2 \leq C_{s,p} \tilde{\alpha}^{4s} \|y\|_{\mathcal{Y}^s}^2 \quad \text{si } s < 2p + 1, \quad \leq \tilde{\alpha}^{8p+4} \|y\|_{\mathcal{Y}^s}^2 \quad \text{si } s \geq 2p + 1.$$

Qu'est-ce que cela signifie ? Que la vitesse de décroissance du terme de « biais » croît lorsque la régularité s croît, mais seulement jusqu'à $s = 2p + 1$: au-delà, quelle que soit la régularité, le terme de biais se comporte alors en $O\left(\tilde{\alpha}^{8p+4}\right)$, indépendamment de la régularité $s \geq 2p + 1$, et cette régularité fût-elle infinie.

Comment interpréter ceci ? On a vu au chapitre précédent que pour une méthode optimale, on doit avoir le biais en $O(\alpha^{2s})$. Ceci n'est vrai que pour $s < 2p + 1$: au-delà, la régularité supplémentaire de la méthode ne permet pas d'améliorer la convergence, qui va « plafonner ». **La plus grande valeur de s telle que la méthode de régularisation est d'ordre optimal est appelée qualification de la méthode.** Ici, on a donc une qualification égale à 1 pour la méthode de Tikhonov classique, et égale à $2p + 1$ pour la méthode de Tikhonov généralisée. On voit naturellement que la qualification augmente quand p augmente.

Remarquer que les calculs ci-dessus sont génériques : ils se ramènent au calcul

- pour la « variance » : du maximum sur α de $c(\alpha)$ tel que dans le théorème 5 d'une part ;
- pour le « biais », pour un a priori $y \in \mathcal{Y}^s$: au calcul du maximum sur σ de $\sigma^{4s} (r(\alpha, \sigma) - 1)^2$.

Le phénomène de *saturation* ci-dessus implique qu'on aura au mieux, comme taux de convergence pour la méthode de Tikhonov classique, un $O(\varepsilon^{\frac{2}{3}})$, et pour la méthode généralisée un $O\left(\varepsilon^{\frac{4p+2}{4p+3}}\right)$.

Résumons tous ces résultats dans un théorème.

Théorème 9. *La méthode de Tikhonov généralisée est d'ordre optimal pour $s \leq 2p + 1$ et d'ordre $2p + 1$ pour $s > 2p + 1$. On dit que la qualification de cette méthode est $2p + 1$.*

4.2 Exemple de l'inversion de l'opérateur intégral

Nous allons traiter cet exemple de façon directe, sans passer par la décomposition en valeurs singulières. $s = 0$ et notre problème de l'opérateur intégral. Rappelons que l'on prend $\mathcal{Y} = \mathcal{Z} = L^2([0, 1], dx)$, muni du produit scalaire dans L^2 , et pour $y \in \mathcal{Y}$ on définit Ψ par (2.10) :

$$\Psi y : [0, 1] \rightarrow \mathbb{R}, \quad t \mapsto \int_0^t y(s) ds.$$

On calcule donc aisément $\Psi^* : L^2([0, 1], dx) \rightarrow L^2([0, 1], dx)$: soit $y, z \in L^2([0, 1])$. On a

$$\langle \Psi y, z \rangle_{\mathcal{Z}} = \int_0^1 z(t) \int_0^t y(s) ds dt = \int_0^1 y(s) \int_s^1 z(t) dt ds = \langle y, \Psi^* z \rangle_{\mathcal{Y}}$$

$$\text{donc } \Psi^* z(s) = \int_s^1 z(t) dt. \text{ Donc } \Psi^* \Psi y(t) = \int_t^1 \int_0^s y(\sigma) d\sigma ds.$$

4.2.1 Méthode de Tikhonov classique

La méthode de Tikhonov *classique* donnée par (4.3) s'écrit donc : $y_{\varepsilon, \alpha} \in L^2([0, 1])$ est l'unique solution de

$$\int_t^1 \int_0^s y_{\varepsilon, \alpha}(x) dx ds + \alpha^2 y_{\varepsilon, \alpha}(t) = \int_t^1 z_{\varepsilon}(s) ds. \quad (4.7)$$

On remarque qu'une telle solution est par construction non seulement dans L^2 mais dans H^1 , ce qui correspond au lemme 1 puisque $y_{\varepsilon, \alpha} \in Im(\Psi^*) \subset H^1([0, 1])$. Ce lemme ainsi que le théorème 9 nous donnent un taux de convergence de la solution en $O(\varepsilon^{1/2})$ si $y \in Im(\Psi^*) \equiv \mathcal{Y}^{1/2} \subset H^1$ et au plus en $O(\varepsilon^{2/3})$ pour $y \in Im(\Psi^* \Psi) \equiv \mathcal{Y}^1 \subset H^2$.

L'équation (4.7) est formellement équivalente, en dérivant l'équation et en l'écrivant pour $u_{\varepsilon, \alpha}(t) = \int_0^t y_{\varepsilon, \alpha}(s) ds$: à résoudre au sens faible

$$u_{\varepsilon, \alpha}(t) - \alpha^2 u_{\varepsilon, \alpha}''(t) = z_{\varepsilon}(t), \quad u_{\varepsilon, \alpha}'(1) = 0, \quad u_{\varepsilon, \alpha}(0) = 0. \quad (4.8)$$

Les deux conditions au bord sont nécessaires pour que le problème soit bien posé. En fait c'est le problème de Poisson en 1D avec conditions aux limites de type Neumann en 1 et Dirichlet en 0. Cette équation peut se résoudre « directement » par des méthodes classiques (Cauchy-Schwartz, formulation faible et théorème de Lax-Milgram). De façon

grossière, on voit qu'ajouter une dérivée régularise le problème : une solution sera telle que $u''_{\varepsilon,\alpha} = \frac{1}{\alpha^2}(-z_\varepsilon + u_{\varepsilon,\alpha}) \in L^2$ donc $y_{\varepsilon,\alpha} \in H^1$.

Nous donnons ici la preuve directe, sans passer par la théorie de la décomposition en valeurs singulières, car elle est instructive à plusieurs titres : elle rendra plus concrètes les manipulations vues dans le cas général sur Ψ et Ψ^* . De plus, elle utilise des outils très classiques des EDP, que vous avez sans doute déjà vu dans un cours sur l'équation de la chaleur, par exemple, ou les équations elliptiques de façon plus générale. Comme vu précédemment, nous définissons tout d'abord l'espace

$$\tilde{H}^1([0, 1]) := \{u \in H^1([0, 1]), \quad u(0) = 0\}, \quad (4.9)$$

sur lequel nous savons, grâce aux inégalités de Poincaré, que la semi-norme $\|u'\|_{L^2}$ définit une norme équivalente à la norme standard de H^1 . Plus précisément on calcule très simplement (déjà fait au ch.1) que, pour $u \in \tilde{H}^1$,

$$\|u\|_{L^2}^2 \leq \frac{1}{2}\|u'\|_{L^2}^2,$$

et plus précisément, grâce à la décomposition en valeurs singulières :

$$\|u\|_{L^2}^2 \leq \sigma_0^2 \|u'\|_{L^2}^2 = \frac{4}{\pi^2} \|u'\|_{L^2}^2.$$

Le lien entre les équations (4.7) et (4.8) n'étant pour l'instant pas démontré mais seulement intuité, commençons par établir la formulation variationnelle du problème (4.8) à partir de l'équation (4.7). Soit $\phi \in C_b^\infty([0, 1])$. Nous partons de l'équation (4.7) et faisons le produit scalaire dans L^2 .

$$\int_0^1 \left(\phi(t) \int_t^1 \int_0^s y(z) dz ds + \phi(t) \alpha^2 y(t) \right) dt = \int_0^1 \phi(t) \int_t^1 z_\varepsilon(s) ds dt.$$

En intervertissant les intégrales en s et en t par Fubini dans le premier terme et dans le membre de droite :

$$\int_0^1 \left(\int_0^s y(z) dz \right) \left(\int_0^s \phi(t) dt \right) ds + \alpha^2 \int_0^1 \phi(s) y(s) ds = \int_0^1 z_\varepsilon(s) \left(\int_0^s \phi(t) dt \right) ds.$$

Définissant $u(t) = \int_0^t y(s) ds$, on a $u \in \tilde{H}^1$ et en prenant la fonction test $\psi(t) = \int_0^t \phi(s) ds \in C_b^\infty([0, 1])$, on a la formulation faible suivante de l'équation (4.8)

$$\forall \psi \in C_b^\infty([0, 1]), \psi(0) = 0, \quad \int_0^1 u(t) \psi(t) dt + \alpha^2 \int_0^1 \psi'(t) u'(t) dt = \int_0^1 z_\varepsilon(t) \psi(t) dt. \quad (4.10)$$

Par densité de $C_b^\infty([0, 1])$ dans \tilde{H}^1 , et tous les termes étant bien définis, on peut étendre la fonctionnelle à $\psi \in \tilde{H}^1$. On définit alors la forme bilinéaire symétrique continue coercive suivante sur \tilde{H}^1

$$a(u, v) := \int_0^1 u(t)v(t)dt + \alpha^2 \int_0^1 u'(t)v'(t)dt$$

et on reformule le problème sous la forme : trouver $u \in \tilde{H}^1$ tel que pour tout $v \in \tilde{H}^1$, on a

$$a(u, v) = \ell_\varepsilon(v),$$

où ℓ_ε est la forme linéaire définie sur \tilde{H}^1 par $\ell_\varepsilon(v) = \int_0^1 z_\varepsilon(t)v(t)dt$. On voit qu'on a toutes les hypothèses pour appliquer le théorème de Lax-Milgram.

Rappel sur le théorème de Lax-Milgram soit H un espace de Hilbert. Une forme bilinéaire continue est dite *coercive* s'il existe une constante $C > 0$ telle que

$$\forall u \in H, \quad a(u, u) \geq C\|u\|_H.$$

Théorème 10 (Théorème de Lax-Milgram). *Soit H un espace de Hilbert, $a(\cdot, \cdot)$ une forme bilinéaire continue et coercive sur H , $\ell(\cdot)$ une forme linéaire continue sur H . Alors il existe une unique solution $u \in H$ au problème*

$$a(u, v) = \ell(v), \quad \forall v \in H.$$

En outre, la solution u dépend continûment de la forme linéaire ℓ :

$$\|u\|_H \leq \frac{1}{C}\|\ell\|_{H'}.$$

(la preuve repose sur le théorème de représentation de Riesz et le point fixe de Banach).

Nous avons maintenant tous les éléments permettant d'énoncer le théorème suivant, qui est en quelque sorte le pendant - partiel - du lemme 1.

Théorème 11. *Soit $z_\varepsilon \in L^2([0, 1])$ et soit $\alpha > 0$. Le problème (4.8) admet une unique solution $u_{\varepsilon, \alpha} \in \tilde{H}^1(\mathbb{R}) \cap H^2(\mathbb{R})$ et cette solution vérifie les inégalités suivantes.*

$$\|u_{\varepsilon, \alpha}\|_{L^2} \leq \|z_\varepsilon\|_{L^2}, \quad \|u'_{\varepsilon, \alpha}\|_{L^2} \leq \frac{C}{\alpha}\|z_\varepsilon\|_{L^2}, \quad \|u''_{\varepsilon, \alpha}\|_{L^2} \leq \frac{2}{\alpha^2}\|z_\varepsilon\|_{L^2}.$$

Preuve. Le théorème de Lax-Milgram nous assure de l'existence et de l'unicité de la solution $u_{\varepsilon, \alpha} \in \tilde{H}^1$ de la formulation faible (4.10). On obtient la formulation au sens p.p. (4.8) en faisant une IPP sur (4.10) : cela entraîne qu'au sens des distributions on a (4.8), et donc que $u''_{\varepsilon, \alpha} = \frac{1}{\alpha^2}(-z_\varepsilon + u_{\varepsilon, \alpha}) \in L^2$, et donc l'égalité au sens des distributions devient vraie p.p. et l'on peut donner un sens aux conditions au bord. De plus on a les inégalités suivantes.

- Avec $\psi = u_{\varepsilon,\alpha}$ dans (4.10) :

$$\|u_{\varepsilon,\alpha}\|_{L^2} \leq \|z_\varepsilon\|_{L^2}.$$

- En notant que $u''_{\varepsilon,\alpha} = \frac{1}{\alpha^2}(u_{\varepsilon,\alpha} - z_\varepsilon)$ et en utilisant l'inégalité ci-dessus :

$$\|u''_{\varepsilon,\alpha}\|_{L^2} \leq \frac{2}{\alpha^2} \|z_\varepsilon\|_{L^2}.$$

- En reprenant la fonction test $\psi = u_{\varepsilon,\alpha}$ dans (4.10) :

$$\|u'_{\varepsilon,\alpha}\|_{L^2}^2 \leq \frac{1}{\alpha^2} \|z_\varepsilon\|_{L^2} \|u_{\varepsilon,\alpha}\|_{L^2} \leq \frac{1}{\alpha^2} \|z_\varepsilon\|_{L^2}^2,$$

et en prenant la racine. ■

Soit $y_{\varepsilon,\alpha} := u'_{\varepsilon,\alpha} \in L^2$. On voit que le théorème nous permet d'écrire

$$\|y_{\varepsilon,\alpha} - y_\alpha\|_{L^2} \leq \frac{\varepsilon}{\alpha},$$

par la deuxième inégalité (pendant de la deuxième inégalité du lemme 1, point 1.). Pour estimer $y_\alpha - y = u'_\alpha - z'$, on applique le théorème 11 à z , ce qui entraîne l'existence et l'unicité d'une solution $u_\alpha \in \tilde{H}^1 \cap H^2$ et de plus par la deuxième inégalité on a

$$\|u'_\alpha\|_{L^2} \leq \frac{C}{\alpha} \|z\|_{L^2}.$$

Pour utiliser cette inégalité, on voudrait pouvoir dériver l'équation (4.8) pour écrire

$$y_\alpha - y = u'_\alpha - z' = \alpha^2 u''_\alpha,$$

et on voudrait pouvoir écrire

$$\|u'''_\alpha\|_{L^2} \leq \frac{C}{\alpha} \|z''\|_{L^2},$$

c'est-à-dire appliquer le théorème avec z'' au lieu de z au second membre. On obtiendrait alors l'équivalent du point 3. du lemme 1. Mais pour pouvoir faire cela, il y a un problème de conditions aux bords. Rappelez-vous : l'hypothèse pour le cas général, dans le lemme 1, était au minimum $y \in Im(\Psi^*)$, soit encore que $z \in Im(\Psi\Psi^*)$, ce qui signifie ici que $z(0) = 0$ et $z'(1) = 0$. Sous ces hypothèses, on aura bien (soit en appliquant le lemme 1, soit en appliquant le théorème 11 avec z'' comme second membre, en trouvant une solution θ_α , puis en intégrant deux fois, d'abord entre t et 1 puis entre 0 et t , de façon à retomber sur u_α , ce qui entraînera bien $\theta_\alpha = u''_\alpha$ par unicité, et donc $y - y_\alpha = \alpha^2 u'''_\alpha = \alpha^2 \theta'_\alpha$)

$$\|y - y_\alpha\|_{L^2} \leq C\alpha \|z''\|_{L^2}.$$

4.2.2 Méthode de Tikhonov généralisée

Au lieu de prendre un terme de régularisation avec l'identité, regardons ce qui se passe pour le cas limite $p = -\frac{1}{4}$: alors $B^*Bx = \sum_{j=0}^{\infty} \langle y, e^j \rangle \sigma_j e^j$, ce qui revient « presque » à l'opérateur Ψ ou à l'opérateur Ψ^* (pas tout à fait cependant en raison du changement des bases, qui revient à un changement de condition au bord). On s'attend à une convergence au plus en $O(\varepsilon^{\frac{1}{2}})$ possible pour $y \in \mathcal{Y}^s$ avec $s = 2p + 1 = \frac{1}{2}$ soit $y \in \tilde{H}^1$. De plus, on remplace α par $\tilde{\alpha} = \alpha^{\frac{1}{2p+1}} = \alpha^2$. Il est donc tout naturel, en considérant qu'un opérateur intégral est « presque » égal à B^*B , de considérer par exemple l'opérateur suivant

$$\alpha \int_t^1 y_{\varepsilon, \alpha}(s) ds + \int_t^1 \int_0^t y_{\varepsilon, \alpha}(s) ds dz = \int_t^1 z_{\varepsilon}(s) ds,$$

que l'on peut dériver, ce qui donne

$$\alpha y + \int_0^t y(s) ds = z_{\varepsilon}, \quad (4.11)$$

soit en dérivant une seconde fois (attention : au sens faible car z'_{ε} n'est plus une fonction)

$$\alpha y'_{\varepsilon, \alpha} + y_{\varepsilon, \alpha} = z'_{\varepsilon}, \quad y(0) = 0. \quad (4.12)$$

Nous sommes partis de la décomposition en valeurs singulières pour proposer cette régularisation et la raccrocher à la théorie de la régularisation de Tikhonov. Ce n'est évidemment pas le plus naturel ! Régulariser par l'ajout d'un petit terme impliquant une dérivée d'ordre supérieur est une technique transverse aux EDP. Par exemple par l'ajout d'un petit terme de diffusion, en utilisant l'effet régularisant de l'équation de la chaleur, etc. Remarquons encore que l'on peut interpréter l'équation comme étant un cas particulier d'une régularisation de type, lorsque $\mathcal{Y} = \mathcal{Z}$:

$$\Psi y_{\varepsilon, \alpha} + \alpha y_{\varepsilon, \alpha} = z_{\varepsilon}.$$

Par rapport à la régularisation de Tikhonov classique, on voit que l'équation est « moins » régularisante : pour $z_{\varepsilon} \in L^2$ on aura seulement $y_{\varepsilon, \alpha} \in L^2$ et non dans H^1 .

On a l'estimation suivante.

Lemme 2. *Si $y_{\varepsilon, \alpha}$ est solution de (4.12), on a*

$$\|y_{\varepsilon, \alpha}\|_{L^2} \leq \frac{1}{\alpha} \|z_{\varepsilon}\|_{L^2}.$$

En prenant dans (4.12) $z_\varepsilon = z \in H^2$, et y_α la solution correspondante, notant $y = z'$ et supposant $y(0) = 0$, on a

$$\|y_\alpha\|_{L^2} \leq \|z'\|_{L^2}, \quad \|y_\alpha - y\|_{L^2} \leq \alpha \|z''\|_{L^2}.$$

Preuve. La première inégalité est obtenue de façon immédiate en multipliant l'équation (sous forme intégrale) contre le poids y et en intégrant, puis en négligeant le second terme du membre de gauche.

$$\alpha \int_0^1 y^2(s) ds + \frac{1}{2} \int_0^1 \left(\int_0^t y(s) ds \right)^2 dt = \int_0^1 z_\varepsilon(t) y(t) dt.$$

La seconde inégalité est obtenue en multipliant l'équation (sous forme (4.12)) par y_α et en intégrant, et en négligeant cette fois le premier terme du membre de gauche. Enfin, la troisième inégalité dérive des deux premières, en considérant que $\bar{y} = y_\alpha - y$ est solution de l'équation (4.12) avec comme second membre $-\alpha y' = -\alpha z''$. ■

Remarque : On peut aussi, à partir de ces inégalités, tâcher de les retrouver sur l'équation de Tikhonov générale $\Psi^* \Psi y + \alpha B^* B y = A^* z$, sans pour autant faire le calcul de la décomposition en valeurs singulières. En effet, prenons $B^* B = (\Psi^* \Psi)^{\frac{1}{2}}$ (défini par la décomposition en valeurs singulières) on a tout d'abord, en appliquant à l'équation l'opérateur $(\Psi^* \Psi)^{-\frac{1}{2}}$:

$$(\Psi^* \Psi)^{\frac{1}{2}} y + \alpha y = (\Psi^* \Psi)^{-\frac{1}{2}} \Psi^* z,$$

d'où l'on déduit la première inégalité, en faisant le produit scalaire contre y et en utilisant la positivité de $(\Psi^* \Psi)^{\frac{1}{2}}$. On n'a pas besoin d'espace particulier pour z . Puis on applique une seconde fois l'opérateur $(\Psi^* \Psi)^{-\frac{1}{2}}$ (sous condition que $z \in \text{Im}(\Psi)$) :

$$y + \alpha (\Psi^* \Psi)^{-\frac{1}{2}} y = (\Psi^* \Psi)^{-1} \Psi^* z = \Psi^\dagger z,$$

et on fait de nouveau le produit scalaire contre y ce qui donne la seconde inégalité.

Pour la troisième inégalité on procède encore de même : on regarde l'équation satisfaite par $y_\alpha - y$ avec comme second membre $-\alpha (\Psi^* \Psi)^{\frac{1}{2}} y$ remplaçant $\Psi^* z$, donc $\|(\Psi^* \Psi)^{-\frac{1}{2}} y\|_{\mathcal{Y}}$ remplaçant $\|\Psi^{-1} z\|_{\mathcal{Z}}$.

De tout cela on déduit, avec les notations habituelles :

$$\|y_{\varepsilon, \alpha} - y\|_{L^2} \leq \|y_{\varepsilon, \alpha} - y_\alpha\|_{L^2} + \|y_\alpha - y\|_{L^2} \leq \frac{\varepsilon}{\alpha} + \alpha \|y'\|_{L^2},$$

par la première et la troisième inégalité. On a exactement ce qu'on avait pensé : il faut $y \in H^1$, et on obtient une inégalité optimale pour $\alpha = O(\sqrt{\varepsilon})$.

Exo : que se passe-t-il si $y(0) \neq 0$? A-t-on encore une estimation possible? Quelle convergence? Peut-on l'interpréter?

4.3 Résolution algorithmique

4.3.1 Exemple de l'opérateur intégral

Pour le cas de l'opérateur intégral ci-dessus, ou pour d'autres opérateurs spécifiques, on peut utiliser l'équation approchée

$$\Psi^* \Psi x + \alpha B^* B x = \Psi^* y,$$

sous l'une ou l'autre de ses formes, pour une résolution numérique directe : ainsi ici, on peut utiliser une méthode adéquate (volumes finis, éléments finis...) pour estimer la dérivée.

C'est numériquement très efficace dans ce cas car **il suffit de faire une seule simulation du problème.**

Par exemple :

$$\alpha \frac{y_{i+1} - y_i}{\Delta x} + y_i = \frac{z_{i+1} - z_i}{\Delta x},$$

qui sera stable à condition d'avoir (condition de type CFL) $\Delta x < \alpha$.

Exercice : l'implémenter et le tester numériquement !

4.3.2 Descente de gradient

La fonctionnelle J_α définie par (4.1) peut être définie aussi pour un opérateur non linéaire : supposons que Ψ soit non linéaire Fréchet dérivable et notons $\Psi'(x)$ sa dérivée de Fréchet, on a

$$J'_\alpha(y) = \Psi'(y)^* \Psi y + \alpha^2 y - \Psi'(y)^* z_\varepsilon.$$

Pour minimiser la fonctionnelle J_α , on s'appuie sur le fait que le gradient désigne - pour la norme considérée - la direction de plus grande pente de la fonctionnelle (*steepest descent*). Il est donc naturel de définir comme suit un algorithme.

