Méthodes de splitting pour les problèmes multi-échelles

B. Bidégaray-Fesquet

Cours de M2R — 2006–2007

1 Introduction

«Splitting» : un terme anglo-saxon, quelle horreur ! Oui, mais je vais l'utiliser tout de même.

En français, ce terme pourraît être traduit par fission, fragmentation, division, fractionnement, séparation, décomposition. Aucun de ces termes n'est vraiment satisfaisant.

On demande souvent de traduire par «pas fractionnaires», ce qui de mon point de vue constitue un contresens, car les pas de temps sont souvent entiers, c'est l'opérateur qui est fractionné.

On utilise parfois «directions alternées», ce qui est une application particulière dont nous parlerons, mais qui est loin de représenter l'ensemble des applications de cette méthode.

1.1 Une équation scalaire sans échelles

Considérons l'équation différentielle ordinaire (EDO) scalaire suivante :

$$\text{(EDO) } \dot{x} = (a+b)x, \qquad x(0) = x^0,$$

où a et b sont des scalaires. On connaît la solution exacte de cette équation :

$$\begin{aligned} x(t) = \exp((a+b)t)x^0 &= \exp(at)\exp(bt)x^0 \qquad \text{(méthode 1)} \\ &= \exp(bt)\exp(at)x^0 \qquad \text{(méthode 2).} \end{aligned}$$

Nous pouvons ainsi séparer l'évolution selon l'équation (EDO) en deux temps :

$$\text{(L1)} \left| \begin{array}{ll} \dot{y} = by, & y(0) = x^0, \\ \dot{x} = ax, & x(0) = y(t), \end{array} \right. \qquad \text{(L2)} \left| \begin{array}{ll} \dot{y} = ay, & y(0) = x^0, \\ \dot{x} = bx, & x(0) = y(t). \end{array} \right.$$

Pour le système (L1), on a clairement

$$x(t) = \exp(at)x(0) = \exp(at)y(t) = \exp(at)\exp(bt)y(0) = \exp(at)\exp(bt)x^0.$$

Le calcul pour (L2) se fait de la même manière et donne le même résultat. On appelle **splitting de Lie** les deux méthodes (L1) et (L2). Pour une équation scalaire, ces deux méthodes sont identiques et reviennent au même que de traiter l'équation en une seule fois.

1.2 Quand le splitting présente un intérêt

Jusqu'à maintenant le splitting n'a pas l'air de présenter un intérêt. En fait l'exemple donné est à peu près le seul pour lequel le splitting ne change pas la solution. Nous allons voir l'intérêt et l'impact du splitting dans de nombreux contextes théoriques

- les systèmes différentiels linéaires : $\dot{x} = (A+B)x$,
- les systèmes différentiels linéaires avec deux échelles différentes : $\dot{x} = (\frac{1}{\varepsilon}A + B)x$,
- les équations aux dérivées partielles non linéaires : $i\partial_t u = \Delta u + f(u)$,

et leur approximation numérique.

Nous allons nous intéresser uniquement à des problèmes académiques. Le splitting est néanmoins utilisé dans de nombreux contextes applicatifs. Nous pouvons citer notamment :

– la chimie complexe (traitement séparé des phénomènes de réaction chimique (équations non linéaires) et de diffusion des espèces. Ce sont des systèmes avec un très grand nombre de variables et des échelles de temps très différentes.
– la météorologie, l'océanographie. Là encore la multiplicité des phénomènes mis en jeu, donne lieu à une multiplicité de termes de natures très différentes.
– la décomposition de domaine. Celle-ci est utilisée lorsqu'il y a couplage de phénomènes sur des domaines différents (adjacents) ou avec des géométries très différentes. Dans la zone d'interaction, il y a redondances des variables et on peut souvent séparer l'opérateur d'évolution en un opérateur facile à intégrer et un autre petit en un certaine sens.
– les méthodes d'ondelettes. À nouveau, les termes diagonaux des opérateurs sont prépondérants et faciles à intégrer et les termes extra-diagonaux sont moralement petit, c'est d'ailleurs là tout l'intérêt de la méthode.
– et bien d'autres ...

Deux intérêts principaux du splitting peuvent d'ores et déjà être identifiés :
– le fait de pouvoir résoudre exactement ou numériquement chacune des sous-équations alors que cela est impossible ou difficile avec l'équation entière,
– le fait de pouvoir traiter séparément des variables ou des opérateurs correspondant à des échelles très différentes.

Un inconvénient apparaît aussi immédiatement comme l'impossibilité de conserver les propriétés fines liées à la structure de certaines équations et mettant en œuvre toute l'équation (quantités conservées).

Dans tout ce cours, nous nous intéressons exclusivement aux semi-discrétisations en temps. La discrétisation en espace peut faire appel à n'importe quelle technique adaptée (différences finies, éléments finis, volumes finis, méthodes spectrales, ...), et éventuellement différentes pour chacune des parties de l'équation. C'est là tout l'intérêt de la chose.

2 Semi-groupes d'évolution

2.1 Une réécriture des schémas de splitting

Pour exprimer le splitting dans notre exemple simpliste, nous avons été obligés d'introduire une variable intermédiaire y. Ceci s'avérerait à l'usage peu pratique pour des contextes plus compliqués. Nous introduisons donc une notation de type semi-groupe d'évolution. L'application qui à x^0 associe $x(t)$ par le flot de l'EDO est le semi-groupe d'évolution que nous noterons $\mathcal{S}(t)$ pour l'équation toute entière :

$$x(t) = \mathcal{S}(t)x^0 = \exp((a+b)t)x^0.$$

Quand l'opérateur est linéaire, on garde aussi souvent la notation avec l'exponentielle. Si on note $\mathcal{A}(t)$ et $\mathcal{B}(t)$, les semi-groupes d'évolution associés aux deux parties de l'équation, les deux splittings (L1) et (L2) consistent à écrire

$$\text{(L1) } x(t) = \mathcal{A}(t)\mathcal{B}(t)x^0, \qquad \text{(L2) } x(t) = \mathcal{B}(t)\mathcal{A}(t)x^0.$$

Avec cette notation, on définit sans effort (et sans introduction de variables supplémentaires) les **splittings de Strang**

$$\text{(S1) } x(t) = \mathcal{A}(\frac{t}{2})\mathcal{B}(t)\mathcal{A}(\frac{t}{2})x^0, \qquad \text{(S2) } x(t) = \mathcal{B}(\frac{t}{2})\mathcal{A}(t)\mathcal{B}(\frac{t}{2})x^0.$$

Évidemment, pour notre premier exemple ces deux splittings sont toujours équivalents à l'équation initiale.

Qu'entend-t-on par **semi-groupe d'évolution** ? L'opérateur $\mathcal{A}(t)$ qui est défini de $\mathbb{R}$ dans $\mathbb{R}$ a les deux propriétés suivantes :

(P1) $\mathcal{A}(0) = \mathrm{I}$, $\quad (cf.\ \exp(0) = 1)$,

(P2) $\mathcal{A}(t+s) = \mathcal{A}(t)\mathcal{A}(s)$, pour tout $t, s \geq 0$ $\quad (cf.\ \exp(a(t+s)) = \exp(at)\exp(as))$.

La propriété (P1) dit que l'évolution pendant un temps nul donne la donnée initiale. La propriété (P2) dit que cela revient au même d'évoluer pendant un temps $t+s$, ou d'évoluer pendant un temps s puis un temps t. Ceci est vrai parce que le système est autonome, sinon cela est faux. Dans notre exemple, on a même un groupe, car $\mathcal{A}(t)$ est défini pour des temps t négatifs, mais cela n'est pas toujours le cas. Certaines équations sont mal posées «en rétrograde».

Les propriétés de semi-groupe peuvent être énoncées dans un cadre beaucoup plus général.