- Initialisation : $y_0 = 0$ par exemple
- $y_{k+1} = y_k - p_k J'_\alpha(y_k)$

Toutes ces méthodes peuvent être qualifiées de méthodes « de type adjoint » : elles font intervenir la définition et le calcul de l'adjoint $\Psi'(y)^*$.

Deux questions se posent : le choix de $p_k > 0$ et le moment où l'on arrête l'algorithme.

Il faut que p_k soit assez petit pour que la fonctionnelle décroisse bien sur la droite reliant y_k et y_{k+1} , et assez grand pour que l'algorithme converge à une vitesse satisfaisante.

L'algorithme, de son côté, doit s'arrêter ou bien lorsque le gradient devient trop faible, ou bien lorsque la valeur de $J_\alpha(y_k)$ paraît suffisamment petite.

On peut rechercher p_k de façon optimale : on choisit alors

$$p_k = \underset{p>0}{\operatorname{argmin}} J_\alpha(y_k - p J'_\alpha(y_k)) := f_k(p).$$

En calculant la dérivée de $f_k(p)$ (**exo : le faire**) on remarque que cela conduit à un point y_k où $J'_\alpha(y_{k+1})$ sera orthogonal à $J'_\alpha(y_k)$.

On peut aussi choisir p_k constant. On a alors le résultat suivant.

Théorème 12. *Avec les hypothèses et les notations vues précédemment (en particulier Ψ injectif compact à image dense), si on choisit $p_k = \frac{1}{\alpha^2 + \sigma_0^2}$, on a*

$$\|y_{k+1} - y_{\varepsilon, \alpha}\|_{\mathcal{Y}} \leq \|y_k - y_{\varepsilon, \alpha}\|_{\mathcal{Y}} \frac{\sigma_0^2}{\alpha^2 + \sigma_0^2} = \frac{\kappa}{\kappa + 1} \|y_k - y_{\varepsilon, \alpha}\|_{\mathcal{Y}},$$

où $\kappa = \frac{\sigma_0^2}{\alpha^2}$ peut être vu comme lié au conditionnement de l'opérateur $\Psi^* \Psi + \alpha^2 Id$: en norme d'opérateur, on définit pour une application continue inversible et les normes subordonnées :

$$\kappa(L) := \|L\| \cdot \|L^{-1}\|.$$

Preuve.

1. Estimation : on a par définition $\langle y_{\varepsilon, \alpha}, e^j \rangle = \frac{\sigma_j}{\alpha^2 + \sigma_j^2} \langle z_\varepsilon, f^j \rangle$ donc on calcule

$$\begin{aligned} y_{k+1} - y_{\varepsilon, \alpha} &= y_k - y_{\varepsilon, \alpha} - \frac{1}{\alpha^2 + \sigma_0^2} (\Psi^* \Psi y_k + \alpha^2 y_k - \Psi^* z_\varepsilon) \\ &= y_k - y_{\varepsilon, \alpha} - \frac{1}{\alpha^2 + \sigma_0^2} \sum_{j=0}^{\infty} \left((\sigma_j^2 + \alpha^2) \langle y_k, e^j \rangle - \sigma_j \langle z_\varepsilon, f^j \rangle \right) e^j \\ &= y_k - y_{\varepsilon, \alpha} - \frac{1}{\alpha^2 + \sigma_0^2} \sum_{j=0}^{\infty} \left((\sigma_j^2 + \alpha^2) \langle y_k, e^j \rangle - (\alpha^2 + \sigma_j^2) \langle y_{\varepsilon, \alpha}, e^j \rangle \right) e^j, \\ &= \sum_{j=0}^{\infty} \left(1 - \frac{\alpha^2 + \sigma_j^2}{\alpha^2 + \sigma_0^2} \right) \langle y_k - y_{\varepsilon, \alpha}, e^j \rangle e^j \\ &= \sum_{j=0}^{\infty} \left(\frac{\sigma_0^2 - \sigma_j^2}{\alpha^2 + \sigma_0^2} \right) \langle y_k - y_{\varepsilon, \alpha}, e^j \rangle e^j \end{aligned},$$

ce qui prouve l'inégalité recherchée.

2. Conditionnement de l'opérateur $\Psi^* \Psi + \alpha^2 Id$: la norme d'opérateur est donnée par la plus grande valeur singulière de l'opérateur, qui est donc $\alpha^2 + \sigma_0^2$ pour $\Psi^* \Psi + \alpha^2 Id$ (proche de σ_0^2 pour α petit) et $\frac{1}{\alpha^2}$ pour $(\Psi^* \Psi + \alpha^2 Id)^{-1}$. Plus le conditionnement est grand, plus la vitesse de convergence sera lente car plus le rapport $\frac{\kappa}{\kappa + 1}$ sera proche de 1.

■

On appelle *méthode de type adjoint* une méthode qui nécessite le calcul de l'adjoint de l'opérateur - ici, pour calculer J'_α il faudra calculer Ψ'^* . Cela peut s'avérer numériquement coûteux. On peut avoir intérêt à changer de produit scalaire de façon à réduire le conditionnement : on fait par exemple appel à un pré-conditionneur.

4.3.3 Méthode de Landweber / descente de gradient à pas constant

En pratique, si le terme de régularisation correspond bien non pas à un a priori (lié par exemple à des mesures précédentes) mais à une volonté de régularisation, autant résoudre de façon itérative le vrai problème des moindres carrés plutôt qu'un problème approché : il faudra simplement définir une règle d'arrêt de la récurrence qui évite l'explosion, liée au caractère mal posé. En effet, l'itération peut être vue elle-même comme une régularisation.

Supposons donc $p_k = p > 0$ constant et $\alpha = 0$: on souhaite résoudre itérativement l'équation normale

$$\Psi^* \Psi y = \Psi^* z.$$

Un autre point de vue sur la descente de gradient est de réécrire l'équation comme un point fixe

$$y = y - p \left(\Psi^* \Psi y - \Psi^* z \right) = \left(Id - p \Psi^* \Psi \right) y + p \Psi^* z,$$

i.e. la solution y de l'équation normale est aussi un point fixe de l'application affine

$$Id - p \Psi^* \Psi + p \Psi^* z.$$

Si p est tel que $\|Id - p \Psi^* \Psi\| < 1$,

on a une contraction, et l'itération à pas constant vue ci-dessus (également appelée itération de Landweber) peut être vue comme l'itération usuelle pour le point fixe d'une application contractante :

$$y_{k+1} = y_k - p \Psi^* \Psi y_k + p \Psi^* z.$$

En réalité, pour $p \leq \frac{1}{\sigma_0^2}$, on a bien que l'application $Id - p \Psi^* \Psi$ est contractante mais pas strictement contractante : en prenant la suite des vecteurs singuliers on se rend compte que

$$\forall 0 < p \leq \frac{1}{\sigma_0^2}, \quad \|Id - p \Psi^* \Psi\| = 1.$$

On n'aura donc pas nécessairement convergence de l'algorithme. On peut cependant montrer qu'on a bien convergence si $z \in D(\Psi^\dagger)$.

Lemme 3. *Si $z \in D(\Psi^\dagger)$ pour Ψ linéaire borné, alors $y_k \rightarrow \Psi^\dagger z$.*

Preuve. Voir [10], lemme 3.2.1. p.62. ■

Comme toujours, ce résultat générique ne donne pas de vitesse de convergence. L'idée est maintenant, comme il n'y a pas de régularisation sur l'équation, d'arrêter l'itération à un moment satisfaisant, *i.e.* k joue le rôle de l'inverse du paramètre de régularisation : on définit donc y_k et $y_{\varepsilon,k}$ par

$$\begin{cases} y_{k+1} &= y_k - p\Psi^*(\Psi y_k - z), \\ y_{\varepsilon,k+1} &= y_{\varepsilon,k} - p\Psi^*(\Psi y_{\varepsilon,k} - z_\varepsilon), \end{cases}$$

avec la même initialisation, par exemple $y_0 = y_{\varepsilon,0} = 0$. On montre alors que

$$\|y_k - y_{\varepsilon,k}\|_{\mathcal{Y}} \leq \sqrt{k} \|z - z_\varepsilon\|_{\mathcal{Z}}.$$

Preuve. [th. 6.1. de [27]] La preuve est fondée sur le fait qu'on prouve facilement, par récurrence

$$y_k = \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j p\Psi^* z.$$

En effet, pour $k = 0$, la somme est vide - $y_0 = 0$. Pour $k = 1$, $y_0 = p\Psi^* z$. En supposant l'hypothèse de récurrence vraie au rang k et en utilisant la définition récursive de y_{k+1} on obtient

$$y_{k+1} = (Id - p\Psi^*\Psi)y_k + p\Psi^* z = \sum_{j=1}^k (Id - p\Psi^*\Psi)^j p\Psi^* z + p\Psi^* z = \sum_{j=0}^k (Id - p\Psi^*\Psi)^j p\Psi^* z.$$

On cherche donc à estimer la norme de l'opérateur $\Phi_k = \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j p\Psi^*$. On utilise les expressions suivantes

$$\begin{aligned} \Phi_k^* &= p\Psi \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j, \\ \Phi_k \Phi_k^* &= \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j p\Psi^* p\Psi \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j, \\ p\Psi^* p\Psi \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j &= Id - (Id - p\Psi^*\Psi)^k, \end{aligned}$$

et donc, du fait que $\|(Id - (Id - p\Psi^*\Psi)^k)\| \leq 1$ pour $p \leq \sigma_0^{-2}$, on a

$$\|\Phi_k\|^2 = \|\Phi_k \Phi_k^*\| = \left\| \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j \left(Id - (Id - p\Psi^*\Psi)^k \right) \right\| \leq \left\| \sum_{j=0}^{k-1} (Id - p\Psi^*\Psi)^j \right\|,$$

et le membre de droite est borné par k . ■

On peut encore prouver (voir [27] ch. 6.1. pour une preuve complète) que le *discrepancy principle* qui choisit y_k comme approximation de y avec

$$k = \operatorname{argmin}_{l \in \mathbb{N}} |\Psi y_{\varepsilon,k} - y_\varepsilon| \leq \tau \varepsilon$$

converge pour $\tau \in (1, 2)$ et constitue une méthode *adaptative* d'ordre optimal. Il faut prendre $\tau > 1$, sans quoi la méthode peut diverger et la limite ne jamais être atteinte.

Chapitre 5

L'assimilation de données

5.1 Retour en dimension finie

Nous considérons tout d'abord le cas de la dimension finie qui nous permet de mener simplement toute une série de calculs. Nous verrons alors dans un second temps ce qui peut se généraliser à la dimension infinie.

5.1.1 Reformulation du cas statique linéaire

Nous reprenons le cadre d'inversion évoqué aux chapitres précédents avec un espace d'état $\mathcal{Y} = \mathbb{R}^{N_{\text{state}}}$ et un espace d'observation $\mathcal{Z} = \mathbb{R}^{N_{\text{obs}}}$. L'opérateur à inverser est pour l'instant l'opérateur d'observation

$$C : \begin{cases} \mathcal{Y} \rightarrow \mathcal{Z} \\ y \mapsto z(y) = C(y) \end{cases} \quad (5.1)$$

qui permet, à partir d'un état $y \in \mathbb{R}^{N_{\text{state}}}$ de générer une mesure correspondante.

En dimension finie, la question de l'inversibilité de C se pose en terme de rang de l'opérateur.

Si $\text{rank}(C) \geq N_{\text{state}}$, , C est injectif donc C^*C est inversible. On déduit de l'équation normale (2.9) et appliquée ici à C que

$$\bar{y} = (C^*C)^{-1}C^*z^\varepsilon.$$

Evidemment en pratique l'adjoint en dimension finie est directement lié à la transposition de l'opérateur C interprété de façon matricielle. Plus précisément, équipons \mathcal{Z} de la norme

$$\forall z \in \mathcal{Z}, \quad \|z\|_{\mathcal{Z},M}^2 = z^\top M z$$

où $M \in \mathbb{S}_{N_{\text{obs}}}^{+*}(\mathbb{R})$ est symétrique définie positive. Cette norme est évidemment équivalente à la norme euclidienne mais pondère différemment les éléments de \mathcal{Z} . L'inverse généralisé s'écrit alors

$$\bar{y} = (C^\top MC)^{-1}MC^\top z^\varepsilon.$$

Cet inverse généralisé correspond à l'estimateur moindres carrés de la fonctionnelle convexe

$$\min_{y \in \mathcal{Y}} \left\{ J(y) = -\frac{1}{2} \|z^\varepsilon - Cy\|_{\mathcal{Z}, M}^2 \right\}$$

On le retrouve facilement en calculant la différentielle $J'(y) \in \mathcal{L}(\mathcal{Y})$ de J par rapport à y , nous obtenons

$$\forall \delta y, \quad J'(y)(\delta y) = (z^\varepsilon - Cy)^\top MC\delta y,$$

tel que

$$\bar{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} J(y) \Leftrightarrow J'(\bar{y}) = 0.$$

Si $\operatorname{rank}(C) < N_{\text{state}}$, alors nous pouvons ajouter une information *a priori* sur l'état pour palier le manque d'injectivité et sélectionner un estimateur. Sur l'espace d'état on considère $N_\diamond \in \mathbb{S}_{N_{\text{state}}}^{+*}(\mathbb{R})$ afin de définir la norme

$$\forall y \in \mathcal{Y}, \quad \|y\|_{\mathcal{Y}, N_\diamond}^2 = y^\top N_\diamond y.$$

L'estimateur au sens de Tikhonov qui généralise la notion d'estimateur moindres carrés permet donc éventuellement de régulariser mais aussi de palier le manque d'information

$$\min_{y \in \mathcal{Y}} \left\{ J(y) = \frac{1}{2} \|y_\diamond - y\|_{\mathcal{Y}, N_\diamond}^2 + \frac{1}{2} \|z^\varepsilon - Cy\|_{\mathcal{Z}, M}^2 \right\}, \quad (5.2)$$

Dans ce cas, on a

$$\forall \delta y, \quad J'(y)(\delta y) = (y_\diamond - y)^\top N_\diamond \delta y + (z^\varepsilon - Cy)^\top MC\delta y,$$

On obtient alors

$$\bar{y} = y_\diamond + (N_\diamond + C^\top MC)^{-1}C^\top M(z^\varepsilon - Cy_\diamond).$$

En introduisant l'opérateur défini positif suivant

$$P = (N_\diamond + C^\top MC)^{-1}, \quad (5.3)$$

puis l'*opérateur de Gain*

$$G = PC^\top M, \quad (5.4)$$

on a alors une écriture

$$\bar{y} = y_\diamond + G(z^\varepsilon - Cy_\diamond) = y_\diamond + PC^\top M(z^\varepsilon - Cy_\diamond). \quad (5.5)$$

qui s'interprète en disant que les mesures ont corrigé l'*a priori* initiale.

Un calcul source d'inspiration pour la suite montre que la connaissance *a priori* et les observations peuvent être traitées de manière symétrique car elles représentent toutes les deux des sources d'information sur le système. En effet notre problème revient à estimer un état \check{y} à partir d'informations « bruitées »

$$\begin{cases} y_\varepsilon = y_\diamond = \check{y} + \text{erreur} \\ z^\varepsilon = Cy_\diamond + \text{erreur} \end{cases} \quad (5.6)$$

On peut alors se ramener au cas injectif en considérant

$$\Psi : \begin{cases} \mathcal{Y} \rightarrow \mathcal{Y} \times \mathcal{Z} \\ y \mapsto {}^\natural z = \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ C \end{pmatrix} y \end{cases} \quad (5.7)$$

L'opérateur pondérant la norme dans $\text{Im}(\Psi)$ devient alors la concaténation des opérateurs N et M pondérant respectivement les normes euclidiennes dans \mathcal{Y} et \mathcal{Z} , c'est à dire

$${}^\natural M = \begin{pmatrix} N_\diamond & 0 \\ 0 & M \end{pmatrix}.$$

Le critère avec *a priori* se réécrit sous la forme $J(y) = \frac{1}{2} \| {}^\natural z^\varepsilon - \Psi y \|_{\mathcal{Y} \times \mathcal{Z}, {}^\natural M}^2$.

Avant d'aller plus loin, nous proposons une réécriture de l'estimateur s'appuyant sur le fameux lemme d'inversion matricielle que nous rappelons ici

Lemme 4 (Lemme d'inversion matricielle). *Soit M_1, M_{12}, M_{21}, M_2 des matrices telles que M_1, M_2 et $M_2 - M_{21}M_1^{-1}M_{12}$ sont inversibles, alors $M_1 - M_{12}M_2^{-1}M_{21}$ est aussi inversible et on a*

$$(M_1 - M_{12}M_2^{-1}M_{21})^{-1} = M_1^{-1} + M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1}.$$

Preuve. La preuve la plus classique consiste à comparer les blocs lors de l'inversion du système $Y = MX$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} M_1 & M_{12} \\ M_{21} & M_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

En effet, ce système peut se réduire à une inversion sur X_1 – respectivement sur X_2 – par l'introduction du complément de Schur suivant $M_1 - M_{12}M_2^{-1}M_{21}$ – respectivement

$M_2 - M_{21}M_1^{-1}M_{12}$. Sachant que la solution de $Y = MX$ est unique, on identifie alors les blocs $(1, 1)$ de M^{-1} .

Une autre preuve plus directe consiste à calculer

$$\begin{aligned} & (M_1^{-1} + M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1})(M_1 - M_{12}M_2^{-1}M_{21}) = \\ & \quad \mathbb{1} + M_1^{-1}M_{12}M_2^{-1}M_{21} + \\ & \quad M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1}(M_1 - M_{12}M_2^{-1}M_{21}), \end{aligned}$$

Ici le dernier terme se simplifie avec l'avant dernier pour donner

$$\begin{aligned} & M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1}(M_1 - M_{12}M_2^{-1}M_{21}) \\ & = M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}(M_{21} - M_{21}M_1^{-1}M_{12}M_2^{-1}M_{21}) \\ & = M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}(M_2 - M_{21}M_1^{-1}M_{12})M_2^{-1}M_{21} \\ & = M_1^{-1}M_{12}M_2^{-1}M_{21}, \end{aligned}$$

autrement dit

$$(M_1^{-1} + M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1})(M_1 - M_{12}M_2^{-1}M_{21}) = \mathbb{1}. \quad \blacksquare$$

L'utilisation du lemme d'inversion matricielle permet alors d'écrire, avec $P_\diamond = N_\diamond^{-1}$ et $W = M^{-1}$,

$$\begin{aligned} G &= (N_\diamond + C^\top MC)^{-1}C^\top M \\ &= (P_\diamond - P_\diamond C^\top(W + CP_\diamond C^\top)^{-1}CP_\diamond)C^\top M \\ &= P_\diamond C^\top(\mathbb{1} - (W + CP_\diamond C^\top)^{-1}CP_\diamond C^\top)M \\ &= P_\diamond C^\top(W + CP_\diamond C^\top)^{-1}(W + CP_\diamond C^\top - CP_\diamond C^\top)M \\ &= P_\diamond C^\top(W + CP_\diamond C^\top)^{-1}, \end{aligned}$$

Nous verrons l'intérêt d'une telle écriture dans la suite mais remarquons tout de suite que si $N_{\text{obs}} \ll N_{\text{state}}$ alors l'inverse $(W + CP_\diamond C^\top)^{-1}$ est moins cher numériquement à calculer que $(N_\diamond + C^\top MC)^{-1}$.

5.1.2 Vers la multiplication des observations

Nous passons désormais au cas où nous avons des mesures répétées sur le système. Autrement dit nous disposons d'une série de mesures $(z_n^\varepsilon)_{n \in \mathbb{N}}$ telle que

$$z_n^\varepsilon = C_n \check{y} + \text{erreur}.$$

et nous aimerais estimer \check{y} . Les opérateurs C_n peuvent changer à chaque itération ou correspondre au même opérateur C . Dans ce cas on envisage typiquement une situation où on a répété l'expérience à différents instants t_n . S'inspirant des calculs précédents, on pose alors dans le cas sans *a priori*

$$\natural z = \begin{pmatrix} z_0 \\ \vdots \\ z_n \end{pmatrix}, \quad \Psi_n = \begin{pmatrix} C_0 \\ \vdots \\ C_n \end{pmatrix},$$

En supposant que les observations ont été obtenues indépendamment les unes des autres, il est légitime de considérer une pondération sur l'espace image de la forme

$$\natural M = \begin{pmatrix} M_0 & & 0 \\ & \ddots & \\ 0 & & M_n \end{pmatrix}.$$

Il s'agit alors de calculer l'inverse généralisé de Ψ_n . On obtient alors

$$\bar{y}_n = \left(\sum_{k=0}^n C_k^\top M_k C_k \right)^{-1} \sum_{k=0}^n C_k^\top M_k z_k^\varepsilon. \quad (5.8)$$

Nous pourrions nous contenter de (5.8). Cependant un des gros inconvénients de cette expression est qu'elle impose de recalculer entièrement l'estimateur pour toute nouvelle observation dont on bénéficierait.

Supposons qu'il existe N_0 tel que

$$\forall n \geq N_0, \quad \Upsilon_n = \sum_{k=0}^n C_k^\top M_k C_k \text{ est défini positif}$$

Nous introduisons alors la série d'opérateurs $(P_n)_{n \in \mathbb{N}}$ définis par

$$P_n^{-1} = \sum_{k=0}^n C_k^\top M_k C_k, \quad (5.9)$$

alors nous voyons que $(P_n)_{n \in \mathbb{N}}$ satisfait la relation de récurrence

$$P_n^{-1} = C_n^\top M_n C_n + P_{n-1}^{-1}. \quad (5.10)$$

L'utilisation du lemme d'inversion matricielle permet de réécrire cette dernière identité. En notant $W_n = M_n^{-1}$, on a en effet

$$P_n = P_{n-1} - P_{n-1} C_n^\top (W_n + C_n P_{n-1} C_n^\top)^{-1} C_n P_{n-1}. \quad (5.11)$$

Ainsi \bar{y}_n se déduit par récurrence de \bar{y}_{n-1} par la relation

$$\begin{aligned}\bar{y}_n &= P_n \left(\sum_{k=0}^n C_k^\top M_k z_k \right) = P_n P_{n-1}^{-1} \bar{y}_{n-1} + P_n C_n^\top M_n z_n^\varepsilon \\ &= \bar{y}_{n-1} + P_n C_n^\top M_n z_n^\varepsilon - G_n C_n \bar{y}_{n-1},\end{aligned}$$

avec la série d'opérateur de gain

$$G_n = P_{n-1} C_n^\top (W_n + C_n P_{n-1} C_n^\top)^{-1}. \quad (5.12)$$

Finalement l'estimateur s'écrit

$$\bar{y}_n = \bar{y}_{n-1} + G_n (z_n^\varepsilon - C_n \bar{y}_{n-1}). \quad (5.13)$$

Il est à noter que le lemme d'inversion matricielle permet de simplifier les expression (5.12), avec formellement

$$\begin{aligned}G_n &= P_n P_{n-1}^{-1} C_n^\top (W_n + C_n P_{n-1} C_n^\top)^{-1} \\ &= P_n (C_n^\top M_n C_n + P_{n-1}^{-1}) P_{n-1} C_n^\top (W_n + C_n P_{n-1} C_n^\top)^{-1} \\ &= P_n (C_n^\top M_n C_n P_{n-1} C_n^\top + C_n^\top) (W_n + C_n P_{n-1} C_n^\top)^{-1} \\ &= P_n C_n^\top M_n (C_n P_{n-1} C_n^\top + W_n) (W_n + C_n P_{n-1} C_n^\top)^{-1} \\ &= P_n C_n^\top M_n.\end{aligned} \quad (5.14)$$

Ainsi on peut formuler des alternatives à (5.10). Par exemple

$$\begin{aligned}P_n &= P_{n-1} - P_{n-1} C_n^\top (W_n + C_n P_{n-1} C_n^\top)^{-1} C_n P_{n-1} \\ &= (\mathbb{1} - G_n C_n) P_{n-1}.\end{aligned} \quad (5.15)$$

L'opérateur P_n étant évidemment symétrique d'après (5.10), la relation (5.15) peut être symétrisée pour donner

$$P_n = (\mathbb{1} - G_n C_n) P_{n-1} (\mathbb{1} - G_n C_n)^\top + G_n W_n G_n^\top. \quad (5.16)$$

Exercice : Appliquer le résultat précédent au calcul de la moyenne de mesures successive sur une grandeur scalaire.

Les calculs formels que nous venons d'effectuer nécessitent en toute rigueur que l'opérateur P_n défini en (5.9) soit inversible ce qui se traduit par la relation de coercivité

$$\exists c > 0 \mid \forall y, \quad \sum_{k=0}^n \|C_k y\|_{\mathcal{Z}, M_k}^2 \geq c \|y\|_{\mathcal{Y}, N}^2. \quad (5.17)$$

Cette relation s'étendra au cadre de la dimension infinie et nous l'appellerons *condition d'observabilité*. Lorsque cette condition n'est pas satisfaite, le simple ajout d'un *a priori* nous redonne en dimension finie la coercivité nécessaire. Dans ce cas l'estimateur devient

$$\bar{y}_n = \left(P_{\diamond}^{-1} + \sum_{k=0}^n C_k^T M_k C_k \right)^{-1} \left(y_{\diamond} + \sum_{k=0}^n C_k^T M_k z_k \right). \quad (5.18)$$

La relation récursive sur les estimateurs successifs subsiste après simple modification de la première itération. On initialise l'opérateur de covariance

$$P_0^- = P_{\diamond}$$

tel que l'estimateur soit

$$y_0^- = y_{\diamond}$$

Puis on corrige immédiatement ce premier estimateur avec

$$\bar{y} = y_{\diamond} + P_0^+ C_0^T M_0 (z^{\varepsilon} - C_0 y_{\diamond}),$$

où

$$(P_0^+)^{-1} = (P_0^-)^{-1} + C_0^T M_0 C_0.$$

Les relations s'enchaînent alors à partir de P_0^+ . Il est à noter que la condition d'observabilité est difficilement satisfaite pour tout n . Le plus souvent on démontre il existe un N tel que pour tout $n \geq N$ l'inégalité (5.17) soit vérifiée. Dans ce cas, l'*a priori* initial permet de manipuler la forme récursive dès $n = 0$, et son impact est évanescence avec l'augmentation du nombre d'observations.

5.1.3 Prise en compte d'une dynamique discrète

On considère désormais un système dynamique discret de la forme

$$\begin{cases} y_{n+1|\zeta} = \Phi_{n+1|n} y_{n|\zeta} + f_n, & n \in \mathbb{N} \\ y_{0|\zeta} = y_{\diamond} + \zeta \end{cases} \quad (5.19)$$

où $\Phi_{n+1|n} \in \mathcal{L}(\mathcal{Y})$ est l'opérateur de transition vers l'état $n+1$ sachant¹ un état à l'instant n . Ce système génère une suite récursive $(y_{n|\zeta})_{n \in \mathbb{N}}$ sachant la composante ζ de la condition initiale fixée. Imaginons maintenant que nous ayons sur une trajectoire particulière ($\bar{y}_n = y_{n|\zeta}$) $_{n \in \mathbb{N}}$, une série d'observations $(z_n^{\varepsilon})_{n \in \mathbb{N}}$ générées à partir d'opérateurs d'observation $(C_n)_{n \in \mathbb{N}}$.

1. le signe | se prononce sachant

Notre objectif est une fois encore de proposer une estimation de $\check{\zeta}$ et surtout de $(\check{y}_n)_{n \in \mathbb{N}}$, à partir des mesures $(z_n^\varepsilon)_{n \in \mathbb{N}}$. Pour ce faire nous conservons l'approche moindres carrés et définissons le critère

$$\min_{\zeta} \left\{ J_n(\zeta) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_\diamond}^2 + \frac{1}{2} \sum_{k=0}^n \|z_k^\varepsilon - C_{k|\zeta} y_{k|\zeta}\|_{\mathcal{Z}, M_k}^2 \right\}.$$

Afin de préciser notre estimateur nous rappelons quelques notations utiles

Définition 8. Pour toute suite récursive issue de (5.19), $(y_{n|\zeta})_{n \in \mathbb{N}}$, nous définissons la résolvante discrète $\mathbb{T} = (\mathbb{T}_{n|m})_{(m,n) \in \mathbb{N}^2}$ par

$$\forall (m, n) \in \mathbb{N}^2, \quad \mathbb{T}_{n|m}(y_{m|\zeta}) = y_{n|\zeta} \quad (5.20)$$

et en particulier

$$\mathbb{T}_{n|0}(y_{0|\zeta}) = y_{n|\zeta}.$$

Ainsi avec la convention $\Phi_{n|n} = \mathbb{1}$, nous avons

$$\forall (m, n) \in \mathbb{N}^2, \quad |m - n| > 1, \quad \Phi_{n|m} = \Phi_{n|n-1} \circ \cdots \circ \Phi_{m+1|m},$$

de telle sorte que pour tout ζ

$$\forall n \in \mathbb{N}^*, \quad y_{n|\zeta} = \mathbb{T}_{n|0}(y_{0|\zeta}) = \Phi_{n|0} y_{0|\zeta} + \sum_{k=0}^{n-1} \Phi_{n|k+1} f_k.$$

Côté observations, remarquons que pour tout $z_n = Cy_{n|\zeta}$, on a l'identité

$$z_n - \sum_{k=0}^{n-1} C \Phi_{n|k} f_k = C \Phi_{n|0} y_{0|\zeta}.$$

Le membre de gauche est une donnée sur l'incertitude ζ observée à travers $C \Phi_{n|0}$. En agrégant les mesures, on a alors

$$\natural z_n^\varepsilon = \begin{pmatrix} y_\diamond \\ z_0^\varepsilon \\ \vdots \\ z_n^\varepsilon - C_n \sum_{k=0}^{n-1} \Phi_{n|k+1} f_k \end{pmatrix},$$

avec

$$\natural M = \begin{pmatrix} N_\diamond & 0 & 0 \\ 0 & M_0 & 0 \\ & \ddots & \\ 0 & 0 & M_n \end{pmatrix}.$$

L'opérateur d'observation qui nous intéresse est par ailleurs

$$\Psi_n = \begin{pmatrix} \mathbb{1} \\ C_0 \\ \vdots \\ C_n \Phi_{n|0} \end{pmatrix}, \quad (5.21)$$

tel que

$$\mathbb{z}_n^\varepsilon = \Psi_n \check{y}_0 + \text{erreur}.$$

Pour tout n , l'estimateur moindres carrés correspond alors à minimiser

$$\left\{ J_n(y_{0|\zeta}) = \frac{1}{2} \|\mathbb{z}_n^\varepsilon - \Psi_n y_{0|\zeta}\|_{\mathbb{Z}, \mathbb{M}_n}^2 \right\}.$$

Aux paragraphes précédents et notamment (5.18), nous avons vu que l'estimateur était donné par

$$\bar{y}_{0|n} = \left(P_\diamond^{-1} + \sum_{k=1}^n (\Psi_{n,k})^\top M_k \Psi_{n,k} \right)^{-1} \left(y_\diamond + \sum_{k=1}^n (\Psi_{n,k})^\top M_k \mathbb{z}_{n,k}^\varepsilon \right).$$

où $\Psi_{n,k}$ correspond au k -ième block de Ψ_n in (5.21) et de même pour $\mathbb{z}_{n,k}^\varepsilon$. Nous avons obtenu qu'il existait une formule récursive telle que

$$\bar{y}_{0|n} = \bar{y}_{0|n-1} + G_{0|n}(\mathbb{z}_{n,n}^\varepsilon - \Psi_{n,n} \bar{y}_{0|n-1}),$$

avec

$$G_{0|n} = P_{0|n}(\Psi_{n,n})^\top M_n,$$

et $P_{0|n}$ donné par (5.10) (ou (5.11)).

Supposons désormais, que plus que la condition initiale, c'est la trajectoire que vous souhaiteriez suivre et imaginons que vous aimerez bénéficier des nouvelles mesures obtenues en « temps réel ». Nous pouvons alors introduire *l'estimateur au sens de Kalman*.

Définition 9. *L'estimateur de Kalman en temps discret $(\hat{y}_n)_{n \in \mathbb{N}}$ du Système (5.19) compte tenu des observations $(z_n^\varepsilon)_{n \in \mathbb{N}}$ est défini à partir de l'estimateur moindres carrés associé à J_n par*

$$\hat{y}_n = \bar{y}_{n|n} = y_n | \underset{\zeta}{\operatorname{argmin}} J_n.$$

Un objectif est maintenant de formuler pour tout n une estimation récursive $\hat{y}_n = \bar{y}_{n|n}$. Pour ce faire, nous introduisons

$$\hat{y}_n^+ = \mathbb{T}_{n|0}(\bar{y}_{0|n}) = \Phi_{n|0} \bar{y}_{0|n} + \sum_{k=0}^{n-1} \Phi_{n|k+1} f_k. \quad (5.22)$$

Et vérifions

$$\hat{y}_n^+ = \hat{y}_n^- + \Phi_{n|0} G_{0|n} (z_n^\varepsilon - C_n \hat{y}_n^-),$$

pour

$$\hat{y}_n^- = \mathbb{T}_{n|0}(\bar{y}_{0|n-1}) = \Phi_{n|n-1} \hat{y}_{n-1}^+ + f_{n-1}, \quad n \geq 1. \quad (5.23)$$

Formellement – *i.e.* en supposant que tous les opérateurs sont inversibles – la formule de récurrence (5.10) nous permet d'obtenir

$$P_{0|n}^{-1} = \Psi_{n,n}^\top M_n \Psi_{n,n} + P_{0,n-1}^{-1} = \Phi_{n|0}^\top C_n^\top M_n C_n \Phi_{n|0} + P_{0|n-1}^{-1}.$$

Ceci suggère d'introduire

$$P_n^+ = \Phi_{n|0} P_{0|n} \Phi_{n|0}^\top, \quad (5.24)$$

et

$$P_n^- = \Phi_{n|n-1} P_{n-1}^+ \Phi_{n|n-1}^\top, \quad (5.25)$$

de laquelle nous déduisons en supposant $\Phi_{n|n-1}$ inversible (ce qui est une hypothèse faible sachant que $\Phi_{n|n-1}$ est souvent une perturbation de l'identité)

$$(P_n^+)^{-1} = C_n^\top M_n C_n + \Phi_{n|n-1}^{-\top} P_{n-1}^+ \Phi_{n|n-1}^{-1} = C_n^\top M_n C_n + (P_n^-)^{-1}. \quad (5.26)$$

En combinant tous ces calcul nous pouvons prouver le théorème suivant.

Théorème 13. *l'estimateur de Kalman en temps discret $(\hat{y}_n)_{n \in \mathbb{N}}$ est une suite à deux état dis prédit et corrigé $(\hat{y}_n^-, \hat{y}_n^+)_{n \in \mathbb{N}}$ où \hat{y}_n est donné par l'étape de correction \hat{y}_n^+*

$$\hat{y}_n^+ = \hat{y}_n = \bar{y}_{n|n} = \mathbb{T}_{n|0}(\bar{y}_{0|n}), \quad (5.27)$$

et l'étape de prédition est

$$\hat{y}_{n+1}^- = \bar{y}_{n+1|n} = \mathbb{T}_{n+1|0}(\bar{y}_{0|n}). \quad (5.28)$$

Ces deux étapes se calculent de manière récursive par

$$\left\{ \begin{array}{l} \text{Initialisation :} \\ \hat{y}_0^- = y_\diamond \\ \text{Correction :} \\ \hat{y}_n^+ = \hat{y}_n^- + P_n^+ C_n^\top M_n (z_n^\varepsilon - C_n \hat{y}_n^-), \quad n \in \mathbb{N} \\ \text{Prédiction :} \\ \hat{y}_{n+1}^- = \Phi_{n+1|n} \hat{y}_n^+ + f_n, \quad n \in \mathbb{N} \end{array} \right. \quad (5.29)$$

où les opérateurs dits de covariance P_n^\pm satisfont l'équation de Riccati discrète suivante

$$\begin{cases} \text{Initialisation :} \\ P_0^- = P_\diamond \\ \text{Correction :} \\ P_n^+ = ((P_n^-)^{-1} + C_n^\top M_n C_n)^{-1}, \quad n \in \mathbb{N} \\ \text{Prédiction :} \\ P_{n+1}^- = \Phi_{n+1|n} P_n^+ \Phi_{n+1|n}^\top, \quad n \in \mathbb{N} \end{cases} \quad (5.30)$$

Nous concluons notre présentation formelle par deux compléments. Le premier complément concerne le calcul de gain.

Proposition 7. *Le gain de Kalman G_n vérifie les deux identités*

$$G_n = P_n^+ C_n^\top M_n = P_n^- C_n^\top (W_n + C_n^\top P_n^- C_n)^{-1}$$

et P_n^+ peut être calculé indépendamment de l'inversibilité de P_n^- par

$$P_n^+ = (\mathbb{1} - G_n C_n) P_n^- (\mathbb{1} - G_n C_n)^\top + G_n W_n G_n^\top.$$

Le deuxième fournit une expression de l'estimateur de Kalman par une suite récurrente $(\hat{y}_n)_{n \in \mathbb{N}}$ à une seule étape.

Proposition 8. *L'estimateur de Kalman $(\hat{y}_n)_{n \in \mathbb{N}}$ suit la dynamique*

$$\begin{cases} \hat{y}_{n+1} &= \Phi_{n+1|n} \hat{y}_n + f_n + G_n (z_{n+1}^\varepsilon - C_{n+1}(\Phi_{n+1|n} \hat{y}_n + f_n)) \\ \hat{y}_0 &= y_\diamond + P_\diamond C_0^\top (W_0 + C_0^\top P_\diamond C_0)^{-1} (z_0^\varepsilon - C_0 y_\diamond) \end{cases} \quad (5.31)$$

Reste désormais à établir sous quelle condition le problème est bien posé. Nous voyons par notre choix de présentation que la minimisation est possible s'il existe un n telle que Ψ_n est injectif. Dans le cas où pour tout n , $C_n = C$ et $\Phi_{n+1|n} = \Phi$, nous retrouvons alors le fameux critère d'observabilité de Kalman qui dit que le système est observable si

$$\text{rang} \begin{pmatrix} C \\ C\Phi \\ \dots \\ C\Phi^n \end{pmatrix} = N_{\text{state}}$$

En effet dans ce cas, l'*a priori* n'est pas nécessaire pour obtenir l'injectivité de Ψ_n et on peut oublier le premier bloc $\mathbb{1}$. Inversement si le problème est mal posé, l'injectivité de Ψ_n est apportée par ce premier bloc $\mathbb{1}$.

5.1.4 Prise en compte d'une dynamique continue

Imaginons désormais le problème continu suivant : On considère une classe de modèles dynamiques continus sur $\mathcal{Y} \sim \mathbb{R}^{N_{\text{state}}}$

$$\begin{cases} \dot{y}|_\zeta(t) = A(t)y|_\zeta(t) + f(t), & t \in \mathbb{R}_+ \\ y|_\zeta(0) = y_0 + \zeta, \end{cases} \quad (5.32)$$

où, pour simplifier la présentation, nous supposons que $A \in \mathcal{C}(\mathbb{R}_+; \mathcal{L}(\mathcal{Y}))$ est uniformément borné par rapport à t de telle sorte que le théorème de Cauchy-Lipchitz implique qu'il existe des solutions globales sur \mathbb{R}_+ . Nous supposons par ailleurs que nous disposons de mesures sur une trajectoire

$$\forall t \in \mathbb{R}_+, \quad z^\varepsilon(t) = C(t)\check{y}(t) + \text{erreur}, \quad (5.33)$$

et nous cherchons à estimer \check{y} . L'opérateur d'observation est alors $\mathcal{C}(\mathbb{R}_+; \mathcal{L}(\mathcal{Y}, \mathcal{Z}))$.

Une façon de définir l'estimateur dans ce cadre – et nous en verrons une autre approche aux paragraphes suivants – est d'imaginer ce problème comme la limite du problème en temps discret. On imagine ainsi que (5.19) est la discrétisation du problème (5.32). Par exemple en imaginant un schéma de type Euler explicite sur une grille régulière en temps, on a

$$\Phi_{n+1|n} = \mathbb{1} + \delta t A(t_n),$$

où pour un schéma de type Euler implicite

$$\Phi_{n+1|n} = (\mathbb{1} - \delta t A(t_{n+1}))^{-1}.$$

Concernant les observations, il suffit de considérer par exemple

$$\forall t \in [t_n, t_{n+1}], \quad C(t) = C(t_{n+1}).$$

Il nous faut enfin donner un sens à l'estimateur continu en tant que limite de l'estimateur discret. Pour cela, on rappelle que l'estimateur en temps discret est l'estimateur moindres carrés associé à

$$\min_{\zeta \in \mathcal{Y}} \left\{ J_n(\zeta) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_0}^2 + \frac{1}{2} \sum_{k=0}^n \|z_k^\varepsilon - C_k y_k|_\zeta\|_{\mathcal{Z}, M_k}^2 \right\}.$$

Ce critère est consistant avec un critère continu de la forme

$$\min_{\zeta \in \mathcal{Y}} \left\{ J(\zeta, t) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_0}^2 + \frac{1}{2} \int_0^t \|z^\varepsilon(t) - C(t)y|_\zeta\|_{\mathcal{Z}, M}^2 dt \right\},$$

pour peu qu'on choisisse par exemple

$$M_0 = 0, \quad M_k = \delta t M, k \in [0, n].$$

On en déduit alors par consistance les équations du filtre de Kalman continu.

Définition 10. L'estimateur de Kalman est défini par

$$\forall t \in \mathbb{R}_+, \quad \hat{y}(t) = y_{\bar{\zeta}|t}(t) \text{ où } \bar{\zeta}|_t = \underset{\zeta \in \mathcal{Y}}{\operatorname{argmin}} J(\zeta, t).$$

Théorème 14. Lorsque les dynamiques suivantes admettent des solutions

$$\begin{cases} \dot{\hat{y}}(t) = A(s)\hat{y}(s) + f(s) + P(s)C(s)^\top M(z^\varepsilon(s) - C(s)\hat{y}(s)) \\ y(0) = y_\diamond \end{cases} \quad (5.34)$$

où $P \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$ est solution de l'équation de Riccati

$$\begin{cases} \dot{P}(s) = P(s)A(s)^\top + A(s)P(s) - P(s)C(s)^\top MC(s)P(s) \\ P(0) = P_\diamond \end{cases} \quad (5.35)$$

elles fournissent l'estimateur de Kalman \hat{y} associé au modèle (5.32), aux mesures (5.33) et au critère J .

Preuve.(formel) Nous allons passer à la limite en δt afin de formuler les équations en temps continu. La dynamique de l'estimateur (5.34) se déduit directement par consistance de la dynamique discrète (5.31). Par ailleurs on a

$$P_n^+ = P_n^- - \delta t P_n^- C(t)^\top M C(t) P_n^- + O(\delta t^2)$$

et

$$P_{n+1}^- = P_n^+ - \delta t A(t) P_n^- + \delta t P_n^- A(t)^\top + O(\delta t^2)$$

d'où on déduit par consistance (5.35). ■

Ce résultat est avant tout formel mais plein d'enseignements. Cependant l'existence rigoureuse de l'estimateur (en fait l'existence de la solution de l'équation de Riccati) reste à obtenir et nécessite plus de travail comme nous allons le voir dans les autres sections. Cependant nous en profitons ici pour justifier formellement le critère de Kalman pour les dynamiques continues qui dit que, pour A et C indépendant du temps, le système (5.32) est observable si il existe n_0 tel que

$$\forall n \geq n_0, \quad \operatorname{rang} \begin{pmatrix} C \\ CA \\ \dots \\ CA^n \end{pmatrix} = N_{\text{state}}.$$

En effet ce critère se déduit directement du critère de rang discret avec $\Phi = \mathbb{1} + \delta t A$ ou $\Phi = \mathbb{1} - \sum_k \delta t A^k$ ou tout autre choix consistant.

5.1.5 Prise en compte d'une erreur de modèle

Il se trouve que le cadre (5.32) ou (5.19) n'est pas encore le plus général même à supposer que tous les opérateurs soient linéaires. En effet, nous avons supposé aux paragraphe précédent que la seule incertitude sur la trajectoire provenait de la condition initiale. Cependant, il est aussi possible d'imaginer que le modèle comporte aussi des incertitudes. La plus simple à prendre en compte est une incertitude sur le terme source f . On considère des incertitudes $\nu \in \mathcal{C}(\mathbb{R}_+; \mathcal{Q}) \cup \mathcal{L}^2(\mathbb{R}_+; \mathcal{Q})$ où ici $\mathcal{Q} \simeq \mathbb{R}^{N_{\text{nois}}}$ qui agissent sur la dynamique au travers de l'opérateur $B \in \mathcal{C}(\mathbb{R}_+; \mathcal{L}(\mathcal{Q}, \mathcal{Y}))$ supposé connu de telle sorte que la dynamique devient

$$\begin{cases} \dot{y}_{|\zeta, \nu}(t) = A(t)y_{|\zeta, \nu}(t) + f(t) + B(t)\nu(t) & t \in \mathbb{R}_+ \\ y_{|\zeta, \nu}(0) = y_\diamond + \zeta. \end{cases} \quad (5.36)$$

De même en temps discret on a une suite $(\nu_n)_{n \in \mathbb{N}}$ d'éléments de \mathcal{Q} et une suite $(B_n)_{n \in \mathbb{N}}$ d'opérateurs de $\mathcal{L}(\mathcal{Q}, \mathcal{Y})$ tels que

$$\begin{cases} y_{n+1|\zeta, (\nu_k)_{k \leq n}} = \Phi_{n+1|n}y_{n|\zeta, (\nu_k)_{k \leq n-1}} + f_n + B_n\nu_n, & n \in \mathbb{N} \\ y_{0|\zeta} = y_\diamond + \zeta \end{cases} \quad (5.37)$$

Nous nous proposons d'étendre les formulations d'estimateur à ce cadre et pour ce faire nous allons suivre une démarche différente (bien qu'équivalente) plus directement centré sur la minimisation d'une fonctionnelle moindres carrés.