2.2 Semi-groupes uniformément continus d'opérateurs linéaires bornés

Soit X un espace de Banach. Une famille à un paramètre $\mathcal{A}(t) : X \to X$, d'opérateurs linéaires bornés, est un **semi-groupe d'opérateurs linéaires bornés** sur X si les propriétés (P1) et (P2) sont vérifiées. Cette famille est **uniformément continue** si de plus

(P3) $\lim_{t \downarrow 0} \|\mathcal{A}(t) - \mathrm{I}\| = 0.$

On définit l'opérateur linéaire A par son domaine

$$D(A) = \left\{ x \in X;\ \lim_{t \downarrow 0} \frac{\mathcal{A}(t)x - x}{t} \text{ existe } \right\}$$

et

$$Ax = \lim_{t \downarrow 0} \frac{\mathcal{A}(t)x - x}{t} = \left. \frac{d_+ A(t)x}{dt} \right|_{t=0} \text{ pour } x \in D(A).$$

Cet opérateur linéaire est le **générateur infinitésimal** du semi-groupe $\mathcal{A}(t)$. Dans cette terminologie se trouve sous-jacent le fait qu'un seul semi-groupe permet de construire A comme ci-dessus. Le lien entre A et $\mathcal{A}$ est bijectif. Si A est un opérateur borné sur X, on peut écrire

$$\mathcal{A}(t) = \exp(tA) = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!}$$

et la série converge bien en norme pour tout $t \geq 0$.

Théorème 1

Il y a équivalence entre

(i) A est un opérateur linéaire borné,

(ii) $\mathcal{A}$ est uniformément continu.

Le semi-groupe $\mathcal{A}(t)$ commute avec son générateur A.

Théorème 2

L'application $t \mapsto \mathcal{A}(t)$ est différentiable en norme et

$$\frac{d\mathcal{A}(t)}{dt} = A\mathcal{A}(t) = \mathcal{A}(t)A.$$

Dans notre exemple, la propriété (P3) s'écrit

$$\lim_{t \downarrow 0} |\exp(at) - 1| = 0.$$

Le domaine de A est

$$D(A) = \left\{ x \in \mathbb{R};\ \lim_{t \downarrow 0} \frac{\exp(at)x - x}{t} \text{ existe } \right\} = \mathbb{R}$$

et

$$Ax = \lim_{t \downarrow 0} \frac{\exp(at)x - x}{t} = ax.$$

Nous allons commencer par étudier des opérateurs linéaires dans des espaces de dimension finie (matrices) et ceci ne nécessite pas un arsenal fonctionnel théorique très compliqué. Nous compléterons donc cette présentation lorsque cela sera nécessaire, en donnant un cadre théorique pour lequel nous savons dire des choses en dimension infinie. Nous généraliserons également ces notions au cas non linéaire.

3 Le cas des systèmes linéaires

On considère maintenant le système linéaire

$$\text{(S)}\ \dot{x} = (A+B)x, \qquad\qquad x(0) = x^0,$$

où $x \in \mathbb{R}^d$ et A et B sont des matrices de $\mathcal{M}_d$. On est clairement dans le cadre général précédemment décrit. La matrice A est la représentation de l'opérateur linéaire A dans la base canonique. L'opérateur $\mathcal{A}(t)$ est représenté par la matrice $\exp(tA)$.

Si les matrices A et B commutent, on a, par exemple, $\exp(tA)\exp(tB) = \exp(t(A+B))$, ce qui limite l'intérêt du splitting. On s'intéressera donc au cas où A et B ne commutent pas et on définira le **commutateur** ou encore **crochet de Lie** par

$$[A, B] = AB - BA.$$

3.1 Splitting de Lie

Pour l'instant les matrices A et B jouent des rôles symétriques. On ne regarde qu'un seul cas de splitting de Lie.

$$\begin{aligned}
\mathcal{A}(t)\mathcal{B}(t) - \mathcal{S}(t) &= \left(\mathrm{I} + tA + \frac{t^2}{2}A^2\right)\left(\mathrm{I} + tB + \frac{t^2}{2}B^2\right) \\
&\quad - \left(\mathrm{I} + t(A+B) + \frac{t^2}{2}(A+B)^2\right) + O(t^3) \\
&= t^2\left(\frac{1}{2}A^2 + AB + \frac{1}{2}B^2\right) - \frac{t^2}{2}\left(A^2 + AB + BA + B^2\right) + O(t^3) \\
&= \frac{t^2}{2}[A, B] + O(t^3).
\end{aligned}$$

Ce développement permet de démontrer l'ordre de la méthode. Nous ferons la preuve plus loin dans un cadre plus général.

3.2 Splitting de Strang

Étudions de même la formule de Strang

$$\begin{aligned}
\mathcal{A}(\frac{t}{2})\mathcal{B}(t)\mathcal{A}(\frac{t}{2}) - \mathcal{S}(t) &= \left(\mathrm{I} + \frac{t}{2}A + \frac{t^2}{8}A^2 + \frac{t^3}{48}A^3\right)\left(\mathrm{I} + tB + \frac{t^2}{2}B^2 + \frac{t^3}{6}B^3\right) \\
&\quad \times \left(\mathrm{I} + \frac{t}{2}A + \frac{t^2}{8}A^2 + \frac{t^3}{48}A^3\right) \\
&\quad - \left(\mathrm{I} + t(A+B) + \frac{t^2}{2}(A+B)^2 + \frac{t^3}{6}(A+B)^3\right) + O(t^4)
\end{aligned}$$

$$\begin{aligned}\mathcal{A}(\frac{t}{2})\mathcal{B}(t)\mathcal{A}(\frac{t}{2}) - \mathcal{S}(t) &= t^3\left(\frac{1}{6}A^3 + \frac{1}{8}A^2B + \frac{1}{4}ABA + \frac{1}{8}BA^2 + \frac{1}{4}B^2A + \frac{1}{4}AB^2 + \frac{1}{6}B^3\right)\\ &\quad - \frac{t^3}{6}\left(A^3 + A^2B + ABA + BA^2 + B^2A + BAB + AB^2 + B^3\right) + O(t^4)\\ &= t^3\left(-\frac{1}{24}A^2B + \frac{1}{12}ABA - \frac{1}{24}BA^2 + \frac{1}{12}B^2A - \frac{1}{6}BAB + \frac{1}{12}AB^2\right)\\ &\quad + O(t^4)\\ &= t^3\left(-\frac{1}{24}[A,[A,B]] + \frac{1}{12}[B,[B,A]]\right) + O(t^4).\end{aligned}$$

3.3 Splittings d'ordre plus élevé

3.3.1 Calcul de l'ordre et convergence

Les deux résultats précédents sont deux cas particuliers du résultat suivant.

Théorème 3

Soit $C \in \mathcal{M}_d$ et f une fonction continue définie sur un voisinage de 0 dans $\mathbb{R}$ à valeurs dans $\mathcal{M}_d$, tels qu'il existe une matrice $R \in \mathcal{M}_d$ et un entier p tels que le développement asymptotique

$$(DA)\ f(t) - \exp(tC) = Rt^{p+1} + O(t^{p+2})$$

soit vrai dans un voisinage de 0. Alors

$$f(\frac{t}{n})^n - \exp(tC) = O((\frac{t}{n})^p).$$

De plus, cette estimation est optimale, sauf si R est identiquement nulle.

Preuve :
On pose $h = t/n$. On a $x^m = f(h)^m x^0$ et $x(mh) = \exp(mhC)x(0)$ et on prend bien sûr $x^0 = x(0)$. L'**erreur locale** est l'erreur de méthode et est définie par

$$\eta^{m+1} = (\exp(hC) - f(h))x(mh).$$

L'**erreur globale** est définie par

$$e^m = x(mh) - x^m.$$

On peut écrire la relation de récurrence

$$\begin{aligned}e^{m+1} &= x((m+1)h) - x^{m+1}\\ &= \exp(hC)x(mh) - f(h)x^m\\ &= f(h)(x(mh) - x^m) + (\exp(hC) - f(h))x(mh)\\ &= f(h)e^m + \eta^{m+1}.\end{aligned}$$

Comme $e^0 = 0$ on a de façon classique

$$\begin{aligned}e^n &= \sum_{k=0}^{n-1} f(h)^k \eta^{n-k},\\ |e^n| &\leq \sum_{k=0}^{n-1} \|f(h)^k\| \|Rh^{p+1} + O(h^{p+2})\| \max_{s\in[0,T]} |x(s)|\end{aligned}$$

On se place dans le cadre de méthodes stables. Pour des temps en $O(1)$, on a donc $\|f(h)^k\| = O(1)$ et $\max_{s\in[0,T]} |x(s)| = O(1)$. Ainsi

$$|e^n| \leq Cn(h^{p+1} + O(h^{p+2})) = tC(h^p + O(h^{p+1})).$$

Enfin $e^n = (\exp(tC) - f(t/n)^n)x^0$, ce qui prouve le résultat. ■

Ce théorème assure que n applications de $f(t/n)$ à x^0 fournit ainsi une méthode d'ordre p pour approcher $x(t)$. Les splittings de Lie et de Strang correspondent aux cas $p = 1$ et $p = 2$ respectivement. En effet

$$\left\| \left(\mathcal{A}(\frac{t}{n})\mathcal{B}(\frac{t}{n}) \right)^n x^0 - x(t) \right\| = O(\frac{t}{n})$$

(cette formule est une version dans le cadre matriciel de la formule de Trotter–Kato que nous verrons plus loin) et

$$\left\| \left(\mathcal{A}(\frac{t}{2n})\mathcal{B}(\frac{t}{n})\mathcal{A}(\frac{t}{2n}) \right)^n x^0 - x(t) \right\| = O((\frac{t}{n})^2).$$

Les méthodes de Lie et de Strang sont respectivement d'ordre 1 et 2.

Ceci justifie l'utilisation du splitting comme méthode numérique, en admettant que l'on sache calculer de manière exacte ou suffisamment précise chacune des exponentielles $\exp(tA/n)$ et $\exp(tB/n)$.

L'ordre 1 et *a fortiori* les ordres supérieurs suffisent à montrer la consistance des méthodes. La stabilité sera assurée systématiquement par la stabilité de chacun des schémas pour résoudre les différentes parties du splitting. Il est bien connu (théorème de Lax-Richtmyer) que consistance et stabilité assurent la convergence de la méthode.

Une question se pose : comment construire des méthodes d'ordre plus élevé ?

3.3.2 Une réponse négative

La première réponse est négative dans le cas où on cherche des coefficients positifs. On cherche des coefficients α_j et β_j tels que la fonction

$$f_k(t, \alpha, \beta) = \exp(\alpha_1 tA) \exp(\beta_1 tB) \dots \exp(\alpha_k tA) \exp(\beta_k tB)$$

vérifie le développement asymptotique du théorème précédent.

Théorème 4

Si $[A,[A,B]]$ et $[B,[B,A]]$ sont linéairement indépendants, il n'existe aucun choix de k et des coefficients α_j et β_j réels positifs pour obtenir (DA) pour $p \geq 3$.

On remarque que les commutateurs qui interviennent sont ceux du reste dans la méthode d'ordre 2 de Strang.

Le résultat est même plus fort, on peut généraliser la forme de la fonction recherchée en

$$f(t) = r_1(tA)s_1(tB) \dots r_k(tA)s_k(tB),$$

où les fonctions r_j et s_j sont analytiques sur $\mathbb{R}$ avec $r_j(0) = s_j(0) = 1$, $r_j'(0) = \alpha_j$ et $s_j'(0) = \beta_j$. Cette généralisation inclut la plupart des approximations numériques de l'exponentielle.

Théorème 5

Si

(i) α_j et β_j sont positifs,

(ii) $[A,B]$, A^2 et B^2 sont linéairement indépendants,

(iii) $[A,[A,B]]$, $[B,[B,A]]$, $[A^2,B]$, $[A,B^2]$, A^3 et B^3 sont linéairement indépendants,

alors il n'existe aucun choix de k et des fonctions r_j et s_j pour obtenir (DA) pour $p \geq 3$.

3.3.3 Formule produit avec des coefficients négatifs

On peut chercher une formule produit d'exponentielles comme ci-dessus en annulant les coefficients jusqu'à l'ordre 3 inclus. Ceci donne par exemple la formule

$$f(t) = \exp(tA)\exp(-\frac{1}{24}tB)\exp(-\frac{2}{3}tA)\exp(\frac{3}{4}tB)\exp(\frac{2}{3}tA)\exp(\frac{7}{24}tB)$$

pour laquelle

$$f(t) - \exp(t(A+B)) = t^4\left(\frac{1}{216}[A,[A,[A,B]]] + \frac{1}{72}[A,[B,[A,B]]] - \frac{1}{2304}[B,[B,[A,B]]]\right) + O(t^5).$$

Pour A et B matrices, ceci ne pose aucun problème. Dans un cadre plus général, où A et B sont des opérateurs linéaires, il faut s'assurer que les semi-groupes d'évolution associés sont de plus des groupes, c'est-à-dire que les exponentielles sont définies pour des temps négatifs. Autrement dit, les équations doivent être bien posées en rétrograde. Dans la formule ci-dessus, ceci doit être valable à la fois pour A et B.

3.3.4 Combinaison d'approximations d'ordre inférieur

On se rappelle que

$$\mathcal{A}(t)\mathcal{B}(t) - \mathcal{S}(t) = \frac{t^2}{2}[A,B] + O(t^3).$$

On a donc

$$\frac{1}{2}(\mathcal{A}(t)\mathcal{B}(t) + \mathcal{B}(t)\mathcal{A}(t)) - \mathcal{S}(t) = O(t^3).$$

En poussant plus loin les développements, on voit apparaître des termes faisant intervenir les commutateurs $[A,[A,B]]$ et $[B,[B,A]]$. On peut essayer de les annuler avec ceux de $\mathcal{A}(t/2)\mathcal{B}(t)\mathcal{A}(t/2) - \mathcal{S}(t)$ et $\mathcal{B}(t/2)\mathcal{A}(t)\mathcal{B}(t/2) - \mathcal{S}(t)$. Ceci donne la formule

$$\begin{aligned} g(t) &= \frac{2}{3}\left(\exp(\frac{1}{2}tA)\exp(tB)\exp(\frac{1}{2}tA) + \exp(\frac{1}{2}tB)\exp(tA)\exp(\frac{1}{2}tB)\right) \\ &\quad -\frac{1}{6}\left(\exp(tA)\exp(tB) + \exp(tB)\exp(tA)\right), \end{aligned}$$

pour laquelle

$$g(t) - \exp(t(A+B)) = -\frac{t^4}{24}[A,B]^2 + O(t^5).$$

3.3.5 Extrapolations de Richardson

L'extrapolation de Richardson est une méthode générale pour accélérer la convergence d'une méthode numérique d'intégration des EDO. Présentons la d'abord dans ce cadre.

On suppose que l'on approche la solution exacte $y(t)$ d'une EDO par une méthode dépendant d'un pas h et calculant une approximation $y(t;h)$ telle que

$$y(t;h) = y(t) + h^p g(t) + O(h^{p+1})$$

qui est donc d'ordre local p. Si au lieu d'utiliser le pas de temps h, on utilise le pas de temps qh, on a

$$y(t;qh) = y(t) + (qh)^p g(t) + O(h^{p+1}).$$

On peut calculer la combinaison linéaire

$$\frac{q^p y(t;h) - y(t;qh)}{q^p - 1} = y(t) + O(h^{p+1})$$

qui donne une nouvelle méthode qui est d'ordre local au moins $p+1$.