Commençons par traiter le cas continu en temps. Soit $M \in \mathcal{G}\ell_{N_{\text{nois}}}(\mathbb{R})$ tel que $\|\nu\|_{\mathcal{Z}, M}^2 = \nu^\top M \nu$ est une norme sur \mathcal{Z} . Nous introduisons désormais le critère à minimiser

$$\min_{\zeta, \nu} \left\{ J(\zeta, \nu, t) = \frac{1}{2}\|\zeta\|_{\mathcal{Y}, N_\diamond}^2 + \frac{1}{2} \int_0^t (\|z^\varepsilon(s) - C(s)y_{|\zeta, \nu}(s)\|_{\mathcal{Z}, M}^2 + \|\nu(s)\|_{\mathcal{Q}, S}^2) ds \right\}, \quad (5.38)$$

où cette fois la minimisation concerne à la fois $\zeta \in \mathcal{Y}$ mais aussi $\nu \in \mathcal{L}^2(\mathbb{R}_+; \mathcal{Q})$. Cette fonctionnelle cherche à trouver la meilleure trajectoire minimisant l'écart aux observations z^ε pour des petites incertitudes de condition initiale et de source. Le critère J est convexe et la preuve de l'existence et unicité du minimum et des minimisants est relativement directe (bien que dans un espace de dimensions infinie pour ν). Nous renvoyons à [59] pour plus de détails. La minimisation du critère J s'effectue sous la contrainte de la dynamique (5.36), ainsi il est classique de faire apparaître le multiplicateur de Lagrange associé à cette contrainte. Les équations d'Euler en théorie variationnelle permettent alors de trouver la dynamique de ce multiplicateur puisqu'il est associé à une contrainte dynamique. Pour simplifier, nous donnons ici cette dynamique, appelée dynamique de l'état adjoint

$$\begin{cases} \dot{q}_{|\zeta, \nu, t}(s) + A(s)^\top q_{|\zeta, \nu, t}(s) = -C(s)^\top M(z^\varepsilon(s) - C(s)y_{|\zeta, \nu, t}(s)), & s \in [0, t] \\ q_{|\zeta, \nu, t}(t) = 0, \end{cases} \quad (5.39)$$

où la condition initiale est en fait une condition finale ! Nous allons voir dans la preuve du théorème qui suit comment cette dynamique permet de calculer les conditions de stationnarité du critère.

Définition 11. *L'estimateur moindres carrés, aussi appelé estimateur 4D-Var, est*

$$\bar{y}|_t = y|_{\bar{\zeta}|_t, \bar{\nu}|_t} \text{ où } (\bar{\zeta}|_t, \bar{\nu}|_t) = \underset{\zeta \in \mathcal{Y}, \nu \in L^2([0,t]; \mathcal{Q})}{\operatorname{argmin}} J(\zeta, \nu, t)$$

Théorème 15. *La dynamique du système (5.36) associé aux minimisants de $J(\cdot, \cdot, t)$ est donnée par*

$$\begin{cases} \dot{\bar{y}}|_t(s) = A(s)\bar{y}|_t + B(s)QB(s)^\top \bar{q}|_t(s), & s \in [0, t] \\ \dot{\bar{q}}|_t(s) + A(s)^\top \bar{q}|_t = -C(s)^\top M(z^\varepsilon(s) - C(s)\bar{y}|_t(s)), & s \in [0, t] \\ \bar{y}|_t(0) = y_\diamond + P_\diamond \bar{q}|_t(0), \\ \bar{q}|_t(t) = 0. \end{cases} \quad (5.40)$$

Preuve. Commençons par différencier le critère par rapport à ζ ,

$$d_\zeta J(\delta\zeta) = \zeta^\top N_\diamond \delta\zeta - \int_0^t (z^\varepsilon(s) - C(s)y|_{\zeta, \nu}(s))^\top MC(s) d_\zeta y|_{\zeta, \nu}(s)(\delta\zeta) ds,$$

où $d_\zeta y|_{\zeta, \nu}$ est la sensibilité de la trajectoire par rapport à l'incertitude initiale ζ , i.e.

$$\begin{cases} \overline{d_\zeta y|_{\zeta, \nu}}(s) = A(s)y|_{\zeta, \nu}(s) d_\zeta y|_{\zeta, \nu}(s), & s \in [0, t] \\ d_\zeta y|_{\zeta, \nu}(0) = \mathbb{1}. \end{cases}$$

Nous obtenons en utilisant l'état adjoint

$$\begin{aligned} d_\zeta J(\delta\zeta) &= \zeta^\top N_\diamond \delta\zeta + \int_0^t \left(\dot{q}|_{\zeta, \nu, t}(s) + d_y A(y(s), s)^\top q|_{\zeta, \nu, t}(s) \right)^\top d_\zeta y(s)(\delta\zeta) ds \\ &= \zeta^\top N_\diamond \delta\zeta + \left[q|_{\zeta, \nu, t}(s)^\top d_\zeta y|_{\zeta, \nu}(s)(\delta\zeta) \right]_0^t - \int_0^t q|_{\zeta, \nu, t}(s)^\top \overline{d_\zeta y|_{\zeta, \nu}}(s)(\delta\zeta) ds \\ &\quad + \int_0^t q|_{\zeta, \nu, t}(s)^\top A(s) d_\zeta y|_{\zeta, \nu}(s)(\delta\zeta) ds \\ &= \zeta^\top N_\diamond(\delta\zeta) - q|_t(0)^\top(\delta\zeta). \end{aligned}$$

Ainsi

$$d_\zeta J(\bar{\zeta}|_t, \bar{\nu}|_t, t) = 0 \Rightarrow \bar{\zeta}|_t = P_\diamond \bar{q}|_t(0).$$

De façon similaire

$$\begin{aligned} d_\nu J(\delta\nu) &= \int_0^t \nu(s)^\top S(\delta\nu(s)) \, ds + \left[q_{|\zeta,\nu,t}(s)^\top d_\nu y_{|\zeta,\nu}(s)(\delta\nu(s)) \right]_0^t \\ &\quad - \int_0^t q_{|\zeta,\nu,t}(s) \overline{\dot{d}_\nu y_{|\zeta,\nu}}(s)(\delta\nu(s)) \, ds \\ &\quad + \int_0^t q_{|\zeta,\nu,t}(s)^\top A(s) y_{|\zeta,\nu}(s) d_\nu y_{|\zeta,\nu}(s)(\delta\nu(s)) \, ds. \end{aligned}$$

Cette fois nous utilisons la sensibilité de la trajectoire par rapport à l'incertitude ν qui vérifie

$$\begin{cases} \overline{\dot{d}_\nu y_{|\zeta,\nu}}(s) = A(s) y_{|\zeta,\nu}(s) d_\nu y_{|\zeta,\nu}(s) + B(s), & s \in [0, t] \\ d_\nu y_{|\zeta,\nu}(0) = 0. \end{cases}$$

Nous obtenons alors

$$d_\nu J(\delta\nu) = \int_0^t (\nu(s)^\top S - q_{|\zeta,\nu,t}(s)^\top B(s)) \delta\nu(s) \, ds,$$

ce qui donne

$$d_\nu J(\bar{\zeta}_{|t}, \bar{\nu}_{|t}, t) = 0 \Rightarrow \forall s \in [0, t], \quad \bar{\nu}_{|t}(s) = Q B(s)^\top \bar{q}_{|\zeta,\nu,t}(s).$$

La conclusion découle directement de ces conditions de stationnarité. ■

La dynamique (5.40) est cependant difficile à résoudre car nous faisons face à un problème dit aux deux bouts et non pas un problème de Cauchy. En pratique, il faut donc se résoudre à adopter une approche itérative en définissant typiquement un algorithme de type descente de gradient sur $J(\cdot, \cdot, t)$. A chaque itération, on résout alors une dynamique (5.36) dite *forward* puis une dynamique dite *backward* pour l'adjoint (5.39). La limite de cette méthode de descente est alors le système (5.40).

Exercice Ecrire une méthode de descente de gradient à pas optimal pour la résolution du 4D-Var.

Il est à noter que cette approche par minimisation permet aussi de définir l'estimateur moindres carrés en l'absence d'erreur de modèle, *i.e.* il suffit de faire $B = 0$ dans tout ce qui précède. Cette approche est plus directe que ce que nous avons fait aux paragraphes précédents. La question reste cependant de trouver l'estimateur de type *Kalman*. Tout d'abord la définition 10 s'étend comme suit.

Définition 12. *L'estimateur de Kalman est défini à partir de l'estimatateur 4D-Var par*

$$\forall t \in \mathbb{R}_+, \quad \hat{y}(t) = \bar{y}_{|t}(t).$$

Théorème 16. *Supposons que la dynamique suivante est bien définie*

$$\begin{cases} \dot{\hat{y}}(t) = A(t)\hat{y}(t) + f(t) + P(t)C(t)^\top M(z^\varepsilon(t) - C(t)\hat{y}(t)) \\ y(0) = y_\diamond \end{cases} \quad (5.41)$$

et $P \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$ est solution de l'équation de Riccati

$$\begin{cases} \dot{P}(t) = P(t)A(t)^\top + A(t)P(t) - P(t)C(t)^\top MC(t)P(t) + B(t)QB(s)^\top \\ P(0) = P_\diamond \end{cases} \quad (5.42)$$

alors \hat{y} est l'estimateur de Kalman associé au modèle (5.36), aux mesures (5.33) et au critère J . Par ailleurs, nous avons

$$\forall s \in [0, t], \quad \bar{y}|_t(s) = \hat{y}(s) + P(s)\bar{q}|_t(s). \quad (5.43)$$

Preuve. Nous supposons que les solutions de (5.34) et (5.35) sont bien globales sur \mathbb{R} . Nous allons prouver (5.43) ce qui justifie directement l'estimateur de Kalman en $s = t$. On a en effet

$$\begin{aligned} \dot{\hat{y}}(s) &= \dot{\bar{y}}|_t(s) - \dot{P}(s)\bar{q}|_t(s) - \dot{P}(s)\dot{\bar{q}}|_t(s) \\ &= A(s)\bar{y}|_t(s) + f(s) - \dot{P}(s)\bar{q}|_t(s) \\ &\quad - P(s)\left(-A(s)^\top\bar{q}|_t + -C(s)^\top M(z^\varepsilon(s) - C(s)\bar{y}|_t(s))\right) \\ &= A(s)(\hat{y}(s) + P(s)\bar{q}|_t(s)) + f(s) - \dot{P}(s)\bar{q}|_t(s) \\ &\quad - P(s)\left(-A(s)^\top\bar{q}|_t + -C(s)^\top M(z^\varepsilon(s) - C(s)(\hat{y}(s) + P(s)\bar{q}|_t(s)))\right) \\ &= A(s)\hat{y}(s) + f(s) + P(s)C(s)^\top M(z^\varepsilon(s) - C(s)\hat{y}(s)) \\ &\quad \left(\dot{P}(s) - P(s)A(s)^\top - A(s)P(s) + P(s)C(s)^\top MC(s)P(s)\right)\bar{q}|_t(s) \end{aligned}$$

qui donne bien

$$\dot{\hat{y}}(t) = A(s)\hat{y}(s) + f(s) + P(s)C(s)^\top M(z^\varepsilon(s) - C(s)\hat{y}(s)).$$

■

Reste à savoir quand l'estimateur de Kalman peut être défini et notamment quand l'équation de Riccati (5.35) admet des solutions. Cette question est difficile puisque (5.35) est une dynamique non-linéaire sur un opérateur $P \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$...

Passons désormais au discret en temps. Ce cadre nous fournira à la fois une formule intrinsèque de l'estimateur de Kalman en temps discret mais surtout une façon de discréteriser (5.34) et (5.35). Le critère discret s'écrit désormais

$$\min_{\zeta, (\nu_k)_{k < n}} \left\{ J_n^+(\zeta, (\nu_k)_{k < n}) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_\diamond}^2 + \frac{1}{2} \sum_{k=0}^n \|z_k^\varepsilon - C_k y_{k|\zeta}\|_{\mathcal{Z}, M_k}^2 + \frac{1}{2} \sum_{k=0}^{n-1} \|\nu_k\|_{\mathcal{Q}, S_k}^2 \right\}, \quad (5.44)$$

avec en tout généralité

$$J_0^+(\zeta) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_\diamond}^2 + \|z_0^\varepsilon - C_0 y_{0|\zeta}\|_{\mathcal{Z}, M_0}^2.$$

Il faut noter que si on veut pouvoir passer à la limite il faut que la suite $(\nu_n)_{n \in \mathbb{N}} \in \ell_2(\mathbb{N}; \mathcal{Q})$. L'état adjoint, correspondant au multiplicateur de Lagrange associé à la contrainte de la dynamique discrète, est cette fois

$$\begin{cases} -q_{k|\zeta, (\nu_k)_{k < n}, n}^+ + \Phi_{k+1|k}^\top q_{k+1|\zeta, (\nu_k)_{k < n}, n}^+ \\ \quad = -C_k^\top M_k (z_k^\varepsilon - C_k y_{k|\zeta, (\nu_j)_{j < k}}), & k \in [0, n] \\ q_{n+1|\zeta, (\nu_k)_{k < n}, n}^+ = 0. \end{cases} \quad (5.45)$$

On définit alors l'estimateur moindres carrés en temps discret à partir de ce critère.

Définition 13. *L'estimateur moindres carrés en temps discret où estimateur 4D – Var en temps discret est défini par*

$$\bar{y}_{|n} = y_{|\bar{\zeta}_{|n}, (\bar{\nu}_k)_{k < n}|n} \text{ où } (\bar{\zeta}_{|n}, (\bar{\nu}_k)_{k < n}|n) = \underset{\zeta \in \mathcal{Y}, (\nu_k)_{k < n} \in \mathcal{Q}^n}{\operatorname{argmin}} J(\zeta, \nu, n)$$

On a alors la caractérisation suivante.

Théorème 17. *L'estimateur 4D – Var en temps discret est donné par*

$$\begin{cases} \bar{y}_{k+1|n}^+ = \Phi_{k+1|k} \bar{y}_{k|n}^+ + f_k + B_k Q_k B_k^\top \bar{q}_{k+1|n}^+, & k \in [0, n] \\ -\bar{q}_{k|n}^+ + \Phi_{k+1|k}^\top \bar{q}_{k+1|n}^+ = -C_k^\top M_k (z_k^\varepsilon - C_k \bar{y}_{k|n}^+), & k \in [0, n] \\ \bar{y}_{0|n}^+ = y_\diamond + P_\diamond \bar{q}_{0|n}^+ \\ \bar{q}_{n+1|n}^+ = 0. \end{cases} \quad (5.46)$$

Preuve. La preuve est l'exacte symétrique en temps discret de celle du Théorème 15. La sensibilité par rapport à la condition initiale s'écrit désormais

$$\begin{cases} d_\zeta y_{k+1} = \Phi_{k+1|k} d_\zeta y_j, & k \in [0, n] \\ d_\zeta y_k = \mathbb{1}. \end{cases}$$

On calcule alors la différentielle

$$\begin{aligned}
d_\zeta J_n^+(\delta\zeta) &= \zeta^\top N_\diamond \delta\zeta - \sum_{k=0}^n (z_k^\varepsilon - C_k y_k)^\top M_k C_k d_\zeta y_k(\delta\zeta) \\
&= \zeta^\top N_\diamond \delta\zeta + \sum_{k=0}^n (-q_{k|n}^+ + \Phi_{k+1|k}^\top q_{k+1|n}^+)^\top d_\zeta y_k(\delta\zeta) \\
&= \zeta^\top N_\diamond \delta\zeta - \sum_{k=0}^n q_{k|n}^+ d_\zeta y_k \cdot \delta\zeta + \sum_{k=0}^n (q_{k+1|n}^+)^\top d_\zeta y_{k+1}(\delta\zeta) \\
&= \zeta^\top N_\diamond \cdot \delta\zeta - (q_{0|n}^+)^\top \delta\zeta.
\end{aligned}$$

Et en utilisant la sensibilité du modèle par rapport à chaque incertitude ν_k ,

$$\begin{cases} d_{\nu_k} y_{j+1} = \Phi_{j+1|j} d_{\nu_k} y_j + \delta_{k,i} B_i, & j \in [k, n] \\ d_{\nu_k} y_k = 0 \end{cases}$$

où $\delta_{k,i}$ est le symbole de Kronecker, on a pour tout $0 \leq k < n$,

$$\begin{aligned}
d_{\nu_k} J_n^+(\delta\nu_k) &= \nu_k^\top S_k \delta\nu_k - \sum_{j=k}^n (z_k^\varepsilon - C_k y_k)^\top M_k C_k d_{\nu_k} y_k(\delta\nu_k) \\
&= \nu_k^\top S_k \delta\nu_k + \sum_{j=k}^n \left(-q_{j|n}^+ + \Phi_{j+1|j}(y_j)^\top q_{j+1|n}^+ \right)^\top d_{\nu_k} y_j(\delta\nu_k) \\
&= \nu_k^\top S_k \delta\nu_k - \sum_{j=k+1}^n q_j d_{\nu_k} y_j(\delta\nu_k) + \sum_{j=k+1}^n (q_{j|n}^+)^\top d_{\nu_k} y_{j+1}(\delta\nu_k) \\
&= \nu_k^\top S_k \delta\nu_k - (q_{k+1|n}^+)^\top B_k(\delta\nu_k).
\end{aligned}$$

On en déduit alors comme précédemment les condition de stationnarité du critère. ■

En temps discret, la définition du critère aurait pu être

$$\min_{\zeta, (\nu_k)_{k \leq n}} \left\{ J_{n+1}^-(\zeta, (\nu_k)_{k \leq n}) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_\diamond}^2 + \frac{1}{2} \sum_{k=0}^n \|z_k^\varepsilon - C_k y_{k|\zeta}\|_{\mathcal{Z}, M_k}^2 + \frac{1}{2} \sum_{k=0}^n \|\nu_k\|_{\mathcal{Q}, S_k}^2 \right\}, \quad (5.47)$$

avec

$$J_0^-(\zeta) = \frac{1}{2} \|\zeta\|_{\mathcal{Y}, N_\diamond}^2.$$

Dans ce cas l'état adjoint correspondant est

$$\begin{cases} -q_{k|\zeta, (\nu_k)_{k \leq n}, n+1}^- + \Phi_{k+1|k}^\top q_{k+1|\zeta, (\nu_k)_{k \leq n}, n+1}^- \\ \quad = -C_k^\top M_k (z_k^\varepsilon - C_k y_{k|\zeta, (\nu_j)_{j < k}}), & k \in [0, n] \\ q_{n+1|\zeta, (\nu_k)_{k \leq n}, n+1}^- = 0 \end{cases} \quad (5.48)$$

et il est très facile d'obtenir un équivalent du Théorème 17.

On peut alors formuler l'estimateur de Kalman discret.

Définition 14. *L'estimateur de Kalman en temps discret $(\hat{y}_n)_{n \in \mathbb{N}}$ est une suite à deux état dis prédit et corrigé $(\hat{y}_n^-, \hat{y}_n^+)_{n \in \mathbb{N}}$ tels que*

$$\hat{y}_n^+ = y_{n|n}^+ \text{ et } \hat{y}_n^- = y_{n|n}^-.$$

On a alors la caractérisation suivante.

Théorème 18. *Supposons que P_n^\pm sont des solutions inversibles de l'équation de Riccati en temps discret*

$$\begin{cases} \text{Initialization :} \\ P_0^- = P_\diamond \\ \text{Correction :} \\ P_n^+ = ((P_n^-)^{-1} + C_n^\top M_n C_n)^{-1}, \quad n \in \mathbb{N} \\ \text{Prediction :} \\ P_{n+1}^- = \Phi_{n+1|n} P_n^+ \Phi_{n+1|n}^\top + B_n Q_n B_n^\top, \quad n \in \mathbb{N} \end{cases} \quad (5.49)$$

Alors l'estimateur de Kalman discret \hat{y}_n^\pm suit la dynamique (5.29). Par ailleurs on a

$$\bar{y}_{k,n+1}^- = \hat{y}_k^- + P_k^- \bar{q}_{k|n+1}^-, \quad (5.50)$$

Preuve. La dynamique de l'estimateur est toujours donné par (5.29), et donc toujours par (5.31) Nous prouvons alors (5.50) par récurrence

$$\bar{y}_{0|n+1}^- = y_\diamond + P_\diamond \bar{q}_{0|n+1}^- = \hat{y}_0^- + P_0^- \bar{q}_{0|n+1}^-.$$

Puis en supposant (5.50) satisfait pour k nous avons

$$\begin{aligned} \hat{y}_{k+1}^- &= \bar{y}_{k+1|n+1}^- - (\Phi_{k+1|k} P_k^+ \Phi_{k+1|k}^\top + B_k Q_k B_k^\top) \bar{q}_{k+1|n+1}^- \\ &= \bar{y}_{k+1|n+1}^- - \Phi_{k+1|k} P_k^+ (\bar{q}_{k|n+1}^- - C_k M(z_k^\varepsilon - C_k \bar{y}_{k|n+1}^-)) \\ &\quad - B_k Q_k B_k^\top \bar{q}_{k+1|n+1}^- \\ &= \Phi_{k+1|k} (\hat{y}_k^- + P_k^- \bar{q}_{k|n+1}^-) + f_k - \Phi_{k+1|k} P_k^+ \bar{q}_{k|n+1}^- \\ &\quad - \Phi_{k+1|k} P_k^+ C_k M_k (z_k^\varepsilon - C_k (\hat{y}_k^- + P_k^- \bar{q}_{k|n+1}^-)) \\ &= \Phi_{k+1|k} \hat{y}_k^- + f_k + \Phi_{k+1|k} P_k^+ C_k M_k (z_k^\varepsilon - C_k \hat{y}_k^-) \\ &\quad + (\Phi_{k+1|k} P_k^- - \Phi_{k+1|k} P_k^+ - \Phi_{k+1|k} P_k^+ C_k M_k C_k P_k^-) \bar{q}_{k|n+1}^- \\ &= \Phi_{k+1|k} \hat{y}_k^- + f_k + \Phi_{k+1|k} P_k^+ (z_k^\varepsilon - C_k M_k \hat{y}_k^-) \\ &\quad + \Phi_{k+1|k} P_k^+ ((P_k^-)^{-1} - (P_k^+)^{-1} - C_k M_k C_k) P_k^- \bar{q}_{k|n+1}^- \\ &= \Phi_{k+1|k} \hat{y}_k^- + f_k + \Phi_{k+1|k} P_k^+ (z_k^\varepsilon - C_k M_k \hat{y}_k^-) \end{aligned}$$

La dernière expression correspond exactement à la dynamique de \hat{y}_k^- . On a alors

$$\bar{y}_{n|n+1}^- = \hat{y}_n^+ \quad \text{et} \quad \bar{y}_{n+1|n+1}^- = \hat{y}_{n+1}^-, \quad (5.51)$$

consistant avec la définition 14. ■

5.1.6 Reconstruction de la condition initiale

Nous avons perçu qu'un des inconvénients de l'approche par filtrage est qu'elle ne s'intéresse pas à la reconstruction de la condition initiale. Cependant la condition initiale peut être envisagée à partir de l'équation adjointe. On rappelle que l'adjoint

$$\dot{\bar{q}}_t(s) + A(s)^\top \bar{q}_t = -C(s)^\top M(z^\varepsilon(s) - C(s)\bar{y}_t(s)), \quad s \in [0, t],$$

permet de calculer

$$\bar{y}_t(0) = y_\diamond + P_\diamond \bar{q}_t(0).$$

Cette dynamique est hélas dépendante de $\bar{y}_t(s)$ que l'estimateur de Kalman ne calcule pas. Cependant en utilisant la relation (5.43) du Théorème 16, nous obtenons pour tout $s \in [0, t]$

$$\begin{aligned} \dot{\bar{q}}_t(s) + A(s)^\top \bar{q}_t &= -C(s)^\top M \left(z^\varepsilon(s) - C(s)(\hat{y}(s) - P(s)\bar{q}_t(s)) \right) \\ &= -C(s)^\top M \left(z^\varepsilon(s) + C(s)P(s)\bar{q}_t(s) \right) \end{aligned} \quad (5.52)$$

ce qui cette fois est calculable à partir de l'estimateur de Kalman et stable en temps rétrograde (Si $A - PC^*C$ est stable en temps $-A^* + C^*CP$ l'est en temps rétrograde). En fait (5.52) est calculable en théorie car (5.52) oblige à stocker $P(s)$ pour tout $s \in [0, t]$. Une alternative consiste à introduire un système augmenté de la forme

$$\frac{d}{dt} \begin{pmatrix} y \\ \zeta \end{pmatrix} = \begin{pmatrix} A(t) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix}$$

de condition initiale

$$\begin{pmatrix} y \\ \zeta \end{pmatrix}(0) = \begin{pmatrix} y_\diamond + \zeta \\ \zeta \end{pmatrix}(0)$$

et à appliquer l'estimation de Kalman à partir de l'opérateur de covariance initiale

$${}^b P_0 = \begin{pmatrix} P_0 & P_0 \\ P_0 & P_0 \end{pmatrix}$$

Exercice : Des calculs directs permettent de montrer que l'estimateur de Kalman (5.41) ainsi que l'estimation de la condition initiale se déduisent directement de l'estimation de Kalman du système augmenté.