Adaptons ceci aux méthodes de splitting. La solution exacte est $\mathcal{S}(t)x_0$. Si on calcule avec une seule itération de splitting, on calcule $f(t)x_0$ qui est l'équivalent de $y(t;h)$. On choisit $q = 1/2$, ce qui revient à faire deux itérations de la méthodes avec un pas de temps moitié : $f(t/2)f(t/2)x_0$. La nouvelle méthode s'écrit :

$$\frac{\frac{1}{2^p}f(t) - f(\frac{t}{2})f(\frac{t}{2})}{\frac{1}{2^p}-1} = \frac{2^p f(\frac{t}{2})f(\frac{t}{2}) - f(t)}{2^p - 1}.$$

Appliquons ceci à un splitting de Lie $\mathcal{L}_1(t) = \mathcal{A}(t)\mathcal{B}(t)$ pour lequel $p = 1$. On obtient le schéma

$$\mathcal{RL}_1(t) = 2\mathcal{A}(\frac{t}{2})\mathcal{B}(\frac{t}{2})\mathcal{A}(\frac{t}{2})\mathcal{B}(\frac{t}{2}) - \mathcal{A}(t)\mathcal{B}(t)$$

qui n'est que d'ordre 2, donc pas meilleur qu'un splitting de Strang tout en demandant plus de calculs. En outre ce schéma a l'inconvénient de comporter des signes moins. On oublie donc.

Appliquons ceci à un splitting de Strang $\mathcal{S}_1(t) = \mathcal{A}(t/2)\mathcal{B}(t)\mathcal{A}(t/2)$ pour lequel $p = 2$. On obtient le schéma

$$\mathcal{RS}_1(t) = \frac{4\mathcal{A}(t/4)\mathcal{B}(\frac{t}{2})\mathcal{A}(t/4)\mathcal{A}(t/4)\mathcal{B}(\frac{t}{2})\mathcal{A}(t/4) - \mathcal{A}(\frac{t}{2})\mathcal{B}(t)\mathcal{A}(\frac{t}{2})}{3}.$$

On peut combiner les deux $\mathcal{A}(t/4)$ centraux et obtient le nouveau schéma

$$\begin{aligned} g(t) &= \frac{4}{3}\exp(\frac{1}{4}tA)\exp(\frac{1}{2}tB)\exp(\frac{1}{2}tA)\exp(\frac{1}{2}tB)\exp(\frac{1}{4}tA) \\ &\quad -\frac{1}{3}\exp(\frac{1}{2}tA)\exp(tB)\exp(\frac{1}{2}tA). \end{aligned}$$

Cette formule est construite à partir de deux formules d'ordre 2 pour être d'ordre 3. En fait, elle est d'ordre 4 et est utilisée en pratique contrairement à la précédente.

3.3.6 Facteurs intégraux

On commence par réécrire le système d'origine :

$$\begin{aligned} \dot{x}(t) &= (A+B)x(t), \\ \exp(-tA)\dot{x}(t) &= \exp(-tA)(A+B)x(t), \\ -A\exp(-tA)x(t) + \exp(-tA)\dot{x}(t) &= \exp(-tA)Bx(t), \\ \frac{d}{dt}\left(\exp(-tA)x(t)\right) &= \exp(-tA)Bx(t). \end{aligned}$$

Il s'agit alors de remplacer $\exp(tA)$ par une approximation, par exemple une méthode de Runge–Kutta implicite ou semi-implicite. Par construction, l'ordre obtenu est formellement celui de la méthode de Runge-Kutta.

Nous verrons plus loin la méthode de splitting de source, qui n'a pas pour but d'obtenir des ordres très élevés et n'a donc pas sa place ici. Celle-ci donne cependant des rôles très différents à A et à B comme pour la méthode des facteurs intégraux.

4 Approximation de l'exponentielle

Tous les calculs précédents ne sont applicables en pratique que pour des matrices dont on sait calculer facilement l'exponentielle. Elles ne sont pas si nombreuses. Entrent dans ce cadre

– les matrices diagonales :
$D = \mathrm{diag}(d_k)$. On a alors $\exp(tD) = \mathrm{diag}(\exp(td_k))$.

– les matrices sous forme de Jordan :

$$J = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & 0 & & \lambda \end{pmatrix}.$$

Il est tout d'abord facile de calculer leurs puissances :

$$J^k = \begin{pmatrix} \lambda^k & k\lambda^{k-1} & \dots & C_k^{k-(d-1)}\lambda^{k-(d-1)} \\ & \ddots & \ddots & \\ & & \ddots & k\lambda^{k-1} \\ & 0 & & \lambda^k \end{pmatrix}.$$

De manière générique, on a pour $j \geq i$, $J^k_{ij} = C_k^{k-(j-i)}\lambda^{k-(j-i)}$. On a alors

$$\begin{aligned} \exp(tJ)_{ij} &= \sum_{k=0}^{\infty} \frac{1}{k!} t^k J^k_{ij} = \sum_{k=j-i}^{\infty} \frac{1}{k!} t^k C_k^{k-(j-i)} \lambda^{k-(j-i)} \\ &= \sum_{k=0}^{\infty} \frac{1}{(k+(j-i))!} t^k t^{j-i} C^k_{k+(j-i)} \lambda^k = \sum_{k=0}^{\infty} \frac{1}{k!(j-i)!} t^k t^{j-i} \lambda^k \\ \exp(tJ)_{ij} &= \frac{t^{j-i}}{(j-i)!} \sum_{k=0}^{\infty} \frac{1}{k!} (t\lambda)^k = \frac{t^{j-i}}{(j-i)!} \exp(t\lambda). \end{aligned}$$

– les matrices idempotentes comme les projecteurs :
$P^n = P$ pour tout $n \geq 1$. On a alors

$$\exp(tP) = \sum_{n=0}^{\infty} \frac{(tP)^n}{n!} = \mathrm{I} + \sum_{n=1}^{\infty} \frac{t^n P}{n!} = \mathrm{I} - P + \sum_{n=0}^{\infty} \frac{t^n}{n!} P = \mathrm{I} + (\exp(t) - 1)P.$$

Dans les autres cas, il faut avoir recours à des approximations à différents ordres de l'exponentielle.

On suppose que l'on discrétise le temps avec un pas constant h : $t_n = nh$. On cherche à calculer une approximation x^n de $x(t_n)$. Il s'agit d'écrire une relation de récurrence qui relie x^{n+1} à x^n. Celle-ci résulte de la discrétisation du système différentiel écrit entre les temps t_n et t_{n+1}.

$$\dot{x} = (A + B)x, \qquad x(t_n) = x^n,$$

ou de son équivalent intégral

$$x(t) = x(t_n) + \int_{t_n}^{t} (a + b)x(\tau)d\tau.$$

Pour pouvoir analyser ce qui suit, donnons le développement limité de x autour de $t = t_n$:

$$x(t_n + h) = (\mathrm{I} + (A + B)h + \frac{1}{2}(A + B)^2h^2 + \frac{1}{6}(A + B)^3h^3 + O(h^4))x(t_n).$$

On a donc

$$\begin{aligned} \mathcal{S}(h) &= (\mathrm{I} + (A + B)h + \frac{1}{2}(A^2 + AB + BA + B^2)h^2 \\ &\quad + \frac{1}{6}(A^3 + A^2B + ABA + AB^2 + BA^2 + BAB + B^2A + B^3)h^3 + O(h^4). \end{aligned}$$

Il suffira de calculer la différence de cette quantité avec celles associées aux différents schémas, qui sont toutes des fonctions continues de h au voisinage de 0 et à valeur dans pour obtenir une expression de type

$$\text{(DA)}\ f(h) - \mathcal{S}(h) = Rh^{p+1} + O(h^{p+2})$$

ce qui est l'opérateur d'erreur locale et le théorème assure que pour $t = nh$

$$f(h)^n - \mathcal{S}(nh) = O(h^p)$$

et la méthode est d'ordre p.