5.1.7 Analyse de stabilité

On pose $\tilde{y} = \check{y} - y$ et on considère dans un premier temps que les données ne sont pas bruitées.

Exercice : Démontrer que $V(y, t) = \langle \tilde{y}, P^{-1}\tilde{y} \rangle_{\mathcal{Y}}$ est une fonctionnelle de Lyapunov.



5.2 Vers les systèmes de dimension infinie

5.2.1 Définition du modèle

Nous pouvons désormais nous interroger sur les extensions des théories précédentes au cadre de la dimension infinie. Pour le cas statique, ce fut l'objet des parties précédentes, nous nous concentrerons donc ici sur le cas dynamique. La première chose à faire est de donner un sens à un système dynamique du type (5.32) ou (5.36) sur un espace d'état \mathcal{Y} désormais considéré comme étant un espace de Hilbert de dimension infinie. Nous allons traiter ici un cas simplifié où A est indépendant du temps. Dans ce cas, si A est borné, l'évolution d'un système de la forme

$$\begin{cases} \dot{y}(t) = Ay(t), & t \in \mathbb{R}_+ \\ y(0) = y_0 \end{cases} \quad (5.53)$$

s'obtient par l'exponentiel d'opérateur défini par

$$e^{tA} = \sum_k \frac{1}{k!} A^k$$

qui est bien défini par le théorème de convergence normale puisqu'en choisissant une norme d'opérateur on a

$$\sum_k \frac{1}{k!} \|A\|^k = e^{t\|A\|} < \infty$$

La question est alors d'étendre cette notion pour A non borné comme c'est typiquement le cas pour les dynamiques associées à des problèmes aux dérivées partielles. Cette extension se fait au travers de la notion de semi-groupe.

Définition 15. Soit \mathcal{Y} un espace de Banach ; on dit que la famille d'opérateurs linéaires $(\Phi(t))_{t \geq 0}$ est un semi-groupe (fortement continu) si :

1. $\forall t \geq 0, \Phi(t) \in \mathcal{L}(\mathcal{Y})$
2. $\Phi(0) = \text{Id}_{\mathcal{L}(\mathcal{Y})}$
3. $\forall (s, t) \geq 0, \Phi(s+t) = \Phi(s) \circ \Phi(t)$
4. $\forall y \in \mathcal{Y}, \lim_{t \rightarrow 0^+} \Phi(t)y = y$

La condition 4 est équivalente à ce que $\forall y \in \mathcal{Y}, t \mapsto \Phi(t)y \in \mathcal{C}^0(\mathbb{R}_+, \mathcal{Y})$.

On dit que le semi-groupe $(\Phi(t))_{t \geq 0}$ est de contraction si $\forall t \geq 0, \|\Phi(t)\|_{\mathcal{L}(X)} \leq 1$. C'est par ailleurs un groupe si $\forall t \geq 0, \|\Phi(t)\|_{\mathcal{L}(X)} = 1$. Relions maintenant le semi-groupe (qui permet de définir un flot) à son générateur (sa dynamique).

Définition 16. On définit le générateur infinitésimal $(A, D(A))$ d'un semi-groupe fortement continu $(\Phi(t))_{t \geq 0}$ comme l'opérateur non borné $A : D(A) \subset \mathcal{Y} \rightarrow \mathcal{Y}$ où

$$D(A) = \left\{ y \in \mathcal{Y}, \lim_{t \rightarrow 0} \frac{\Phi(t)y - y}{t} \text{ existe} \right\} \text{ et } \forall y \in D(A), Ay = \lim_{t \rightarrow 0} \frac{\Phi(t)y - y}{t}$$

On a alors le résultat suivant.

Théorème 19. Soit $A : D(A) \subset \mathcal{Y} \rightarrow \mathcal{Y}$ un opérateur non borné sur \mathcal{Y} . On a l'équivalence :

- $(A, D(A))$ est le générateur infinitésimal d'un semi-groupe de contraction.
- $D(A)$ est dense et pour toute condition initiale $y_0 \in D(A)$, il existe une unique solution $t \mapsto y(t) \in \mathcal{C}^1(\mathbb{R}_+, X)$ de (5.53).

De plus, sous cette hypothèse, la solution $y(t)$ est à valeurs dans $D(A)$ et vérifie $\|y(t)\|_{\mathcal{Y}} \leq \|y_0\|_{\mathcal{Y}}$ ainsi que $\|y'(t)\|_{\mathcal{Y}} \leq \|Ay(t)\|_{\mathcal{Y}} \leq \|Ay_0\|_{\mathcal{Y}}$ (inégalités d'énergie).

Une fois la notion de semi-groupe bien défini, on peut donc considérer les solutions de (5.32) par la formule dite de Duhamel

$$y|_{\zeta}(t) = \Phi(t)\zeta + \int_0^t \Phi(t-s)f(s)ds,$$

qui étend la célèbre formule de la variation de la constante quand $\Phi(t) = e^{tA}$.

Il nous faut maintenant donner des conditions simples pour que $(A, D(A))$ soit le générateur infinitésimal d'un semi-groupe de contraction. Pour ce faire, on introduit la notion d'opérateur dissipatif.

Définition 17. Un opérateur $(A, D(A))$ est dissipatif si

$$\forall y \in D(A), \quad \forall \lambda > 0, \quad \|y - \lambda Ay\| \geq \|y\|.$$

On montre par ailleurs que

Proposition 9. Soit \mathcal{Y} un espace de Hilbert, $(A, D(A))$ est dissipatif si et seulement si

$$\forall y \in D(A), \quad \Re(\langle Ay, y \rangle_{\mathcal{Y}}) \leq 0.$$

On remarque que si $(A, D(A))$ est dissipatif alors $\forall \lambda > 0$, $(\mathbb{1} - \lambda A)$ est injectif. Si de plus $(\mathbb{1} - \lambda A)$ est surjectif alors on dit que $(A, D(A))$ est maximal-dissipatif (ou m-dissipatif). On peut montrer par ailleurs que $(\mathbb{1} - \lambda A)$ est surjectif si et seulement s'il existe un λ_0 tel que $(\mathbb{1} - \lambda A)$ est surjectif. Enfin dans le cas où $(\mathbb{1} - \lambda A)$ est un isomorphisme, on appelle $\Lambda_\lambda = (\mathbb{1} - \lambda A)^{-1}$ la résolvante de A .

Une des caractérisations les plus agréables des opérateurs m-dissipatif est alors la suivante.

Proposition 10. *Soit \mathcal{Y} un espace de Hilbert, on a*

- si $(A, D(A))$ est dissipatif autoadjoint à domaine dense alors il est m-dissipatif,
- si $(A, D(A))$ est anti-adjoint à domaine dense alors il est m-dissipatif.

Nous pouvons alors relier les opérateurs m-dissipatifs au générateurs infinitésimaux par le célèbre théorème de Hille Yoshida.

Théorème 20 (Hille-Yosida). *Soit \mathcal{Y} un espace de Banach et $A : D(A) \subset \mathcal{Y} \rightarrow \mathcal{Y}$ un opérateur non borné. On a l'équivalence :*

- $(A, D(A))$ est m-dissipatif à domaine dense,
- $(A, D(A))$ est le générateur infinitésimal d'un semi-groupe de contraction.

5.2.2 L'approche variationnelle

Nous voyons que dans le cadre de l'approche variationnelle, la minimisation de la fonctionnelle s'écrit toujours sous la forme

$$\min_{\zeta, \nu} \left\{ J(\zeta, \nu, t) = \frac{\alpha}{2} \|\zeta\|_{\mathcal{Y}}^2 + \frac{1}{2} \int_0^t \beta (\|z^\varepsilon(s) - C(s)y_{|\zeta, \nu}(s)\|_{\mathcal{Z}}^2 + \gamma \|\nu(s)\|_{\mathcal{Q}}^2) \, ds \right\}, \quad (5.54)$$

où la pondération dans les différents espaces s'effectue par trois scalaires $\alpha, \beta, \gamma \in \mathbb{R}_+$. L'origine de l'appellation 4D-Var provient d'ailleurs de ce cadre d'application aux systèmes physiques modélisés par des équation aux dérivées partielles d'évolution [39]. On souhaite toujours utilisé l'état adjoint et la théorie des semi-groupes nous offre un cadre tout à fait adapté. Il nous faut tout de même préciser les régularité des opérateurs d'observation C et d'erreur de modèle B . Le plus simple est de considérer $C \in L^\infty(\mathbb{R}_+, \mathcal{L}(\mathcal{Y}, \mathcal{Z}))$ et $B \in L^\infty(\mathbb{R}_+, \mathcal{L}(\mathcal{Q}, \mathcal{Y}))$.

En particulier, l'existence d'un minimum atteint est toujours possible [59], et la définition de la dynamique adjointe (5.39) est possible. En effet si $(A, D(A))$ est générateur d'un semi-groupe alors $(A^*, D(A^*))$ l'est. Donc $-A^*$ permet de définir une dynamique en temps rétrograde.

Concernant la discrétisation de cette approche, nous percevons ici tout l'intérêt d'avoir d'abord traité le cas de la dimension finie. En effet, une stratégie de discrétisation adéquate consiste à discréteriser le problème de dimension infinie initial puis d'y appliquer la théorie variationnelle en dimension finie.

5.2.3 L'approche par filtrage optimal

Pour définir le filtre de Kalman, il faut d'abord être en mesure de définir rigoureusement l'opérateur de covariance. Pour ce faire (et ceci est d'ailleurs valable en dimension finie) on le définit intrinsèquement comme l'opérateur borné de \mathcal{Y} associé à $q \mapsto \bar{y}_{t,\lambda}(t)$

$$\bar{y}_{t,\lambda}(t_1) = P(t_1)\lambda, \quad \forall \lambda \in \mathcal{Y},$$

où

$$\begin{cases} \dot{\bar{y}}_{s,\lambda}(s) = A(s)\bar{y}_{s,\lambda}(s) + B(s)QB(s)^*\bar{q}_{t,\lambda}(s), & \text{in } (0, t), \\ -\dot{\bar{q}}_{t,\lambda}(s) = A(s)^*\bar{q}_{t,\lambda}(s) - C(s)^*MC(s)\bar{y}_{t,\lambda}(s), & \text{in } (0, t), \\ \bar{y}_{t,\lambda}(0) = P_\diamond \bar{q}_{t,\lambda}(0), \\ \bar{q}_{t,\lambda}(s) = \lambda. \end{cases}$$

L'équation de Riccati

$$\begin{cases} \dot{P}(s) = P(s)A(s)^* + A(s)P(s) - \beta P(s)C(s)^*C(s)P(s) + \gamma B(s)B(s)^* \\ P(0) = \alpha \mathbb{1}, \end{cases} \quad (5.55)$$

admet une solution en différents sens sur $[0, T]$

- *mild sense*, i.e. pour tout $\lambda \in \mathcal{Y}$, l'application $t \in [0, T] \rightarrow P(t)\lambda$ est continue et

$$\begin{aligned} P(t)\lambda &= \alpha \Phi(t)\Phi(t)^*\lambda + \gamma \int_0^t \Phi(t-s)B(s)B(s)^*\Phi(t-s)^*\lambda \, ds \\ &\quad - \beta \int_0^t \Phi(t-s)P(s)C(s)^*C(s)P(s)\Phi(t-s)^*\lambda \, ds, \end{aligned} \quad (5.56)$$

- *weak sense*, i.e. pour tout $p, q \in D(A^*)$, l'application $t \in (0, T) \rightarrow \langle A^*p, P(t)q \rangle_{\mathcal{Y}}$ est différentiable et

$$\begin{aligned} \frac{d}{dt} \langle p, P(t)q \rangle_{\mathcal{Y}} &= \langle A^*p, P(t)q \rangle_{\mathcal{Y}} + \langle P(t)p, A^*q \rangle_{\mathcal{Y}} \\ &\quad - \beta \langle CP(t)p, CP(t)q \rangle_{\mathcal{Z}} + \gamma \langle B^*p, B^*q \rangle_{\mathcal{Q}}. \end{aligned} \quad (5.57)$$

- *strong sense*

$$P \in \mathcal{C}^1(\mathbb{R}, S^+(\mathcal{Y}))$$

où $S^+(\mathcal{Y})$ est l'espace des opérateurs bornées symétriques

Une fois l'opérateur de Riccati défini, il est alors possible de donner un sens (au moins au sens *mild solution*) à l'estimateur au sens de Kalman de dynamique

$$\begin{cases} \dot{\hat{y}}(t) = A\hat{y}(t) + f(t) + \beta P(t)C^*(z^\varepsilon(t) - C\hat{y}(t)), & t \in \mathbb{R}_+ \\ \hat{y}(0) = y_\diamond \end{cases} \quad (5.58)$$

dans le sens où

$$\forall t \in \mathbb{R}_+, \quad \hat{y}(t) = \Phi(t)y_0 + \int_0^t \Phi(t-s) \left(f(s) + \beta P(s)C^*(z^\varepsilon(s) - C\hat{y}(s)) \right) ds.$$

Reste alors à proposer une discrétisation adaptée. Là encore, l'idée est de d'utiliser la théorie en dimension finie et de l'appliquer à la discrétisation du problème de dimension infinie initial. Nous percevons alors un problème de fond pour l'approche par filtrage de Kalman. L'opérateur P après discrétisation est une matrice pleine et non creuse ce qui rend cette approche très vite incalculable pour des systèmes dynamiques issus de la discrétisation d'EDP. En particulier les méthodes par différences-finies ou éléments-finis en dimension supérieurs à 1 ne pourront pas être estimés. On appelle cette limitation la « malédiction de la dimensionnalité » en hommage à R. Bellman [14]. Pour plus de détails sur toutes ces notions présentées succinctement nous renvoyons notamment à [16, 15, 8].

5.2.4 Exemples d'application

Un problème jouet : Equation de transport avec conditions périodiques – Soit l'équation de transport avec des conditions périodiques :

$$\begin{cases} \frac{\partial}{\partial t}y_{|\zeta}(x,t) - \frac{\partial}{\partial x}\left(by_{|\zeta}(x,t)\right) = 0, & x \in (0,\ell), t \geq 0, \\ y_{|\zeta}(\ell,t) = y_{|\zeta}(0,t), \\ y(x,0) = y_{0|\zeta}(x) = y(x) + \zeta(x). \end{cases} \quad (5.59)$$

La résolution de ce type d'équation est triviale par la méthode des caractéristiques. Soit une courbe caractéristique $\mathcal{C}_\xi = \{(x_\xi(t), t), t \in \mathbb{R}_+ \mid x_\xi(0) = \xi\}$. On a sur cette courbe

$$\frac{d}{dt}y(x_\xi(t), t) = x'_\xi(t) \frac{\partial y}{\partial x}(x_\xi(t), t) + \frac{\partial y}{\partial t}(x_\xi(t), t).$$

Donc $y_{|\zeta}$ est constante sur les droites \mathcal{C}_ξ d'équation

$$x'_\xi(t) = -b \Rightarrow x_\xi(t) = -bt + \xi \bmod \ell.$$

On en déduit que

$$y_{|\zeta}(x, t) = y_{0|\zeta}((x + bt) \bmod \ell).$$

Pour l'espace d'état, on peut considérer $\mathcal{Y} = L^2(0, \ell)$ et la variable d'état est

$$y : t \in \mathbb{R}_+ \mapsto y(\cdot, t) \in L^2(0, \ell).$$

L'opérateur de la dynamique est

$$A : \begin{cases} H^1(0, \ell) \rightarrow L^2(0, \ell) \\ y \mapsto \frac{\partial}{\partial x}\left(by(x, t)\right) \end{cases}$$

avec $D(A) = \{y \in H^1(0, \ell), y(0) = y(\ell)\}$. On démontre simplement que A est anti-adjoint, puisque

$$\begin{aligned} \forall (y_1, y_2) \in D(A)^2, \quad \langle y_1, Ay_2 \rangle &= \int_0^\ell \frac{\partial}{\partial x} (by_1(x)) y_2(x) dx \\ &= [by_1(x)y_2(x)]_0^\ell - \int_0^\ell by_1(x) \frac{\partial}{\partial x} (y_2(x)) dx \\ &= -\langle y_1, Ay_2 \rangle, \end{aligned}$$

Ainsi

$$\forall y \in D(A), \quad \Re(\langle Ay, y \rangle) = 0.$$

et $(A, D(A))$ génère un groupe.

Imaginons que les mesures consistent en une observation de la solution sur un sous-intervalle $[\ell_1, \ell_2] \subset [0, \ell]$:

$$z^\varepsilon(x, t) = \mathbb{1}_{[\ell_1, \ell_2]} \check{y} + \text{bruit}.$$

Concernant cet opérateur d'observation, nous voyons facilement que C^* est donné par

$$C^* : \begin{cases} L^2(\ell_1, \ell_2) \rightarrow L^2(0, \ell) \\ z \mapsto \mathbb{1}_{(\ell_1, \ell_2)}(x)z(x). \end{cases}$$

Par ailleurs on démontre la condition d'observabilité donnée par la proposition suivante.

Proposition 11. *Le temps passé dans $[\ell_1, \ell_2]$ est dénoté $\tau_\xi = \frac{1}{b} \int_0^t \mathbb{1}_{[\ell_1, \ell_2]}(x_\xi(s)) ds$. Si $\tau_{min} = \min_\xi \tau_\xi > 0$, alors $\forall t \geq \tau_{min}$, la condition d'observabilité suivante est satisfaite :*

$$\exists \alpha > 0 \quad \left| \int_0^t \|y\|_{L^2(\ell_1, \ell_2)} ds \right| \geq \alpha \|y\|_{L^2(0, \ell)}.$$

Concernant le bruit de modèle, nous nous limiterons à $B = 0$. Mais on pourrait tout aussi bien regarder un bruit de modèle intervenant à travers

$$B : \begin{cases} H^2(0, \ell) \rightarrow L^2(0, \ell) \\ y \mapsto \varepsilon \Delta y \end{cases}$$

et dans ce cas $B \in \mathcal{L}(D(A^2), \mathcal{Y})$ est non borné.

On peut alors définir l'estimateur de Kalman sur ce système. L'écriture de l'estimateur sous forme opérateur a été donnée en (5.58).

Dans [41, 8], il est aussi évoqué la possibilité de formuler le filtre de Kalman à partir de la convolution avec un noyau π sous la forme

$$P(t)y : x \mapsto \pi(x, \cdot, t) * y = \int_0^\ell \pi(x, \tilde{x}, t)y(\tilde{x}) d\tilde{x}.$$

Dans ce cas, en reprenant (5.57) pour $(p, q) \in D(A)^2$, on a

$$\frac{d}{dt} \langle p, P(t)q \rangle_{\mathcal{Y}} = \int_0^\ell \int_0^\ell \frac{\partial \pi}{\partial t}(x, \tilde{x}, t) q(\tilde{x}) p(x) d\tilde{x} dx.$$

Puis par intégration par parties

$$\begin{aligned} \langle A^* p, P(t)q \rangle_{\mathcal{Y}} &= - \int_0^\ell \int_0^\ell b p'(x) \pi(x, \tilde{x}, t) q(\tilde{x}) d\tilde{x} dx \\ &= \int_0^\ell \int_0^\ell b \frac{\partial \pi}{\partial x}(x, \tilde{x}, t) p(x) q(\tilde{x}) d\tilde{x} dx - \left[b \pi(x, \tilde{x}, t) p(x) q(\tilde{x}) \right]_0^\ell, \end{aligned}$$

de même

$$\begin{aligned} \langle P(t)p, A^* q \rangle_{\mathcal{Y}} &= - \int_0^\ell \int_0^\ell \pi(x, \tilde{x}, t) p(\tilde{x}) b q'(x) d\tilde{x} dx \\ &= \int_0^\ell \int_0^\ell b \frac{\partial \pi}{\partial \tilde{x}}(x, \tilde{x}, t) p(x) q(\tilde{x}) d\tilde{x} dx - \left[b \pi(x, \tilde{x}, t) p(x) q(\tilde{x}) \right]_0^\ell, \end{aligned}$$

et enfin

$$\begin{aligned} \langle CP(t)p, CP(t)q \rangle_{\mathcal{Z}} &= \int_{\ell_1}^{\ell_2} \int_{\ell_1}^{\ell_2} \int_{\ell_1}^{\ell_2} \pi(x, \tilde{x}, t) p(\tilde{x}) \pi(\tilde{x}, \tilde{\tilde{x}}, t) q(\tilde{\tilde{x}}) dx d\tilde{x} d\tilde{\tilde{x}} \\ &= \int_{\ell_1}^{\ell_2} \int_{\ell_1}^{\ell_2} \left(\int_{\ell_1}^{\ell_2} \pi(\tilde{x}, x, t) \pi(\tilde{x}, \tilde{\tilde{x}}, t) d\tilde{x} \right) p(x) q(\tilde{\tilde{x}}) dx d\tilde{\tilde{x}}, \end{aligned}$$

qui nous donne π solution forte de

$$\begin{cases} \frac{\partial \pi}{\partial t} - b \frac{\partial \pi}{\partial x} - b \frac{\partial \pi}{\partial \tilde{x}} = \mathbb{1}_{(\ell_1, \ell_2)}(x, \tilde{x}) \int_{\ell_1}^{\ell_2} \pi(\tilde{x}, x, t) \pi(\tilde{x}, \tilde{\tilde{x}}, t) d\tilde{x}, & x \in (0, \ell), t \geq 0, \\ \pi(x, 0, t) = \pi(x, \ell, t) \\ \pi(0, \tilde{x}, t) = \pi(\ell, \tilde{x}, t) \\ \pi(x, \tilde{x}, 0) = \pi_{\diamond}(x, \tilde{x}). \end{cases} \quad (5.60)$$

Nous voyons ici que le noyau est un fonction de deux variables suivant une équation de transport intégro-différentielle. En particulier [41] indique qu'il serait intéressant de démontrer directement l'existence de solution du type de (5.60). Ceci permettrait de compléter la preuve succincte de [41] de l'existence d'un noyau de convolution associé à P . En tout état de cause elle permet de mieux comprendre l'impact de l'opérateur de covariance puisque

nous pouvons désormais écrire une forme forte de l'estimateur de Kalman

$$\begin{cases} \frac{\partial \hat{y}}{\partial t}(x, t) - \frac{\partial}{\partial x} \left(b\hat{y}(x, t) \right) = -\beta \int_{\ell_1}^{\ell_2} \pi(x, \tilde{x}, t) (z^\varepsilon(\tilde{x}, t) - \hat{y}(\tilde{x}, t)) d\tilde{x}, & x \in (0, \ell), t \geq 0, \\ \hat{y}(\ell, t) = \hat{y}(0, t), \\ y(x, 0) = y_0(x). \end{cases} \quad (5.61)$$

Imaginons maintenant que les mesures consistent en une observation de la solution sous la forme de coefficient de Fourier, par exemple

$$z^\varepsilon(x, t) = a_m = \int_0^\ell \sin\left(m\pi \frac{x}{\ell}\right) \check{y} + \text{bruit},$$

ou

$$z^\varepsilon(x, t) = b_m = \int_0^\ell \cos\left(m\pi \frac{x}{\ell}\right) \check{y} + \text{bruit}.$$

Nous savons bien que l'observabilité n'est acquise que si nous observons tous les coefficients de Fourier en même temps et que nous concaténons ces observations. Dans le cas contraire seule une partie régulière de la solution est reconstruisable d'où dans ce cas l'importance de l'*a priori* dans le processus d'estimation. En considérant qu'un seul a_m est observé, le noyau associé à P vérifie

$$\frac{\partial \pi}{\partial t} - b \frac{\partial \pi}{\partial x} - b \frac{\partial \pi}{\partial \tilde{x}} = \int_0^\ell \sin\left(m\pi \frac{\tilde{x}}{\ell}\right) \pi(\tilde{x}, x, t) d\tilde{x} \int_0^\ell \sin\left(m\pi \frac{\tilde{x}}{\ell}\right) \pi(\tilde{x}, \tilde{x}, t) d\tilde{x}, \quad x \in (0, \ell), t \geq 0,$$

et l'observateur

$$\begin{cases} \frac{\partial \hat{y}}{\partial t}(x, t) - \frac{\partial}{\partial x} \left(b\hat{y}(x, t) \right) = \\ \beta \int_0^\ell \pi(x, \tilde{x}, t) \sin\left(m\pi \frac{\tilde{x}}{\ell}\right) \left(z^\varepsilon(t) - \int_0^\ell \sin\left(m\pi \frac{\tilde{x}}{\ell}\right) \hat{y}(\tilde{x}, t) d\tilde{x} \right) d\tilde{x}, & x \in (0, \ell), t \geq 0, \\ \hat{y}(\ell, t) = \hat{y}(0, t), \\ y(x, 0) = y_0(x). \end{cases}$$

Un processus de dépolymérisation – Regardons désormais l'équation de dépolymérisation

$$\begin{cases} \frac{\partial}{\partial t} y_{|\zeta}(x, t) - \frac{\partial}{\partial x} \left(b y_{|\zeta}(x, t) \right) = 0, & x \in (0, \ell), t \geq 0, \\ y_{|\zeta}(\ell, t) = 0, \\ y_{|\zeta}(x, 0) = y_0(x). \end{cases} \quad (5.62)$$

où $y(x, t)$ représente une concentration de polymère de taille x . La borne supérieure ℓ est choisie de telle sorte qu'il n'y ait pas de polymère de taille supérieure à ℓ . Ce problème est bien posé, on le voit aisément avec les caractéristiques du système et d'ailleurs la solution de ce problème est

$$\forall x \in (0, \ell), t \geq 0, \quad y(x, t) = y_0(x + bt)$$

L'opérateur de la dynamique est cette fois

$$A : \begin{cases} H^1(0, \ell) \rightarrow L^2(0, \ell) \\ y \mapsto \frac{\partial}{\partial x} (by(x, t)) \end{cases}$$

avec désormais $D(A) = \{y \in H^1(0, \ell), y(\ell) = 0\}$. On démontre que A est désormais dissipatif. En effet on a cette fois

$$\begin{aligned} \forall (y_1, y_2) \in D(A)^2, \quad \langle y_1, Ay_2 \rangle &= \int_0^\ell \frac{\partial}{\partial x} (by_2(x)) y_1(x) dx \\ &= [by_1(x)y_2(x)]_0^\ell - \int_0^\ell by_2(x) \frac{\partial}{\partial x} (y_1(x)) dx \\ &= -\langle Ay_1, y_2 \rangle - by_1(0)y_2(0) \end{aligned}$$

donc

$$\forall y \in D(A), \quad \langle y, Ay \rangle = \frac{1}{2}(\langle y, Ay \rangle + \langle y, Ay \rangle) = -\frac{b}{2}y(0)^2 \leq 0.$$

Ainsi $(A, D(A))$ génère un semi-groupe de contraction.

On suppose que l'on dispose sur ce problème de mesures de Thioflavin T (ThT) fluorescence comme décrit au chapitre 1. Ces données fournissent une mesure de la masse totale depolymérisée. A une transformation linéaire prête, cela signifie que nous avons accès au premier moment de la concentration

$$z_{\text{tht}}^\varepsilon(t) = c_1 \int_0^\infty x \check{y}(x, t) dx + \text{erreur.}$$

Les techniques de Static Light Scattering (SLS) (voir aussi le chapitre 1), donnent par ailleurs une image du deuxième moment

$$z_{\text{sls}}^\varepsilon(t) = c_1 \int_0^\infty x^2 \check{y}(x, t) dx + c_2 + \text{erreur,}$$

où $c_1 \geq 0$, $c_2 \in \mathbb{R}$. Pour simplifier nous supposerons que $c_2 = 0$ et $c_1 = 1$ et nous noterons

$$\forall y \in \mathcal{Y}, \quad \mu_m = \int_0^\infty x^n y(x, t) dx.$$

Commençons par décrire dans ce cadre, l'approche 4D-Var pour une mesure de type μ_m . L'adjoint de $(A, D(A))$ est donné par

$$A^* : \begin{cases} H^1(0, \ell) \rightarrow L^2(0, \ell) \\ y \mapsto -b \frac{\partial}{\partial x} (y(x, t)) \end{cases}$$

où $D(A^*) = \{y \in H^1(0, \ell), y(0) = 0\}$. De même C^* est donné par

$$C^* : \begin{cases} \mathbb{R} \rightarrow L^2(0, \ell) \\ z \mapsto x^m z \end{cases}$$

L'état adjoint sur $[0, t]$ admet alors pour formulation forte

$$\begin{cases} \frac{\partial}{\partial t} q_{|\zeta, t}(x, s) - b \frac{\partial}{\partial x} q_{|\zeta, t}(x, s) = -\gamma x^m (z^\varepsilon(t) - \mu_m(t)), & x \in (0, \ell), s \in (0, t) \\ q_{|\zeta, t}(0, s) = 0, & s \in (0, t), \\ q_{|\zeta, t}(x, t) = 0, & x \in (0, \ell). \end{cases} \quad (5.63)$$

On voit de nouveau à partir des caractéristiques que ce système est bien posé. Pour l'estimateur moindres carrés, on a alors le problème aux deux bouts suivants

$$\begin{cases} \frac{\partial}{\partial t} \bar{y}_{|t}(x, s) - b \frac{\partial}{\partial x} (\bar{y}_{|t}(x, s)) = 0, & x \in (0, \ell), s \in (0, t) \\ \frac{\partial}{\partial t} \bar{q}_{|t}(x, s) - b \frac{\partial}{\partial x} (\bar{q}_{|t}(x, s)) = -\gamma x^m (z^\varepsilon(t) - \bar{\mu}_m(t)), & x \in (0, \ell), s \in (0, t) \\ \bar{y}_{|t}(\ell, s) = 0, & s \in (0, t), \\ \bar{q}_{|t}(0, s) = 0, & s \in (0, t), \\ \bar{y}_{|t}(x, 0) = y_0(x) + \alpha \bar{q}_{|t}(x, 0), & x \in (0, \ell), \\ \bar{q}_{|t}(x, t) = 0, & x \in (0, \ell). \end{cases} \quad (5.64)$$

Concernant l'estimateur de Kalman, on a cette fois

$$\begin{cases} \frac{\partial \hat{y}}{\partial t}(x, t) - b \frac{\partial}{\partial x} (\hat{y}(x, t)) = \\ \quad + \beta \int_0^\ell \pi(x, \tilde{x}, t) \tilde{x}^m (z^\varepsilon(t) - \hat{\mu}_m) d\tilde{x}, \quad x \in (0, \ell), t \geq 0, \\ \hat{y}(\ell, t) = 0, \quad t \geq 0 \\ \hat{y}(x, 0) = y_0(x), \quad x \in (0, \ell). \end{cases}$$

où

$$\begin{cases} \frac{\partial \pi}{\partial t} - b \frac{\partial \pi}{\partial x} - b \frac{\partial \pi}{\partial \tilde{x}} = \int_0^\ell \tilde{x}^m \pi(\tilde{x}, x, t) d\tilde{x} \int_0^\ell \tilde{x}^m \pi(\tilde{x}, \tilde{x}, t) d\tilde{x}, & (x, \tilde{x}) \in (0, \ell)^2, t \geq 0, \\ \pi(\ell, \tilde{x}, t) = 0, & \tilde{x} \in (0, \ell), t \geq 0, \\ \pi(x, \ell, t) = 0, & x \in (0, \ell), t \geq 0, \\ \pi(x, \tilde{x}, 0) = \pi_\diamond(x, \tilde{x}) & (x, \tilde{x}) \in (0, \ell)^2. \end{cases}$$

Exercice : Du point de vue de la modélisation, on pourrait remettre en cause la borne supérieure ℓ . Ceci peut être fait en ajoutant un bruit de modèle sous forme d'une condition de Dirichlet inconnue en ℓ . Quelles sont les conséquences sur ce que nous venons d'exposer et quel est l'opérateur d'erreur de modèle résultant.

5.3 Les observateurs de type Luenberger

Une question est donc de proposer des observateurs calculables pour des systèmes d'EDP. Une stratégie apparaît avec les observateurs de Luenberger [42]. L'idée n'est plus de s'appuyer sur un principe variationnel sous-jacent mais simplement de définir l'opérateur de gain tel que la propriété de stabilité de l'erreur d'estimation soit satisfaita, *i.e.* que l'erreur d'estimation décroisse au cours du temps pour t et idéalement qu'elle tende assymptotiquement vers 0 pour des données non bruitées :

$$\hat{y}(t) \xrightarrow[t \rightarrow \infty]{} \check{y}(t)$$

De plus on souhaite que l'estimateur soit robuste au bruit de mesure.

Dans un cadre linéaire, la dynamique de l'erreur $\tilde{y} = \check{y} - \hat{y}$ est très simple à écrire :

$$\begin{cases} \dot{\tilde{y}} = (A - GC)\tilde{y} + G(\check{z} - z^\varepsilon), \\ \tilde{y}(0) = \check{y}(0) - y_\diamond. \end{cases} \quad (5.65)$$

On peut donc concevoir un observateur efficace si on peut concevoir G tel que $A - GC$ est stable, éventuellement exponentiellement, vers 0.

Le cas où $A = A^*$ – Dans ce cas, on donne ici un résultat très utile permettant de proposer un observateur très simple à calculer.

Théorème 21. Soit $A = A^*$ et $C \in \mathcal{L}(\mathcal{Y}, \mathcal{Z})$. On suppose qu'il existe t_0 et m tels que pour toute solution de

$$\begin{cases} \dot{y}|_\zeta = Ay_\zeta \\ y(0) = \zeta \end{cases}$$

on a la condition d'observabilité

$$\forall t \geq t_0, \quad \int_0^t \|Cy_\zeta\|_{\mathcal{Z}} \geq m\|\zeta\|_{\mathcal{Y}},$$

alors le système

$$\begin{cases} \dot{\tilde{y}} = (A - C^*C)\tilde{y}, \\ \tilde{y}(0) = \zeta, \end{cases}$$

est exponentiellement stable à 0.

Regardons comment ce théorème s'applique dans les exemples que nous avons présentés précédemment. On reprend alors l'équation avec conditions aux limites périodiques. Dans ce cas, nous avons A anti-adjoint. Ainsi avec l'opérateur d'observation sur un sous-domaine qui vérifie la condition d'observabilité, nous pouvons assurer que l'observateur suivant est asymptotiquement convergent

$$\begin{cases} \frac{\partial \hat{y}}{\partial t}(x, t) - \frac{\partial}{\partial x} \left(b\hat{y}(x, t) \right) = -\beta \mathbb{1}_{(\ell_1, \ell_2)}(x) (z^\varepsilon(x, t) - \hat{y}(x, t)), & x \in (0, \ell), t \geq 0, \\ \hat{y}(\ell, t) = \hat{y}(0, t), & t \geq 0, \\ \hat{y}(x, 0) = y_0(x). & x \in (0, \ell). \end{cases} \quad (5.66)$$

Cet observateur doit être comparé à (5.61) où on voit que le noyau a été réduit à un Dirac $\delta(x - \tilde{x})$. Évidemment l'estimateur (5.66) est beaucoup plus simple à mettre en oeuvre et ne souffre d'aucune « malédiction de la dimensionnalité ». Une question difficile est d'arriver à démontrer que cet estimateur est efficace tout autant que (5.61). Ceci se fait par l'étude de la vitesse de stabilisation de $(A - C^*C)$. Cette question est le plus souvent ouverte et des réponses partielles sont obtenues par l'analyse du spectre de $(A - C^*C)$, i.e. les (λ, y_λ) tels que

$$(A - C^*C)y_\lambda = \lambda y_\lambda.$$

En effet toute erreur se décomposant sur les éléments propres se stabilisera à la vitesse $\sup_\lambda \Re(\lambda)$.

Un autre exemple peut être donné avec l'opérateur d'observation correspondant au a_m . Dans ce cas on a

$$\begin{cases} \frac{\partial \hat{y}}{\partial t}(x, t) - \frac{\partial}{\partial x} \left(b\hat{y}(x, t) \right) = -\beta \sin\left(m\pi \frac{x}{\ell}\right) \left(z^\varepsilon - \int_0^\ell \sin\left(m\pi \frac{\tilde{x}}{\ell}\right) \bar{y}^k(\tilde{x}, s) d\tilde{x} \right), & x \in (0, \ell), t \geq 0, \\ \hat{y}(\ell, t) = \hat{y}(0, t), & t \geq 0, \\ \hat{y}(x, 0) = y_0(x), & x \in (0, \ell). \end{cases}$$

Nous avons vu que l'observabilité n'était pas acquise dans ce cas. Cependant nous voyons que cet observateur permet quand même de faire en sorte que l'énergie de l'erreur vérifiant

$$\frac{d}{dt} \|\tilde{y}\|^2 = - \left(\int_0^\ell \sin\left(m\pi \frac{x}{\ell}\right) \tilde{y} dx \right)^2,$$

décroît. Pour aller plus loin, on peut alors essayer de montrer que cet observateur stabilise la partie observable de l'erreur [31].

L'observateur de Luenberger fait donc le choix de la convergence assymptotique plutôt que celui de la minimisation d'un critère moindre carré. L'intérêt est désormais d'avoir un observateur « calculable ». Evidemment ce gain a pour contre-partie de ne donner qu'une propriété assymptotique, et notamment la reconstruction de la condition initiale doit être repensée. Dans notre cas périodique, la question n'est pas très importante du fait de la périodicité en temps des solutions. Mais que faire dans le cas général ?

Heureusement dans le cas de stabilisation décrit par le théorème (21), A générant un groupe, il est possible de considérer une dynamique rétrograde. Ceci a conduit à la définition d'observateurs dits « back and force » permettant eux de reconstruire la condition initiale [2, 51].

Théorème 22. Soit $A = A^*$ et $C \in \mathcal{L}(\mathcal{Y}, \mathcal{Z})$. On suppose qu'il existe t_0 et m tels que pour tout solution de

$$\begin{cases} \dot{y}|_\zeta = Ay_\zeta \\ y(0) = \zeta \end{cases}$$

on a la condition d'observabilité

$$\forall t \geq t_0, \quad \int_0^t \|Cy_\zeta\|_{\mathcal{Z}} \geq m\|\zeta\|_{\mathcal{Y}}.$$

Alors l'observateur itératif vérifiant $\hat{y}_b^{-1}(0) = y_\diamond$ et pour tout $k \in \mathbb{N}$

$$\begin{cases} \dot{\hat{y}}^k(s) = A\hat{y}^k(s) + \gamma C^*(z^\varepsilon - C\hat{y}^k(s)), & s \in [0, t] \\ \hat{y}^k(0) = \hat{y}_b^{k-1}(0), \end{cases}$$

suiti de

$$\begin{cases} \dot{\hat{y}}_b^k(s) = A\hat{y}_b^k(s) - \gamma C^*(z^\varepsilon - C\hat{y}_b^k(s)), & s \in [0, t] \\ \hat{y}_b^k(t) = \hat{y}^k(t), \end{cases}$$

vérifie

$$\|\check{y}(0) - \hat{y}^k(0)\| \xrightarrow[k \rightarrow \infty]{} 0,$$

et reconstruit ainsi la condition initiale.

Ce théorème s'étend au cas C non borné et au cas où l'observabilité n'est pas acquise [31].

Autre cas de reversibilité – Le défaut du cadre précédent est qu'il ne s'applique qu'à des configurations où $A = A^*$. Pour notre problème de dépolymérisation, nous avons bien une équation hyperbolique mais les conditions aux limites nous écartent a priori du cadre précédent. De plus on peut se demander l'intérêt dans notre cas de dépolymérisation de fonder notre estimateur sur une propriété asymptotique alors que la solution disparaît en 0 en temps fini. En fait, nous pouvons quand même nous appuyer sur un résultat du type du théorème 22 pour peu que nous soyons capable de réintroduire l'information sortante en 0 dans l'observateur rétrograde. On peut donc proposer à partir de la convention $\hat{y}_b^{-1}(x, 0) = y_\diamond$:

$$\begin{cases} \frac{\partial \hat{y}^k}{\partial t}(x, s) - \frac{\partial}{\partial x} \left(b \hat{y}^k(x, s) \right) = \beta x^m \left(z^\varepsilon - \int_0^\ell \tilde{x}^m \bar{y}^k(\tilde{x}, s) d\tilde{x} \right), & x \in (0, \ell), s \in (0, t), \\ \hat{y}(\ell, s) = 0, & s \in (0, t), \\ y(x, 0) = \hat{y}_b^{k-1}(x, 0), & x \in (0, \ell). \end{cases}$$

suivi de

$$\begin{cases} \frac{\partial \hat{y}_b^k}{\partial t}(x, s) - \frac{\partial}{\partial x} \left(b \hat{y}_b^k(x, s) \right) = \beta x^m \left(z^\varepsilon - \int_0^\ell \tilde{x}^m \bar{y}_b^k(\tilde{x}, s) d\tilde{x} \right), & x \in (0, \ell), t \in (0, t), \\ \hat{y}_b^k(0, s) = \hat{y}^k(0, t-s), & s \in (0, t), \\ y_b^k(x, t) = \hat{y}^k(x, t), & x \in (0, \ell). \end{cases}$$

Exercice : Démontrer que cet estimateur « back and forth » permet de reconstruire la condition initiale (On discutera en fonction de m).

Ceci illustre comment l'analyse précise des EPDs permet de concevoir des observateurs spécifiques efficaces. Ainsi ce que nous perdons en générnicité par rapport à Kalman, nous le gagnons en praticabilité.

Chapitre 6

Estimation de la densité en statistique

Le problème de l'estimation de la densité peut s'énoncer comme suit :

soient X_1, \dots, X_n des variables aléatoires indépendantes identiquement distribuées de densité de probabilité $N(x)$ continue par rapport à la mesure de Lebesgue sur \mathbb{R} . Comment peut-on estimer $N(x)$?

Dans un cadre déterministe, c'est un problème qui peut sembler « déjà résolu » : en effet, la question est de déterminer une fonction à partir... de mesures sur cette fonction ! Ici, c'est l'échantillonnage X_1, \dots, X_n qui joue le rôle de la mesure bruitée N_ε vérifiant, dans le cadre du cours,

$$\|N - N_\varepsilon\|_{\mathcal{Z}} \leq \varepsilon.$$

En effet, si l'on prend $\mathcal{Y} = \mathcal{Z} = L^2(\mathbb{R})$ et $\Psi = Id$, il s'agit d'un problème trivial et bien posé.

Cependant, comme nous le verrons ci-dessous, le fait d'échantillonner est heuristiquement équivalent à prendre une mesure bruitée N_ε dans $H^{-\frac{1}{2}}(\mathbb{R})$ [45, 46] : on voit alors qu'il s'agit d'un problème mal posé, de degré $\frac{1}{2}$.

6.1 Un point de vue déterministe : suites régularisantes

Dans ce paragraphe, nous notons D l'opérateur de dérivation sur les distributions de \mathbb{R} dans \mathbb{R} , \mathcal{I} un opérateur d'intégration (par ex. $\mathcal{I}y(t) := \int_0^t y(s)ds$) et considérons le problème d'estimation de la dérivée k -ième. Reprenant les notations usuelles du cours, nous posons

donc

$$\Psi := \mathcal{I}^k, \quad \mathcal{Y} = L^p([0, 1]), \quad \mathcal{Z} = \mathcal{W}^{-\theta, p}([0, 1]),$$

avec $0 \leq \theta \leq 1$, $p \geq 1$. Nous avons donc étendu le cas hilbertien du cours à un cadre plus général d'espaces de Sobolev.