4.1 Schéma d'Euler

Pour l'équation (S) de départ, le schéma d'Euler consiste à approcher l'intégrale par la méthode des rectangles à gauche et s'écrit

$$x^{n+1} = \mathcal{E}(A+B,h)x^n = x^n + h(A+B)x^n = (\mathrm{I} + h(A+B))x^n.$$

Si on applique le schéma d'Euler à (L1), on a

$$\begin{aligned} \mathcal{E}(A,h)x^n &= x^n + hBx^n = (\mathrm{I}+hB)x^n, \\ x^{n+1} &= \mathcal{E}(A,h)\mathcal{E}(B,h)x^n = \mathcal{E}(B,h)x^n + hA\mathcal{E}(B,h)x^n = (\mathrm{I}+hA)\mathcal{E}(B,h)x^n \\ &= (\mathrm{I}+hA)(\mathrm{I}+hB)x^n = (\mathrm{I}+h(A+B)+h^2AB)x^n. \end{aligned}$$

On a

$$\mathcal{L}_1\mathcal{E}(h) = \mathrm{I} + h(A+B) + h^2AB$$

et de manière analogue

$$\mathcal{L}_2\mathcal{E}(h) = \mathrm{I} + h(A+B) + h^2BA.$$

Dans le cadre commutatif, les deux splittings sont équivalents. Du point de vue de l'ordre, on a

$$\begin{aligned} \mathcal{L}_1\mathcal{E}(h) - \mathcal{S}(h) &= \frac{1}{2}h^2([A,B] - (A^2+B^2)) + O(h^3), \\ \mathcal{L}_2\mathcal{E}(h) - \mathcal{S}(h) &= \frac{1}{2}h^2(-[A,B] - (A^2+B^2)) + O(h^3). \end{aligned}$$

Les schémas de splitting avec approximation d'Euler sont d'ordre 1. On peut penser aussi voir l'effet du splitting par rapport à la méthode d'Euler appliquée au système complet

$$\begin{aligned} \mathcal{L}_1\mathcal{E}(h) - \mathcal{L}_1(h) &= -\frac{1}{2}h^2(A^2+B^2) + O(h^3), \\ \mathcal{L}_2\mathcal{E}(h) - \mathcal{L}_2(h) &= -\frac{1}{2}h^2(A^2+B^2) + O(h^3). \end{aligned}$$

La dérive est la même pour les deux types de splitting. On est bien sûr encore d'ordre 1.

4.2 Schéma d'Euler rétrograde

Pour l'équation (S) de départ, le schéma d'Euler rétrograde consiste à approcher l'intégrale par la méthode des rectangles à droite et s'écrit

$$\begin{aligned} x^{n+1} &= \mathcal{R}(A+B,h)x^n = x^n + h(A+B)\mathcal{R}(A+B,h)x^n, \\ \Rightarrow x^{n+1} &= (\mathrm{I} - h(A+B))^{-1}x^n. \end{aligned}$$

Il suffit de prendre h suffisamment petit pour que ceci ait un sens : $h \leq 1/\rho(A+B)$.

On peut effectuer un développement limité et

$$\mathcal{R}(A+B,h) = \mathrm{I} + h(A+B) + h^2(A+B)^2 + O(h^3).$$

Si on applique le schéma d'Euler rétrograde à (L1), on a

$$\begin{aligned} \mathcal{R}(B,h)x^n &= (\mathrm{I}-hB)^{-1}x^n, \\ x^{n+1} &= \mathcal{R}(A,h)\mathcal{R}(B,h)x^n = (\mathrm{I}-hA)^{-1}(\mathrm{I}-hB)^{-1}x^n. \end{aligned}$$

À nouveau ceci est symétrique en A et B dans le cas commutatif. Sinon, on a deux opérateurs

$$\begin{aligned} \mathcal{L}_1\mathcal{R}(h) &= (\mathrm{I}-hA)^{-1}(\mathrm{I}-hB)^{-1} = \mathrm{I} + h(A+B) + h^2(A^2+AB+B^2) + O(h^3), \\ \mathcal{L}_2\mathcal{R}(h) &= (\mathrm{I}-hB)^{-1}(\mathrm{I}-hA)^{-1} = \mathrm{I} + h(A+B) + h^2(A^2+BA+B^2) + O(h^3). \end{aligned}$$

Du point de vue de l'ordre, on a

$$\begin{aligned}\mathcal{L}_1\mathcal{R}(h) - \mathcal{S}(h) &= \frac{1}{2}h^2([A,B] + (A^2+B^2)) + O(h^3),\\ \mathcal{L}_2\mathcal{R}(h) - \mathcal{S}(h) &= \frac{1}{2}h^2(-[A,B] + (A^2+B^2)) + O(h^3).\end{aligned}$$

Les schémas de splitting avec approximation d'Euler rétrograde sont d'ordre 1. On peut penser aussi voir l'effet du splitting par rapport à la méthode d'Euler appliquée au système complet

$$\begin{aligned}\mathcal{L}_1\mathcal{R}(h) - \mathcal{L}_1(h) &= \frac{1}{2}h^2(A^2+B^2) + O(h^3),\\ \mathcal{L}_2\mathcal{R}(h) - \mathcal{L}_2(h) &= \frac{1}{2}h^2(A^2+B^2) + O(h^3).\end{aligned}$$

La dérive est à nouveau la même pour les deux types de splitting et l'ordre est 1.

4.3 Couplage Euler–Euler rétrograde

Une fois le splitting effectué, rien n'oblige à utiliser le même schéma pour les deux parties. Avec les schémas déjà définis, on a quatre solutions :

$$\begin{aligned}\mathcal{L}_1\mathcal{R}\mathcal{E}(h) &= (\mathrm{I}-hA)^{-1}(\mathrm{I}+hB) = \mathrm{I} + h(A+B) + h^2(A^2+AB) + O(h^3),\\ \mathcal{L}_1\mathcal{E}\mathcal{R}(h) &= (\mathrm{I}+hA)(\mathrm{I}+hB)^{-1} = \mathrm{I} + h(A+B) + h^2(AB+B^2) + O(h^3),\\ \mathcal{L}_2\mathcal{R}\mathcal{E}(h) &= (\mathrm{I}-hB)^{-1}(\mathrm{I}+hA) = \mathrm{I} + h(A+B) + h^2(BA+B^2) + O(h^3),\\ \mathcal{L}_2\mathcal{E}\mathcal{R}(h) &= (\mathrm{I}+hB)(\mathrm{I}+hA)^{-1} = \mathrm{I} + h(A+B) + h^2(A^2+BA) + O(h^3).\end{aligned}$$

La **méthode de Crank–Nicolson** peut être décrite à l'aide des ces couplages de splittings différents. On veut résoudre $\dot{x} = Ax$, que l'on écrit

$$\dot{x} = \frac{1}{2}Ax + \frac{1}{2}Ax.$$

On utilise une méthode d'Euler pour une partie et d'Euler rétrograde pour l'autre. On est clairement dans le cadre commutatif et

$$\mathcal{C}(h) = (\mathrm{I} + \frac{1}{2}hA)(\mathrm{I} + \frac{1}{2}hA)^{-1} = (\mathrm{I} + \frac{1}{2}hA)^{-1}(\mathrm{I} + \frac{1}{2}hA).$$

Ceci est équivalent à l'écriture habituelle de cette méthode qui correspond à utiliser la méthode du trapèze pour approcher la formulation intégrale :

$$\frac{x^{n+1}-x^n}{h} = A\frac{x^{n+1}+x^n}{2}.$$

Cette méthode est bien connue pour être d'ordre 2. Pour le démontrer, il faut pousser un peu plus loin le développement

$$\begin{aligned}\mathcal{C}(h) &= (\mathrm{I} + \frac{1}{2}hA)(\mathrm{I} + \frac{1}{2}hA + \frac{1}{4}h^2A^2 + \frac{1}{8}h^3A^3 + O(h^4))\\ &= \mathrm{I} + hA + \frac{1}{2}h^2A^2 + \frac{1}{4}h^3A^3 + O(h^4).\end{aligned}$$

On a

$$\mathcal{C}(h) - \mathcal{A}(h) = \frac{1}{12}h^3A^3 + O(h^4).$$

La méthode est exactement d'ordre 2.