Le problème inverse que nous considérons s'écrit donc : Supposant $y \in \mathcal{W}^{m, p}([0, 1])$ avec $m \in \mathbb{N}$, notant $z = \mathcal{I}^k y$, et mesurant $z_\varepsilon \in \mathcal{W}^{-\theta, p}([0, 1])$ tel que

$$\|z_\varepsilon - z\|_{\mathcal{W}^{-\theta, p}([0, 1])} \leq \varepsilon,$$

comment définir un estimateur $y_{\varepsilon, \alpha}$ de y dans \mathcal{L}^p ?

6.1.1 Rappel sur les espaces de Sobolev

1. Pour $f : \mathbb{R} \rightarrow \mathbb{R}$ On note

$$\|f\|_{\mathcal{L}^p} := \left(\int_{\mathbb{R}} |f(x)|^p dx \right)^{1/p}.$$

On définit \mathcal{L}^p l'espace de Banach $\mathcal{L}^p((0, +\infty), dx)$ équipé de la norme $\|\cdot\|_{\mathcal{L}^p}$.

2. On note

$$\mathcal{W}^{m, p} := \left\{ f : [0, +\infty) \rightarrow \mathbb{R} \mid f, \dots, f^{(n)} \in \mathcal{L}^p \right\}$$

équipé de la norme

$$\|f\|_{\mathcal{W}^{m, p}(dx)} := \sum_{k=0}^m \|f^{(k)}\|_{\mathcal{L}^p(dx)},$$

$\mathcal{W}^{m, p}(dx)$ est un espace de Banach.

3. On définit de même $\mathcal{H}^m(dx) = \mathcal{W}^{m, 2}(dx)$.

4. On définit :

$$\mathcal{W}^{-1, p}(dx) := \left\{ f \in \mathcal{D}'(0, +\infty) : f = g + h', g, h \in \mathcal{L}^p(dx) \right\}$$

équipé de la norme

$$\|f\|_{\mathcal{W}^{-1, p}((1)dx)} := \inf_{f=g+h'} (\|g\|_{\mathcal{L}^p(dx)} + \|h\|_{\mathcal{L}^p(dx)}),$$

$\mathcal{W}^{-1, p}(dx)$ est un espace de Banach.

5. Enfin on définit pour $\theta \in [0, 1]$ l'espace de Banach $\mathcal{W}^{-\theta, p}(dx)$ par interpolation complexe :

$$\mathcal{W}^{-\theta, p}(dx) := [\mathcal{L}^p(dx), \mathcal{W}^{-1, p}(dx)]_\theta.$$

équipé de la norme standard pour l'interpolation complexe. C'est un espace de Banach. Pour la définition de $\mathcal{W}^{-\theta, p}([0, 1])$ nous renvoyons par exemple à [21] p.320 ou à [61] pour une définition par interpolation complexe entre L^p et $\mathcal{W}^{-1, p}$.

6.1.2 Régularisation par convolution avec une suite régularisante

Une façon très classique d'approcher une fonction par une suite de fonctions très régulières est d'utiliser une suite régularisante. Nous revisitons cette méthode dans le cadre de la théorie des problèmes inverses.

Soit $\rho \in \mathcal{C}_c^\infty(\mathbb{R})$, on définit $\rho_\alpha(x) = \frac{1}{\alpha}\rho(\frac{x}{\alpha})$, et on pose

$$y_{\varepsilon,\alpha} := \rho_\alpha * D^k z_\varepsilon = \rho_\alpha^{(k)} * z_\varepsilon, \quad y_\alpha(t) := \rho_\alpha * D^k z(t) = \rho_\alpha * y(t) = \int \rho_\alpha(s)y(t-s)ds.$$

Comme toujours, on décompose

$$\|y_{\varepsilon,\alpha} - y\|_{\mathcal{L}^p} \leq \|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{L}^p} + \|y_\alpha - y\|_{\mathcal{L}^p}.$$

Nous rappelons ci-dessous une série d'inégalités qui sont le pendant, dans les espaces de Sobolev, des inégalités vues dans le cadre général des espaces de Hilbert dans le chapitre 4, lemme 1. Nous supposons de plus

$$\int_{\mathbb{R}} \rho(x)dx = 1, \quad \int_{\mathbb{R}} x^k \rho(x)dx = 0, \quad \text{for } 1 \leq k \leq m_0. \quad (6.1)$$

6.1.3 Inégalités

Lemme 5 (Inégalités de convolution). *Soit $m \in \mathbb{N}$, $p \geq 1$, $\alpha \in (0, 1)$, $\rho \in \mathcal{C}_c^\infty(\mathbb{R})$ et m satisfaisant les hypothèses (6.1). On définit la fonction $\rho_\alpha(x) = \frac{1}{\alpha}\rho(\frac{x}{\alpha})$.*

i) *Si f est dans $W^{1,p}(\mathbb{R}_+)$, on a*

$$\|f - \rho_\alpha * f\|_{\mathcal{L}^p(\mathbb{R}_+)} \leq c_1 \alpha \|f\|_{W^{1,p}(\mathbb{R}_+)}, \quad (6.2)$$

où $c_1 = \|x\rho\|_{\mathcal{L}^1(\mathbb{R})}$

ii) *Soit $m \leq m_0$, si la fonction f est dans $\mathcal{W}^{m+1,p}(\mathbb{R}_+)$, on a*

$$\|f - \rho_\alpha * f\|_{\mathcal{L}^p(\mathbb{R}_+)} \leq c_2 \alpha^{m+1} \|f\|_{\mathcal{W}^{m+1,p}(\mathbb{R}_+)}, \quad (6.3)$$

où $c_2 = \frac{1}{m!} \|x^{m+1} \rho(x)\|_{\mathcal{L}^1(\mathbb{R})}$.

iii) *De plus on a*

$$\|\rho_\alpha * f^{(n)}\|_{\mathcal{L}^p(\mathbb{R}_+)} = \|\rho_\alpha^{(n)} * f\|_{\mathcal{L}^p(\mathbb{R}_+)} \leq c_3 \alpha^{-n} \|f\|_{\mathcal{L}^p(\mathbb{R}_+)}, \quad (6.4)$$

où $c_3 = \|\rho^{(n)}\|_{\mathcal{L}^1(\mathbb{R})}$.

iv) *Soit $s \in [0, 1]$, si la fonction f est dans $\mathcal{W}^{-s,p}(\mathbb{R}_+)$ et $\rho, \rho' \in \mathcal{L}^1(\mathbb{R})$, on a*

$$\|\rho_\alpha * f\|_{\mathcal{L}^p(\mathbb{R}_+)} \leq c_4 \alpha^{-s} \|f\|_{\mathcal{W}^{-s,p}(\mathbb{R}_+)}, \quad (6.5)$$

où c_4 dépend de $\|\rho\|_{\mathcal{L}^1(\mathbb{R})}$ et $\|\rho'\|_{\mathcal{L}^1(\mathbb{R})}$.

v) Soit $s \in [0, 1]$, si la fonction f est dans $\mathcal{W}^{-s,p}(\mathbb{R}_+)$ et $\rho^{(n)}, \rho^{(n+1)} \in \mathcal{L}^1(\mathbb{R}_+)$, on a

$$\|\rho_\alpha * f^{(n)}\|_{\mathcal{L}^p(\mathbb{R}_+)} \leq c_5 \alpha^{-(n+s)} \|f\|_{\mathcal{W}^{-s,p}(\mathbb{R}_+)}, \quad (6.6)$$

où c_5 dépend de $\|\rho^{(n)}\|_{\mathcal{L}^1(\mathbb{R})}$, $\|\rho^{(n+1)}\|_{\mathcal{L}^1(\mathbb{R})}$.

Preuve.

i) et ii)] On remarque que

$$f(x) - f_\alpha(x) = \int \rho_\alpha(t)(f(x) - f(x-t))dy = \int \rho(t)(f(x) - f(x-\alpha t))dt,$$

et on fait un développement de Taylor avec reste intégral :

$$f(x) - f(x-\alpha t) = \sum_{k=1}^m f^{(k)}(x) \frac{(-\alpha t)^k}{k!} + \frac{1}{m!} (-\alpha t)^{m+1} \int_0^1 u^m f^{(m+1)}(x - \alpha ut) du,$$

d'où en intégrant et en utilisant (6.1)

$$(f(x) - f_\alpha(x))^p = \left(\frac{1}{m!} \int \rho(t)(\alpha t)^{m+1} \int_0^1 u^m f^{(m+1)}(x - \alpha ut) du dt \right)^p$$

et donc par inégalité de Hölder (on écrit $\rho t^{m+1} = (\rho t^{m+1})^{\frac{1}{p} + \frac{1}{p'}}$)

$$\begin{aligned} |f(x) - f_\alpha(x)|^p &\leq \frac{\alpha^{p(m+1)}}{m!^p} \left(\int \rho(t)|t|^{m+1} dt \right)^{\frac{p}{p'}} \left(\int_0^1 \int |u^{mp}| \cdot |f^{(m+1)}(x - \alpha ut)|^p \rho(t)|t|^{m+1} du dt \right)^{\frac{1}{p'}} \\ \int |f(x) - f_\alpha(x)|^p dx &\leq \frac{\alpha^{p(m+1)}}{m!^p} M_{m+1}(\rho)^{\frac{p}{p'}} \int_0^1 \int \int |f^{(m+1)}(x')|^p \rho(t)t^{m+1} du dt dx' \\ &\leq C(\rho)^p \alpha^{p(m+1)} \|f\|_{\mathcal{W}^{m+1,p}}^p, \end{aligned}$$

avec $C(\rho) = \int \rho(t)|t|^{m+1} dt$.

iii) Pour $n = 0$: toujours la même décomposition $\rho = \rho^{\frac{1}{p} + \frac{1}{p'}}$

$$|\rho_\alpha * f(x)| \leq \|\rho\|_{\mathcal{L}^1}^{\frac{1}{p'}} \left(\int |\rho(s)| \cdot |f(x - \alpha s)|^p dx \right)^{\frac{1}{p}}.$$

Donc

$$\int |\rho_\alpha * f(x)|^p dx \leq C(\rho) \left(\int |\rho(s)| \cdot |f(x - \alpha s)|^p dx ds \right) = C(\rho) \|f\|_{\mathcal{L}^p}^p.$$

On généralise alors car $\rho_\alpha^{(n)} = \alpha^{-n}(\rho^{(n)})_\alpha$.

- iv) On le prouve pour $\theta = 1$ puis par interpolation. On écrit $f = g + h'$ et on utilise les inégalités (iii) :
- v) Idem avec $\rho^{(n)}$.

■

De ces inégalités on déduit : par l'inégalité (v)

$$\|y_{\varepsilon,\alpha} - y_\alpha\|_{\mathcal{L}^p} = \|\rho_\alpha * (z_\varepsilon - z)^{(k)}\|_{\mathcal{L}^p} \leq C(\rho) \frac{\varepsilon}{\alpha^{k+\theta}},$$

et par l'inégalité (ii)

$$\|y_\alpha - y\|_{\mathcal{L}^p} = \|\rho_\alpha * y - y\|_{\mathcal{L}^p} \leq C(\rho) \alpha^m \|y\|_{\mathcal{W}^{m,p}}$$

Erreur optimale en $O(\varepsilon^{\frac{m}{k+\theta+m}})$ pour α en $O(\varepsilon^{\frac{1}{k+\theta+m}})$. Ce taux est cohérent avec le taux de l'estimation de la dérivée d'une fonction vu précédemment : il suffit de prendre $m = 2s$, $\theta = 0$ et $k = 1$.

6.2 Estimateurs à noyau d'une densité

Cette partie est très largement inspirée du cours d'Alexandre Tsybakov [60].

Soit X_1, \dots, X_n n variables aléatoires indépendantes identiquement distribuées de densité de probabilité $f(x)$ continue par rapport à la mesure de Lebesgue sur \mathbb{R} . On souhaite estimer f à partir de réalisations x_1, \dots, x_n des X_i . On note $F(x)$ la fonction de répartition de X_i , i.e.

$$F(x) = \int_0^x f(s) ds.$$

On définit la fonction de répartition empirique par

$$F_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \leq x},$$

la loi forte des grands nombres dit que

$$F_n(x) \xrightarrow{p.s.} F(x).$$

Si on dérive (au sens des distributions) F_n , on obtient donc une « mesure bruitée » de f donnée par la *mesure empirique* f_n :

$$f_n(x) := \frac{1}{n} \sum_{i=1}^n \delta_{x=x_i}.$$

Prenant f_n comme un équivalent d'une mesure f_ε , on comprend mieux le lien avec les problèmes inverses déterministes : estimer f revient en effet à estimer la dérivée F' de F à partir d'une mesure F_n qui joue le rôle de F_ε . Il faut ici de plus tenir compte du cadre probabiliste.

Par le théorème de la limite centrale, on sait que, sous des hypothèses assez générales, $\sqrt{n}(|F_n(x) - F(x)|)$ converge en loi vers une loi normale à x fixé. Cela conduit à l'équivalence entre ε en déterministe et $\frac{1}{\sqrt{n}}$ en estimation de la densité. On verra ci-dessous que les estimations obtenues sont effectivement consistantes avec les estimations déterministes en prenant $\varepsilon = n^{-\frac{1}{2}}$.

On procède tout d'abord exactement comme pour la régularisation d'une fonction déterministe : avec les mêmes notations que dans le paragraphe précédent, on définit

$$f_{\alpha,n} = \rho_\alpha * f_n.$$

Par souci de cohérence avec les habitudes en statistiques, on adopte les notations suivantes.

- Le paramètre de régularisation α est noté h et nommé *fenêtre* (*width* ou *bandwidth* en anglais),
- le *noyau* ρ est noté K ,
- l'estimateur dit *estimateur à noyau* $f_{\alpha,n}$ est noté \hat{f}_n .

On note donc

$$\hat{f}_n(x; X_1, \dots, X_n) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right).$$

Il s'agit d'une variable aléatoire $n-$ variée¹. Dans un cadre déterministe, nous chercherions à évaluer $\int |f(x) - \hat{f}_n(x)|^p dx$ avec $p \in [1, \infty]$. Dans le cadre probabiliste dans lequel nous nous situons maintenant, il faut évaluer la probabilité de chaque tirage : on définit donc, pour chaque x , le *risque quadratique* (Mean Squared Error ou MSE) défini par

$$MSE(x) := \mathbb{E}_{f^{\otimes n}} [(f(x) - \hat{f}_n(x))^2] = \int \dots \int \left(f(x) - \hat{f}_n(x; x_1, \dots, x_n) \right)^2 \prod_{i=1}^n f(x_i) dx_i.$$

Pour faire une équivalence plus proche de l'analyse déterministe, il faudrait que nous regardions

$$\mathbb{E}_{f^{\otimes n}} [\|f(x) - \hat{f}_n(x)\|_2^2] = \int \dots \int \sqrt{\int (f(x) - \hat{f}_n(x; x_1, \dots, x_n))^2 dx} \prod_{i=1}^n f(x_i) dx_i,$$

1. Rappel : la densité f d'une v.a. X est définie au sens faible par $\mathbb{E}_f [\phi(X)] := \int \phi(x) f(x) dx$ pour toute fonction test ϕ . Pour X_1, \dots, X_n n v.a. i.i.d. de loi f , leur densité est la densité produit $f^{\otimes n}$, qui vérifie $\mathbb{E}_{f^{\otimes n}} [\phi(X_1, \dots, X_n)] = \int \dots \int \phi(x_1, \dots, x_n) f(x_1) \dots f(x_n) dx_1 \dots dx_n$

mais cela rajouterait un peu de complexité inutile pour exposer les idées principales de l'estimation par noyau. On cherche donc à estimer une borne uniforme en x sur $MSE(x)$. Comme en déterministe, on va décomposer la différence $f - \hat{f}_n$ en $f - \hat{f} + \hat{f} - \hat{f}_n$. D'une part, l'analogie avec ce qui précède pousse donc à définir \hat{f} comme on a défini f_α ci-dessus, par

$$\hat{f} := K_h * f.$$

D'autre part, la vision probabiliste fait remarquer que f est une fonction (donc déterministe) tandis que \hat{f}_n est une variable aléatoire $n-$ variée. Il est donc naturel de définir \hat{f} comme l'espérance de cette variable aléatoire, *i.e.*

$$\hat{f} := \mathbb{E}_{f^{\otimes n}} [\hat{f}_n(x; X_1, \dots, X_n)].$$

En réalité, ces deux définitions sont identiques, par le simple calcul suivant

$$\begin{aligned} \mathbb{E}_{f^{\otimes n}} [\hat{f}_n(x; X_1, \dots, X_n)] &= \int \dots \int \hat{f}_n(x, x_1, \dots, x_n) \prod_{i=1}^n f(x_i) dx_i \\ &= \int \dots \int \frac{1}{n} \left(\sum_{i=1}^n K_h(x - x_i) \right) \prod_{i=1}^n f(x_i) dx_i \\ &= \frac{1}{n} \sum_{i=1}^n \int K_h(x - x_i) f(x_i) dx_i \prod_{j \neq i} \int f(x_j) dx_j \\ &= \frac{1}{n} \sum_{i=1}^n \int K_h(x - x_i) f(x_i) dx_i = K_h * f \end{aligned}$$

puisque $\int f(x_j) dx_j = 1$. De plus, on a de façon immédiate la décomposition biais-variance suivante :

$$MSE(x) = b^2(x) + \sigma^2(x)$$

avec le biais, quantité déterministe définie exactement comme dans tout ce qui précède :

$$b(x) = f(x) - \hat{f}(x),$$

et la variance, dont la dénomination trouve enfin sa justification dans l'écriture ci-dessous (il s'agit de la variance relative à la variable aléatoire $\hat{f}_n(x; X_1, \dots, X_n)$)

$$\sigma^2(x) = \mathbb{E} \left[(\hat{f}_n(x) - \hat{f}(x))^2 \right] = \mathbb{E} \left[(\hat{f}_n(x) - \mathbb{E}_{f^{\otimes n}} [\hat{f}_n(x)])^2 \right].$$

Grâce à toute l'analyse qui précède, et sous les mêmes hypothèses, le lemme 5 de la partie 6.1.3 (inégalité ii]) nous donne immédiatement l'estimation suivante pour le biais :

$$b^2(x) \leq C(K, s) h^{2m} \|f\|_{W^{m,\infty}(\mathbb{R})}^2.$$

En ce qui concerne la variance, on a la proposition suivante.

Proposition 12. Si $K \in L^2(\mathbb{R})$ et $f \in L^\infty(\mathbb{R})$ on a

$$\|\sigma^2\|_{L^\infty} \leq \frac{1}{nh} \|K\|_{L^2}^2 \|f\|_{L^\infty}.$$

Preuve.

$$\begin{aligned}\sigma^2(x) &= \mathbb{E} \left[\left(\hat{f}_n(x) - \mathbb{E}_{f^{\otimes n}}[\hat{f}_n(x)] \right)^2 \right] \\ &= \mathbb{E} \left[\frac{1}{n^2} \left(\sum_{i=1}^n K_h(x - X_i) - \mathbb{E}_f[K_h(x - X_i)] \right)^2 \right] \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \mathbb{E} \left[\left(K\left(\frac{x-X_i}{h}\right) - \mathbb{E}_f[K\left(\frac{x-X_i}{h}\right)] \right)^2 \right] \\ &= \frac{1}{nh^2} \int \left(K\left(\frac{x-y}{h}\right) - K * f(x) \right)^2 f(y) dy \\ &\leq \frac{1}{nh^2} \int K\left(\frac{x-y}{h}\right)^2 f(y) dy \leq \frac{1}{nh} \|f\|_{L^\infty} \int K(z)^2 dz.\end{aligned}$$

■

Rassemblant ces inégalités, on obtient l'estimation suivante pour le risque quadratique :

$$\|MSE\|_{L^\infty} \leq C \left(h^{2m} + \frac{1}{nh} \right),$$

qui est optimale pour $h = O(n^{-\frac{1}{2m+1}})$, auquel cas l'estimation d'erreur pour \sqrt{MSE} est en $O(n^{-\frac{m}{2m+1}})$.

Notant, comme le théorème de la limite centrale nous y engage, $\varepsilon = n^{-\frac{1}{2}}$, c'est donc une estimation d'erreur en $O(\varepsilon^{\frac{m}{m+\frac{1}{2}}})$: on reconnaît l'estimation du paragraphe 6.1 en prenant $k = 0$ (puisque nous considérons ici $\Psi = Id$) et $\theta = \frac{1}{2}$. Cela justifie l'heuristique de « problème mal posé de degré 1/2 » pour l'estimation de la densité, annoncée en introduction de ce chapitre.

Pour l'implémentation numérique, des solveurs comme la fonction `ksdensity` de Matlab permettent des calculs rapides et proposent des algorithmes pour des choix automatiques de la fenêtre h . L'analyse statistique a développé de nombreuses méthodes - *adaptatives* est l'équivalent des méthodes *a posteriori* des problèmes inverses déterministes, *les inégalités oracle* étant le pendant des estimations optimales.

Chapitre 7

Retour sur les applications en dynamique des populations

Ce chapitre est la suite du chapitre 1, dans lequel les problèmes étaient énoncés mais non encore traités. Pour un traitement rigoureux, il fallait en effet avoir abordé au minimum la théorie générale des problèmes inverses, au chapitre 2. Disposant donc maintenant des outils nécessaires, nous pouvons à présent poser et résoudre un certain nombre de problèmes inverses - et aussi, donner un éclairage sur les problèmes inverses encore ouverts lors de l'écriture de ce cours. Par rapport aux chapitres précédents, qui constituaient la partie la plus fondamentale du cours, ce chapitre correspond donc à un cours de recherche, *i.e.* il présente des résultats récents faisant appel aux théories bien établies qui précédent.

7.1 La division cellulaire

Cette partie fait suite à la partie 1.1, et plus spécialement aux équations écrites en 1.1.3, au comportement en temps long vu en 1.1.4, et aux problèmes énoncés de façon assez générale dans la partie 1.1.5.

7.1.1 Interprétation des données

Rappelons les données présentées dans la partie 1.1.1 : on peut mesurer ou bien un échantillon de tailles d'individus à un instant donné, ou bien les tailles et âges des individus à des temps successifs.

Dans le cadre d'un modèle donné, par exemple dans le cadre du modèle structuré en taille, on interprète un échantillon (x_1, \dots, x_k) de tailles d'individus à un instant donné comme étant la réalisation de k variables aléatoires indépendantes identiquement distribuées de loi $\frac{n(t,x)}{\int n(t,x)dx}$, où $n(t,x)$ est l'unique solution d'un problème de croissance-

fragmentation de type (1.2), où le taux de croissance $g(x)$, le taux de division $B(x)$ et le noyau de fragmentation $k(y, x)$ restent à estimer.

On peut simplifier ce problème en utilisant l'analyse asymptotique vue en 1.1.4 : dans des conditions expérimentales particulières (milieu non limitant en terme d'espace ou de nourriture), et sous certaines hypothèses sur les paramètres de l'équation, on a vu que la population tendait exponentiellement vite vers un comportement de type stationnaire donné par le problème aux valeurs propres (1.4). On peut donc supposer $\frac{n(t,x)}{\int n(t,x)dx}$ suffisamment proche de $N(x)$ solution de (1.4) pour qu'on les confonde.

L'analyse asymptotique permet aussi une utilisation particulièrement agréable des données au cours du temps : en se plaçant à partir d'un temps suffisamment grand, on considère que *toute* donnée de taille mesurée à *tout* temps est la réalisation d'une variable aléatoire indépendante des autres données et de loi $N(x)$. Sur ces données temporelles, si les outils de mesure et d'analyse d'image le permettent, on peut aussi isoler de façon spécifique les cellules en train de se diviser et les cellules qui viennent de naître : en se plaçant de même en temps suffisamment grand, on peut alors supposer que la mesure d'une taille de cellule au moment où elle se divise, par exemple, est la réalisation d'une variable aléatoire de loi $\frac{B(x)N(x)}{\int B(x)N(x)dx}$.

Dans cette interprétation statistique des données sur des échantillons se cachent plusieurs problèmes mathématiques profonds, et pour certains ouverts. En effet, prouver que ces échantillons - qui ne sont certainement pas indépendants puisque les cellules sont toutes reliées les unes aux autres selon un arbre généalogique - se comportent *comme* s'il s'agissait de variables indépendantes est un problème probabiliste difficile, partiellement résolu dans [32] par exemple.

Une fois admise la validité de l'hypothèse d'un échantillon i.i.d., les méthodes statistiques d'estimation de la densité comme par exemple les méthodes d'estimation par noyaux, cf. ch. 6 nous permettent d'avoir accès à une mesure bruité de la densité $N(x)$, ainsi qu'à la vitesse de croissance $g(x)$ et au noyau de fragmentation $k(y, x)$.

7.1.2 Estimation sur le problème en âge : un problème jouet instructif

Dans cette partie, nous supposons que la population de cellules considérée

– obéit à l'équation de renouvellement (1.1) sans taux de mort, soit

$$\begin{cases} \frac{\partial}{\partial t}n(t, a) + \frac{\partial}{\partial a}n(t, a) = -B(a)n(t, a), \\ n(t, a=0) = 2 \int_0^\infty B(a)n(t, a)da, \quad n(t=0, a) = n^0(a), \end{cases} \quad (7.1)$$

– a déjà atteint l'état asymptotique (cf partie 1.1.4) *i.e.* $\frac{n(t,a)}{\int n(t,a)da} \approx N(a)$ où $(\lambda, N(a))$

est l'unique solution du problème aux valeurs propres

$$\begin{cases} \lambda N(a) + \frac{\partial}{\partial a} N(a) = -B(a)N(a), \\ N(a=0) = 2 \int_0^\infty B(a)N(a)da, \quad \int N(a)da = 1, \quad N(a) > 0, \quad \lambda > 0, \end{cases} \quad (7.2)$$

– peut être mesurée, de sorte que nous avons accès à une mesure bruitée de $(\lambda, N(a))$ unique solution de (7.2). On modélise cette erreur de mesure par exemple par $N_\varepsilon(a)$ vérifiant

$$\|N - N_\varepsilon\|_{L^2(\mathbb{R}_+)} \leq \varepsilon.$$

Le problème inverse considéré s'écrit comme suit.

Estimer $B(a)$ à travers une mesure bruitée $(\lambda_\varepsilon, N_\varepsilon(a))$ vérifiant

$$\|N - N_\varepsilon\|_{L^2(\mathbb{R}_+)} \leq \varepsilon, \quad |\lambda - \lambda_\varepsilon|\varepsilon,$$

où (λ, N) est la solution de (7.2).

On remarque tout d'abord que B n'apparaît dans l'équation que multiplié par N . On change donc d'inconnue dans le problème inverse et on note

$$h(a) := B(a)N(a),$$

qu'on cherche à estimer par une fonction $h_{\varepsilon,\alpha}$. Une fois h estimé, on pourra poser par exemple

$$B_{\varepsilon,\alpha} := \mathbb{1}_{N_\varepsilon \geq \sigma > 0} \frac{B_{\varepsilon,\alpha}}{N_\varepsilon} + f(a)\mathbb{1}_{N_\varepsilon < \sigma},$$

où σ est un seuil adéquat au moins supérieur à ε . **On ne pourra donc pas estimer B aux endroits où N est nul, à moins d'y avoir un a priori par ailleurs, symbolisé ici par une fonction $f(a)$ donnée.**

Posé en h , le problème inverse devient

Estimer $h(a)$ solution de

$$\begin{cases} \lambda N(a) + \frac{\partial}{\partial a} N(a) = -h(a), \\ N(a=0) = 2 \int_0^\infty h(a)da, \quad \int N(a)da = 1, \quad N(a) > 0, \quad \lambda > 0, \end{cases}$$

à travers une mesure bruitée $(\lambda_\varepsilon, N_\varepsilon(a))$ vérifiant

$$\|N - N_\varepsilon\|_{L^2(\mathbb{R}_+)} \leq \varepsilon, \quad |\lambda - \lambda_\varepsilon|\varepsilon.$$

Le premier constat est que dans cette formulation le problème inverse est devenu linéaire, ce qu'il n'était pas. De plus, on reconnaît ici le problème d'estimation de la dérivée :

si l'on dispose d'une méthode de régularisation qui permette d'estimer $N'(a)$, la solution h s'en déduit immédiatement en posant par exemple

$$h_{\varepsilon,\alpha} := -\lambda_\varepsilon N_\varepsilon - N'_{\varepsilon,\alpha}(a),$$

ou alternativement (si l'on souhaite plus de régularité sur $h_{\varepsilon,\alpha}$ par exemple)

$$h_{\varepsilon,\alpha} := -\lambda_\varepsilon N_{\varepsilon,\alpha} - N'_{\varepsilon,\alpha}(a),$$

et on aura pour elle les mêmes estimations que pour la dérivée. **L'énoncé exact et la preuve d'une estimation, sur la base des chapitres exposant la théorie générale des problèmes inverses linéaires, est laissé en exercice.**

7.1.3 Estimation sur le problème en taille

Dans cette partie, nous supposons que la population de cellules considérée

- obéit à l'équation de croissance-fragmentation (1.2)
- a déjà atteint l'état asymptotique (cf partie 1.1.4) i.e. $\frac{n(t,x)}{\int n(t,x)dx} \approx N(x)$ où $(\lambda, N(x))$ est l'unique solution du problème aux valeurs propres (1.4)
- peut être mesurée, et que $g(x)$, λ et $k(y, x)$ sont connus par ailleurs, de sorte que nous avons accès à une mesure bruitée de $(\lambda, N(x))$ unique solution de (1.4).

Le problème inverse considéré ici s'écrit donc comme suit.

Estimer $B(x)$ à travers une mesure bruitée $N_\varepsilon(x)$ vérifiant

$$\|N - N_\varepsilon\|_{L^2(\mathbb{R}_+)} \leq \varepsilon$$

où $N(x)$ est solution de (7.2), λ , g et $k(y, x)$ étant par ailleurs connus.

De même que pour le problème en âge, on remarque que B n'apparaît dans l'équation que multiplié par N , et on pose donc

$$h(x) := B(x)N(x),$$

qui est solution de

$$\lambda N(x) + \frac{\partial}{\partial x}(g(x)N(x)) = -h(x) + 2 \int_x^\infty k(y, x)h(y)dy.$$

Limitons-nous tout d'abord au cas de la mitose égale, où les deux cellules filles mesurent exactement la moitié de la taille de la cellule mère, ce qui correspond à $k(y, x) := \delta_{x=\frac{y}{2}}$.

Comme pour le problème en âge, nous remarquons que le problème inverse formulé ainsi est devenu linéaire. Nous pouvons réécrire ce problème sous la forme

$$\mathcal{L}(h)(x) = \mathcal{F}(N)(x),$$

où l'on définit

$$\mathcal{L}(h)(x) := 4h(2x) - h(x), \quad \mathcal{F}(N)(x) := \lambda N(x) + \frac{\partial}{\partial x}(g(x)N(x)). \quad (7.3)$$

L'opérateur $\mathcal{F} := \lambda Id + \frac{d}{dx}(gId)$ s'inverse en suivant la ligne du problème en âge : sitôt qu'on a un estimateur de l'opérateur de dérivation, on en déduit un estimateur de $\mathcal{F}(N)$, que l'on peut noter ou bien $\mathcal{F}_\alpha(N_\varepsilon)$ ou bien $\mathcal{F}(N_{\varepsilon,\alpha})$ selon le type de méthode de régularisation (*i.e.* selon que la régularisation porte sur l'opérateur \mathcal{F} ou sur la fonction N_ε).

L'opérateur \mathcal{L} nécessite une étude spécifique : en effet, son caractère bien ou mal posé dépend fortement de l'espace où l'on se trouve. On a besoin du lemme suivant.

Lemme 6 ([26], proposition A.1. et lemme A.2.). *Soit \mathcal{L} défini par (7.3) sur l'espace*

$$\mathcal{Y} = L^2(\mathbb{R}_+, x^p dx) := \left\{ f : \mathbb{R}_+ \rightarrow \mathbb{R}, \quad \int_0^\infty f^2(x)x^p dx < +\infty \right\}.$$

Alors pour $f \in \mathcal{Y}$, $\mathcal{L}(f) \in \mathcal{Y}$ et si $p \neq 3$, \mathcal{L} est une bijection linéaire continue de \mathcal{Y} dans \mathcal{Y} . De plus, on a

$$\begin{aligned} \text{si } \mathcal{Y} = L^2(x^p dx), \quad p < 3 \quad \text{implique} \quad \mathcal{L}^{-1}(f)(x) &= \sum_{k=1}^{\infty} 2^{-2n} f(2^{-n}x) \\ \text{si } \mathcal{Y} = L^2(x^p dx), \quad p > 3 \quad \text{implique} \quad \mathcal{L}^{-1}(f)(x) &= - \sum_{k=0}^{\infty} 2^{-2n} f(2^{-n}x). \end{aligned}$$

Dans l'espace des distributions $\mathcal{D}'(\mathbb{R}_+)$, le noyau de l'opérateur \mathcal{L} est l'ensemble des distributions du type $\frac{f(\log(x))}{x^2}$ avec $f \in \mathcal{D}'(\mathbb{R})$ une distribution log(2) périodique.

Pour la preuve, nous renvoyons à l'annexe de l'article [26], proposition A.1. et lemme A.2.

A partir de ce résultat, conjugué avec l'étude générale des problèmes inverses linéaires, il est facile d'énoncer un résultat d'estimation pour h dans un espace $\mathcal{Y} = L^2(\mathbb{R}_+, x^p dx)$. Cependant, en raison du bruit de mesure, selon qu'on choisit une estimation dans un espace avec $p < 3$ ou avec $p > 3$, on voit qu'on n'aura pas la même solution : la différence entre les deux sera égale à

$$\sum_{k=-\infty}^{\infty} 2^{-2n} \mathcal{F}(N_{\varepsilon,\alpha})(2^{-n}x)$$

qui n'a aucune raison d'être nulle (tandis que si l'on remplace $\mathcal{F}(N_{\varepsilon,\alpha})$ par $\mathcal{F}(N)$ dans cette formule, cette fonction est bien nulle, en raison des propriétés de l'équation même). Même, cette somme peut parfaitement n'appartenir à aucun espace $L^2(x^p dx)$! Or, on cherche en fait une solution qui a une bonne décroissance en 0 et en l'infini : en effet, la fonction recherchée $h(x) = B(x)N(x)$ représente la répartition des tailles des cellules à la division, répartition qui tend rapidement vers 0 en 0 comme en l'infini. Cela signifie qu'on souhaite pouvoir estimer h dans un espace de type $\mathcal{Y} = L^2((x^{p_1} + x^{p_2})dx)$ par exemple, avec $p_1 \leq 0$ et $p_2 > 3$ grand. On propose donc un nouvel estimateur, donné par

$$h_{\varepsilon,\alpha}(x) := \mathbb{1}_{x < x_0} \mathcal{L}_{p < 3}^{-1} \left(\mathcal{F}(N_{\varepsilon,\alpha}) \right) + \mathbb{1}_{x > x_0} \mathcal{L}_{p > 3}^{-1} \left(\mathcal{F}(N_{\varepsilon,\alpha}) \right).$$

La preuve d'une estimation précise pour $h_{\varepsilon,\alpha}$ est laissée en exercice ; on pourra aussi s'aider des articles [26, 19].

7.2 La polymérisation des protéines

Cette partie fait directement suite au problème énoncé dans la partie 1.2.3, dont le cadre était l'objet de la partie 1.2. A partir de tout ce qui précède, la résolution est laissée en exercice.

Bibliographie

- [1] A S Ackleh, H T Banks, K Deng, and S Hu. Parameter estimation in a coupled system of nonlinear size-structured populations. *Mathematical biosciences and engineering MBE*, 2(2) :289–315, 2005.
- [2] D Auroux and Jacques Blum. A nudging-based data assimilation method : the Back and Forth Nudging (BFN) algorithm. *Nonlinear Processes In Geophysics*, 15(2) :305–319, January 2008.
- [3] A.B. Bakushinskii. Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *USSR Comp. Math. Phys.*, 24(4) :181–182, 1984.
- [4] J. M. Ball and J. Carr. The discrete coagulation-fragmentation equations : existence, uniqueness, and density conservation. *J. Statist. Phys.*, 61(1-2) :203–234, 1990.
- [5] J. M. Ball, J. Carr, and O. Penrose. The Becker-Döring cluster equations : basic properties and asymptotic behaviour of solutions. *Comm. Math. Phys.*, 104(4) :657–692, 1986.
- [6] H T Banks and B.G. Fitzpatrick. Estimation of growth rate distributions in size structured population models. *Quart. Appl. Math.*, 49 :215–235, 2005.
- [7] H.T. Banks, F. Charles, M. Doumic, K.L. Sutton, and W.C. Thompson. Label structured cell proliferation models. *App. Math. Letters*, 23(12) :1412–1415, 2010.
- [8] John S Baras and Alain Bensoussan. On observer problems for systems governed by partial differential equations. Technical Report SCR TR 86-47, DTIC, 1987.
- [9] B. Basse, B.C. Baguley, E.S. Marshall, W.R. Joseph, B. van Brunt, G. Wake, and D. J. N. Wall. A mathematical model for analysis of the cell cycle in cell lines derived from human tumors. *J. Math. Biol.*, 47(4) :295–312, 2003.
- [10] Johann Baumeister and Antonio Leitão. *Topics in inverse problems*. Publicações Matemáticas do IMPA. [IMPA Mathematical Publication]. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 25º colóquio brasileiro de matemática. [25th brazilian mathematics colloquium] edition, 2005.
- [11] R. Becker and W. Döring. Kinetische behandlung der keimbildung in übersättigten dämpfen. *Annalen der Physik*, 416(8) :719–752, 1935.

- [12] George I. Bell. Cell growth and division : III. Conditions for Balanced Exponential Growth in a Mathematical Model. *Biophysical Journal*, 8(4) :431 – 444, 1968.
- [13] George I. Bell and Ernest C. Anderson. Cell growth and division : I. a mathematical model with applications to cell volume distributions in mammalian suspension cultures. *Biophysical Journal*, 7(4) :329 – 351, 1967.
- [14] R E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A., 1961.
- [15] A. Bensoussan. *Filtrage optimal des systèmes linéaires*. Dunod, 1971.
- [16] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter. *Representation and control of infinite-dimensional systems. Vol. II. Systems & Control : Foundations & Applications*. Birkhäuser Boston Inc., Boston, MA, 1993.
- [17] M.F. Bishop and F.A. Ferrone. Kinetics of nucleation-controlled polymerization. a perturbation treatment for use with a secondary pathway. *Biophysical Journal*, 46(5) :631 – 644, 1984.
- [18] T. Bourgeron, Z. Xu, M. Doumic, and M.T. Teixeira. The asymmetry of telomere replication contributes to replicative senescence heterogeneity. *Scientific Reports*, 5 :15326, 2015.
- [19] Thibault Bourgeron, Marie Doumic, and Miguel Escobedo. Estimating the division rate of the growth-fragmentation equation with a self-similar kernel. *Inverse Problems*, 30(2) :025007, 2014.
- [20] H. Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications].
- [21] H. Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, 2010.
- [22] J.-F. Collet, T. Goudon, F. Poupaud, and A. Vasseur. The Becker–Döring system and its Lifshitz–Slyozov limit. *SIAM J. on Appl. Math.*, 62(5) :1488–1500, 2002.
- [23] R. Danchin. Cours de topologie et d'analyse fonctionnelle master première année. <http://perso-math.univ-mlv.fr/users/danchin.raphael/cours/resume09.pdf>, 2013.
- [24] M. Doumic, T. Goudon, and T. Lepoutre. Scaling Limit of a Discrete Prion Dynamics Model. *Comm. in Math. Sc.*, 7(4) :839–865, 2009.
- [25] M. Doumic, M. Hoffmann, N. Krell, and L. Robert. Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli*, in press, 2014.
- [26] M. Doumic, B. Perthame, and J.P. Zubelli. Numerical solution of an inverse problem in size-structured population dynamics. *Inverse Problems*, 25(electronic version) :045008, 2009.

- [27] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Springer, 1996.
- [28] J.Z. Farkas. Stability conditions for a nonlinear size-structured model. *Nonlin. Anal. Real World App.*, 6 :962–969, 2005.
- [29] G.H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numer. Math.*, 14 :403–420, 1970.
- [30] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13 :49–52, 1902.
- [31] Ghislain Haine. Recovering the observable part of the initial data of an infinite-dimensional linear system with skew-adjoint generator. *Mathematics of Control Signals and Systems*, 26(3) :435–462, January 2014.
- [32] M. Hoffmann and A. Olivier. Nonparametric estimation of the division rate of an age dependent branching process. *ArXiv e-prints*, December 2014.
- [33] J. Keller. Inverse problems. *Amer. Math. Monthly*, 83 :107–118, 1976.
- [34] W.O. Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. Roy. Society of London, Series A*, 115(772) :700–721, 1927.
- [35] W.O. Kermack and A.G. McKendrick. Contribution to the mathematical theory of epidemics. ii. the problem of endemicity. *Proc. Roy. Society of London, Series A*, 138(834) :55–83, 1932.
- [36] T. P. J. Knowles, C. A. Waudby, G. L. Devlin, S. I. A. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, M. E. Welland, and C. M. Dobson. An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science*, 326(5959) :1533–1537, 2009.
- [37] H. E. Kubitschek. Growth during the bacterial cell cycle : Analysis of cell size distribution. *Biophysical Journal*, 9(6) :792–809, 1969.
- [38] P. Laurençot and S. Mischler. From the discrete to the continuous coagulation-fragmentation equations. *Proc. Roy. Soc. Edinburgh Sect. A*, 132(5) :1219–1248, 2002.
- [39] Francois-Xavier Le Dimet and O Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : Theoretical aspects. *Tellus A*, 38(2) :97–110, 1986.
- [40] I.M. Lifshitz and V.V. Slyozov. The kinetics of precipitation from supersaturated solid solutions. *Journal of Physics and Chemistry of Solids*, 19(1–2) :35 – 50, 1961.
- [41] J.-L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Avant propos de P. Lelong. Dunod, Paris, 1968.
- [42] David G Luenberger. An introduction to observers. *IEEE Transactions on Automatic Control*, 16 :596–602, 1971.

- [43] A.G. McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44 :98–130, 1926.
- [44] J. A. J. Metz and O. Diekmann, editors. *The dynamics of physiologically structured populations*, volume 68 of *Lecture Notes in Biomathematics*. Springer-Verlag, Berlin, 1986. Papers from the colloquium held in Amsterdam, 1983.
- [45] M. Nussbaum. Asymptotic equivalence of density estimation and white noise. *Ann. Statist.*, 24 :2399–2430, 1996.
- [46] M. Nussbaum and S. Pereverzev. The degrees of ill-posedness in stochastic and deterministic noise models. *Preprint WIAS 509*, 1999.
- [47] F. Oosawa and S. Asakura. *Thermodynamics of the polymerization of protein*. Academic Press, 1975.
- [48] B. Perthame. *Transport equations in biology*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2007.
- [49] S. Prigent, A. Ballesta, F. Charles, N. Lenuzza, P. Gabriel, L.M. Tine, H., and M. Doumic. An efficient kinetic model for assemblies of amyloid fibrils and its application to polyglutamine aggregation. *PLoS ONE*, 7(11) :e43273, 11 2012.
- [50] Stephanie Prigent, Wafaâ Haffaf, H. T. Banks, M. Hoffmann, Human Rezaei, and Marie Doumic. Size distribution of amyloid brils. Mathematical models and experimental data. *International Journal of Pure and Applied Mathematics*, 93(6) :845–878, June 2014.
- [51] Karim Ramdani, M Tucsnak, and G. Weiss. Recovering the initial state of an infinite-dimensional system using observers. *Automatica*, 46(10) :1616–1625, 2010.
- [52] S I Rubinow. A maturity-time representation for cell populations. *Biophys J.*, 8 :1055–1073, 1968.
- [53] E. Schock. Approximate solution of ill-posed equations : arbitrarily slow convergence vs. superconvergence. In G. Hämmerlin and K.H. Hoffmann, editors, *Constructive Methods for the Practical Treatment of Integral Equations*. Birkhäuser, 1985.
- [54] F R Sharpe and A J Lotka. A problem in age-distribution. *Philosophical Magazine*, 21 :435–438, 1911.
- [55] J.W. Sinko and W. Streifer. A new model for age-size structure of a population. *Ecology*, 48(6) :910–918, 1967.
- [56] J.W. Sinko and W. Streifer. A model for populations reproducing by fission. *Ecology*, 52(2) :330–335, 1971.
- [57] E. Stewart, R. Madden, G. Paul, and F. Taddei. Aging and death in an organism that reproduces by morphologically symmetric division. *Curr. Biol.*, 20(12) :1099–103, 2010.

- [58] Sattar Taheri-Araghi, Serena Bradde, John T. Sauls, Norbert S. Hill, Petra Anne Levin, Johan Paulsson, Massimo Vergassola, and Suckjoon Jun. Cell-size control and homeostasis in bacteria. *Current Biology*, 11679(1-7), 2015.
- [59] E. Trélat. *Contrôle optimal : théorie & applications*. Mathématiques Concrètes. Vuibert, 2005.
- [60] A. Tsybakov. Apprentissage statistique et estimation non-paramétrique. 2012.
- [61] C. Villani. Cours de deuxième année donné à l'école normale supérieure de Lyon. <http://perso-math.univ-mly.fr/users/danchin.raphael/cours/resume09.pdf>, 2013.
- [62] P Wang, L Robert, J Pelletier, W L Dang, F Taddei, A Wright, and S Jun. Robust growth of Escherichia coli. *Curr. Biol.*, 20(12) :1099–103, 2010.
- [63] M. Wulkow. The simulation of molecular weight distributions in polyreaction kinetics by discrete galerkin methods. *Macromol. Theory Simul.*, 5 :396–416, 1996.
- [64] W-F Xue, S W Homans, and S E Radford. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *PNAS*, 105 :8926–8931, 2008.