4.4 Vers l'ordre 2 pour un splitting approché

Pour passer à l'ordre 2, on peut envisager différentes solutions :
- monter à l'ordre 2 sur le schéma de splitting,
- monter à l'ordre 2 sur le schéma d'approximation de l'exponentielle,
- monter à l'ordre 2 sur le schéma de splitting et le schéma d'approximation de l'exponentielle.

On commence donc par appliquer une schéma d'Euler sur une formule de Strang. On n'étudie que (S1). L'autre cas est évidemment symétrique en A et B. On calcule

$$\mathcal{S}_1\mathcal{E}(h) = (\mathrm{I} + \frac{h}{2}A)(\mathrm{I} + hB)(\mathrm{I} + \frac{h}{2}A) = \mathrm{I} + h(A+B) + h^2(\frac{1}{2}AB + \frac{1}{2}BA + \frac{1}{4}A^2) + O(h^3),$$

ce qui donne une erreur locale de

$$\mathcal{S}_1\mathcal{E}(h) - \mathcal{S}(h) = h^2(-\frac{1}{4}A^2 - \frac{1}{2}B^2) + O(h^3).$$

L'ordre de la méthode est 1. Il n'y a pas de miracle. Comme le splitting de Strang est d'ordre 2 dans sa version exponentielle, on a bien évidemment $\mathcal{S}_1\mathcal{E}(h) - \mathcal{S}_1(h) = O(h^2)$ et pas mieux.

On a bien évidemment des résultats semblables avec la méthode d'Euler rétrograde combinée à un splitting de Lie.

Maintenant, on essaie d'appliquer le schéma de Crank–Nicolson au splitting de Lie. On calcule

$$\begin{aligned}
\mathcal{L}_1\mathcal{C}(h) &= (I - \frac{h}{2}A)^{-1}(I + \frac{h}{2}A)(I - \frac{h}{2}B)^{-1}(I + \frac{h}{2}B) \\
&= (\mathrm{I} + hA + \frac{1}{2}h^2A^2 + O(h^3))(\mathrm{I} + hB + \frac{1}{2}h^2B^2 + O(h^3)) \\
&= \mathrm{I} + h(A+B) + h^2(\frac{1}{2}A^2 + \frac{1}{2}B^2 + AB) + O(h^3).
\end{aligned}$$

On a exactement comme on pouvait s'y attendre

$$\mathcal{L}_1\mathcal{C}(h) = \mathcal{L}_1(h) + O(h^3).$$

On n'améliore pas la méthode de Lie en l'approchant à l'ordre 2. C'est moral.

Commutons maintenant les matrices dans $\mathcal{L}_1\mathcal{C}(h)$ pour obtenir

$$\mathcal{P}(h) = (I - \frac{h}{2}A)^{-1}(I + \frac{h}{2}B)(I - \frac{h}{2}B)^{-1}(I + \frac{h}{2}A).$$

Ceci s'appelle la **méthode de Peaceman–Rachford**. Cela consiste à effectuer un demi-pas de temps d'Euler pour A, suivi d'un pas de temps de Crank–Nicolson pour B et enfin un demi-pas de temps d'Euler rétrograde pour A. On calcule

$$\begin{aligned}
\mathcal{P}(h) &= (\mathrm{I} + \frac{1}{2}hA + \frac{1}{4}h^2A^2 + \frac{1}{8}h^3A^3 + O(h^4))(\mathrm{I} + hB + \frac{1}{2}h^2B^2 + \frac{1}{4}h^3B^3 + O(h^4))\ (\mathrm{I} + \frac{1}{2}hA) \\
&= \mathrm{I} + h(A+B) + \frac{1}{2}h^2(A^2 + AB + BA + B^2) \\
&\quad + \frac{1}{4}h^3(A^3 + A^2B + ABA + AB^2 + B^2A + B^3) + O(h^4).
\end{aligned}$$

L'erreur locale est donnée par

$$\begin{aligned}
\mathcal{P}(h) - \mathcal{S}(h) &= \frac{1}{4}h^3(A^3 + A^2B + ABA + AB^2 + B^2A + B^3) \\
&\quad - \frac{1}{6}h^3(A^3 + A^2B + ABA + BA^2 + BAB + AB^2 + B^2A + B^3) + O(h^4) \\
&= \frac{1}{12}h^3(A^3 + A^2B + ABA + AB^2 + B^2A + B^3) - \frac{1}{6}h^3(BA^2 + BAB) + O(h^4).
\end{aligned}$$

On obtient une méthode exactement d'ordre 2.

Une autre façon d'obtenir l'ordre 2 est bien sûr de combiner une méthode de Strang avec un méthode de Crank–Nicolson.

$$\begin{aligned}
\mathcal{S}_1\mathcal{C}(h) &= (\mathrm{I}+\frac{h}{4}A)(\mathrm{I}-\frac{h}{4}A)^{-1}(\mathrm{I}+\frac{h}{2}B)(\mathrm{I}-\frac{h}{2}B)^{-1}(\mathrm{I}+\frac{h}{4}A)(\mathrm{I}-\frac{h}{4}A)^{-1} \\
&= (\mathrm{I}+\frac{1}{2}hA+\frac{1}{8}h^2A^2+\frac{1}{32}h^3A^3+O(h^4))(\mathrm{I}+hB+\frac{1}{2}h^2B^2+\frac{1}{4}h^3B^3+O(h^4))\times \\
&\quad \times(\mathrm{I}+\frac{1}{2}hA+\frac{1}{8}h^2A^2+\frac{1}{32}h^3A^3+O(h^4)) \\
&= \mathrm{I}+h(A+B)+\frac{1}{2}h^2(A^2+AB+BA+B^2) \\
&+ h^3(\frac{3}{16}A^3+\frac{1}{8}(A^2B+BA^2)+\frac{1}{4}(ABA+AB^2+B^2A+B^3))+O(h^4).
\end{aligned}$$

L'erreur locale est donnée par

$$\begin{aligned}
\mathcal{S}_1\mathcal{C}(h)-\mathcal{S}(h) &= h^3(\frac{3}{16}A^3+\frac{1}{8}(A^2B+BA^2)+\frac{1}{4}(ABA+AB^2+B^2A+B^3)) \\
&\quad -\frac{1}{6}h^3(A^3+A^2B+ABA+BA^2+BAB+AB^2+B^2A+B^3)+O(h^4) \\
&= \frac{1}{48}h^3(A^3-2(A^2B+BA^2)+4(ABA+AB^2+B^2A+B^3)-8BAB) \\
&\quad +O(h^4).
\end{aligned}$$

La méthode n'est pas plus précise que l'ordre 2.

4.5 Interprétations de la méthode de Peaceman–Rachford

De nombreux travaux traitent de la méthode de Peaceman–Rachford. Ils l'introduisent de manières diverses.

4.5.1 Régularisation

Cette méthode est parfois présentée comme une régularisation de la méthode d'Euler appliquée au système d'origine :

$$\frac{x^{n+1}-x^n}{h}=(A+B)x^n$$

est remplacé par

$$(\mathrm{I}-\frac{h}{2}B)(\mathrm{I}-\frac{h}{2}A)\frac{x^{n+1}-x^n}{h}=(A+B)x^n$$

ou de manière équivalente

$$(\mathrm{I}-\frac{h}{2}B)(\mathrm{I}-\frac{h}{2}A)x^{n+1}=(\mathrm{I}-\frac{h}{2}B)(\mathrm{I}-\frac{h}{2}A)x^n+h(A+B)x^n=(\mathrm{I}+\frac{h}{2}B)(\mathrm{I}+\frac{h}{2}A)x^n.$$

On retrouve bien la méthode de Peaceman–Rachford.

4.5.2 Prédicteur–correcteur

La méthode du prédicteur correcteur consiste ici à commencer par prédire une valeur au temps $t^{n+1/2}$ en calculant

$$x^{n+1/2}=(\mathrm{I}-\frac{h}{2}A)^{-1}(\mathrm{I}-\frac{h}{2}B)^{-1}x^n.$$

Nous avons déjà vu que ceci donne une approximation d'ordre local 2 de la valeur de $x^{n+1/2}$. On utilise ensuite cette valeur pour calculer x^{n+1} par une méthode explicite utilisant la matrice $A+B$ complète :

$$\frac{x^{n+1}-x^n}{h}=(A+B)x^{n+1/2}.$$

On a donc

$$x^{n+1} = x^n + h(A+B)(\mathrm{I} - \frac{h}{2}A)^{-1}(\mathrm{I} - \frac{h}{2}B)^{-1}x^n.$$

La méthode est donc donnée par

$$\begin{aligned}
\mathcal{PC}(h) &= \mathrm{I} + h(A+B)(\mathrm{I} - \frac{h}{2}A)^{-1}(\mathrm{I} - \frac{h}{2}B)^{-1} \\
&= \left((\mathrm{I} - \frac{h}{2}B)(\mathrm{I} - \frac{h}{2}A) + h(A+B)\right)(\mathrm{I} - \frac{h}{2}A)^{-1}(\mathrm{I} - \frac{h}{2}B)^{-1} \\
&= (\mathrm{I} + \frac{h}{2}B)(\mathrm{I} + \frac{h}{2}A)(\mathrm{I} - \frac{h}{2}A)^{-1}(\mathrm{I} - \frac{h}{2}B)^{-1}.
\end{aligned}$$

Ceci n'est pas exactement la méthode de Peaceman–Rachford telle que nous l'avons donnée, mais l'ordre est clairement le même.

4.5.3 Triangularisation

Cette méthode est introduite dans le cas où $A + B$ est auto-adjointe et où on décompose de telle manière que $A = B^*$ est triangulaire inférieure. On peut cependant décrire cette méthode dans un contexte plus général. Il s'agit ici moralement de propager $(A + B)/2$ sur un pas de temps puis $(A+B)/2$ sur un deuxième pas de temps. (Ici $x^{n+1/2}$ ne désigne pas une approximation de $x(t^{n+1/2})$.)

$$\begin{aligned}
\frac{x^{n+1/2} - x^n}{h} &= \frac{1}{2}(Ax^n + Bx^{n+1/2}), \\
\frac{x^{n+1} - x^{n+1/2}}{h} &= \frac{1}{2}(Ax^{n+1} + Bx^{n+1/2}).
\end{aligned}$$

On voit facilement que

$$\begin{aligned}
x^{n+1/2} &= (\mathrm{I} - \frac{h}{2}B)^{-1}(\mathrm{I} + \frac{h}{2}A)x^n, \\
x^{n+1} &= (\mathrm{I} - \frac{h}{2}A)^{-1}(\mathrm{I} + \frac{h}{2}B)x^{n+1/2}.
\end{aligned}$$

On retrouve cette fois exactement la méthode de Peaceman–Rachford.

Toutes ces interprétations sont susceptibles de différer si les matrices dépendent du temps. Dans cette dernière interprétation par exemple A est toujours appliquée à des x^n «entiers» et B à des $x^{n+1/2}$ «demi-entiers».

4.6 Et si A et B commutent ?

Nous avons vu que toutes les approximations à base d'exponentielles sont exactes si A et B commutent. Est-ce encore vrai lorsque l'on approche les exponentielles ? Sinon, peut-on espérer au moins un gain en ordre ? Dans tous les cas l'erreur locale sera la même que l'on compare avec $\mathcal{S}(h)$ ou avec la méthode de Lie, de Strang, ...puisqu'elles sont équivalentes. Reprenons toutes les méthodes déjà vues une-à-une.

4.6.1 Schéma d'Euler–Lie

On calcule l'erreur locale

$$\mathcal{L}_1\mathcal{E}(h) - \mathcal{S}(h) = \mathcal{L}_2\mathcal{E}(h) - \mathcal{S}(h) = -\frac{1}{2}h^2(A^2 + B^2) + O(h^3).$$

Si on ne considère pas l'ordre le traitement de A et B est le même. Les deux méthodes $\mathcal{L}_1\mathcal{E}(h)$ et $\mathcal{L}_2\mathcal{E}(h)$ sont donc clairement identiques. Aucun gain en ordre n'est observé.

4.6.2 Schéma d'Euler rétrograde–Lie

On calcule l'erreur locale

$$\mathcal{L}_1\mathcal{R}(h) - \mathcal{S}(h) = \mathcal{L}_2\mathcal{R}(h) - \mathcal{S}(h) = \frac{1}{2}h^2(A^2 + B^2)) + O(h^3).$$

La conclusion est la même que pour Euler.

4.6.3 Schéma d'Euler–Strang

Cette fois-ci le traitement de A et de B sont différents pour les deux méthodes de Strang, on s'attend donc à des erreurs différentes. C'est bien le cas :

$$\begin{aligned}\mathcal{S}_1\mathcal{E}(h) - \mathcal{S}(h) &= h^2(-\frac{1}{4}A^2 - \frac{1}{2}B^2) + O(h^3),\\ \mathcal{S}_2\mathcal{E}(h) - \mathcal{S}(h) &= h^2(-\frac{1}{4}B^2 - \frac{1}{2}A^2) + O(h^3).\end{aligned}$$

4.6.4 Schéma de Crank–Nicolson–Lie

Le schéma de Crank-Nicolson Lie est exactement le même que celui de Peaceman–Rachford si A et B commutent, puisque l'on a déduit l'un de l'autre en commutant.

Ce schéma devient donc du même ordre de Peaceman–Rachford, c'est-à-dire au moins 2.

4.6.5 Schéma de Peaceman–Rachford

La formule d'erreur locale pour la méthode de Peaceman–Rachford est nettement simplifiée si A et B commutent. En effet $A^2B + ABA - 2BA^2 = 0$ et $AB^2 + B^2A - BAB = 0$ donc

$$\mathcal{P}(h) - \mathcal{S}(h) = \frac{1}{12}h^3(A^3 + B^3) + O(h^4).$$

L'ordre n'est cependant pas amélioré.

5 Systèmes raides

Toutes les estimations d'erreur précédentes contiennent des $O(h)$ qui dépendent en pratique des normes $\|A\|$ et $\|B\|$. Nous avons jusqu'à maintenant supposé implicitement que ces normes étaient d'ordre $O(1)$ et que le temps sur lequel on regardait l'évolution étaient également d'ordre $O(1)$.

Nous allons maintenant et continuer à regarder l'équation sur des temps $O(1)$ et à discrétiser l'équation avec des pas de temps h. Nous introduisons en outre un deuxième petit paramètre ε, plus petit que le pas de temps : $\varepsilon \ll h$.

Le système est sensé être raide, c'est-à-dire plus particulièrement de la forme

$$\dot{x} = (A + B)x,$$

avec A raide et B non raide, à savoir

$$\|A\| = O(1/\varepsilon), \qquad \|B\| = O(1).$$

5.1 Raide mais stable

On se place dans le cas scalaire le plus simple avec

$$\dot{x} = \frac{1}{\varepsilon}ax.$$

Sa solution est $x(t) = \exp(at/\varepsilon)x(0)$. Si $a > 0$ et $t = O(1)$ ceci tend vers l'infini and $\varepsilon \to 0$. On est donc dans un cas instable. Dans ce cas, toute méthode numérique est vouée à l'échec.

Il faut donc que les valeurs propres en $O(1/\varepsilon)$ de A soient de parties réelles négatives. On a alors $\lim_{\varepsilon \to 0} x(t) = 0$ pour l'équation scalaire.

Un pas de temps tel que $h \gg \varepsilon$ ne permettra pas de décrire correctement le comportement transitoire mais ce n'est pas ce que nous essayons de décrire.

Dans le contexte matriciel, cela veut dire qu'une itération de la matrice A va correspondre «moralement» à annuler les composantes selon les directions propres correspondant aux valeurs propres d'ordre $O(1/\varepsilon)$, c'est-à-dire à projeter sur un sous-espace vectoriel perpendiculaire à ces directions.

Ces quelques remarques vont guider les hypothèses supplémentaires effectuées sur la matrice A.

On notera dans la suite par C toute constante qui ne dépend ni du pas de temps h, ni de la raideur ε.

5.2 Cadre de l'étude

5.2.1 Hypothèses sur la matrice raide

Le but étant de mettre en évidence des pertes d'ordre par rapport au cas standard, on peut se placer dans un cas légèrement particulier. On suppose que A est diagonalisable et que ses valeurs propres se divisent en deux sous-ensembles bien séparés, les grandes valeurs propres de partie réelle négative :

$$\mathrm{Re}(\frac{1}{\varepsilon}\lambda_k h) \ll -1$$

et les petites valeurs propres :

$$\mu_k h = O(h).$$

Soit P la matrice de passage de la base canonique dans laquelle est écrite x à celle des vecteurs propres de A. On pose $y = Px$ et $A = P^{-1}DP$ où $D = \mathrm{diag}(d_k)$ est diagonale. Le système complet devient :

$$\dot{y} = Dy + PBP^{-1}y.$$

On suppose de plus qu'il existe une constante C (indépendante de ε) telle que

$$\mathrm{Re}(d_k) \leq C, \qquad \|P\| \leq C \qquad \text{et} \qquad \|P^{-1}\| \leq C.$$

Ceci autorise en revanche les valeurs propres de partie réelle très négatives. Sous ces hypothèses, on a

$$\|hD\| \gg 1 \qquad \text{et} \qquad \|hPBP^{-1}\| = O(h).$$

Comme $\|P^{-1}\| \leq C$, un résultat d'ordre sur y sera valide automatiquement pour $x = P^{-1}y$.

Pour la suite et pour alléger les notations, on supposera donc que le problème est directement posé dans la base où A est diagonale :

$$\dot{x} = Dx + Bx.$$

5.2.2 Composantes raides et non raides

On distingue entre les composantes raides et non raides du vecteur solution $x(t)$. Ceci a un sens puisque l'on est dans la base des vecteurs propres de A. On note R_k l'opérateur de restriction à la $k$ième composante. Sans hypothèses particulières sur A (symétrie), ceci n'est pas une projection sur un sous-espace car ceci suppose des hypothèses d'orthogonalité. On dit que x_k est une **composante raide** si elle est associée à une valeur propre raide λ_k/ε :

$$\dot{x}_k(t) = \frac{1}{\varepsilon}\lambda_k x_k(t) + R_k B x(t).$$

On dit que x_k est une **composante non raide** si elle est associée à une valeur propre non raide μ_k :

$$\dot{x}_k(t) = \mu_k x_k(t) + R_k B x(t).$$

5.2.3 Estimations sur la solution exacte

Les hypothèses donnent que $\dot{x} = (D+B)x$ a une solution

$$|x(t)| \leq C\exp(Ct),$$

et comme $t = O(1)$, $|x(t)| \leq C$.

On peut donner une bien meilleure estimation pour une composante raide. Pour cela, on utilise la formulation intégrale de l'équation différentielle

$$\begin{aligned} x_k(t+h) &= \exp(h\lambda_k/\varepsilon)x_k(t) + \int_0^h \exp(s\lambda_k/\varepsilon)R_k B x(t+h-s)ds, \\ |x_k(t+h)| &\leq |\exp(h\lambda_k/\varepsilon)||x_k(t)| + C\varepsilon\left|\frac{1-\exp(h\lambda_k/\varepsilon)}{\lambda_k}\right| \max_{s\in[0,h]}\|x(t+s)\|. \end{aligned}$$

Comme $\mathrm{Re}(\lambda_k) < 0$ et $h \gg \varepsilon$, l'exponentielle est négligeable et

$$|x_k(t+h)| \leq C\varepsilon\frac{1}{|\lambda_k|} = O(\varepsilon).$$

A partir d'une donnée initiale $x_k(0) \neq 0$, ce qui est la situation générique, on obtient un comportement transitoire rapide.

5.2.4 Première estimation sur la solution splittée

Les hypothèses séparées sur D et B ne sont pas pires que celles sur $D+B$, on a donc également et quelque soit la façon dont on effectue le splitting

$$|x^m| \leq C\exp(Chm),$$

qui reste borné sur un intervalle de temps tel que $hm = O(1)$.

5.2.5 Erreur locale et globale

On se place dans le monde idéal (et vrai pour D) où on sait calculer les exponentielles de matrices. On n'étudie pas ici l'impact supplémentaire d'une méthode d'approximation des exponentielles.

Dans la preuve du théorème de calcul de l'ordre à partir de l'erreur locale, beaucoup de quantités étaient en $O(1)$, ce qui permettait de conclure. Nous allons recalculer les erreurs locales et en déduire les erreurs globales pour chaque schéma de splitting.

La définition de l'erreur locale peut être étendue pour tous les temps t et pas seulement ceux de la forme $t_m = mh$. On note

$$\eta(t+h) = (\mathcal{S}(h) - f(h))x(t) = x(t+h) - f(h)x(t).$$

On retrouve la définition précédente : $\eta^m = \eta(mh)$.

L'erreur globale, en revanche, n'a toujours de sens qu'aux temps de discrétisation $t_m = mh$:

$$e^m = e(mh) = x(mh) - x^m = (S(mh) - f(h)^m)x^0.$$

Comme on va distinguer les composantes raides et non raides, on utilise une version composante par composante de la relation de récurrence

$$\text{(RR) } e_k^{m+1} = R_k f(h) e^m + \eta_k^{m+1}.$$

5.3 Le splitting de Lie (L1)

Rappelons que le splitting (L1) consiste à commencer par un pas de temps non raide suivi d'un pas de temps raide :

$$\mathcal{L}_1(h) = \exp(hD)\exp(hB).$$

Composantes raides Soit x_k une composante raide. Commençons par estimer

$$x_k^{m+1} = \exp(\frac{1}{\varepsilon}\lambda_k h) R_k \exp(hB) x^m.$$

Comme $\exp(hB)$ est borné, ceci est exponentiellement petit, et à tous les ordres dominants en h :

$$x_k^{m+1} = O(\varepsilon),$$

de même que pour l'estimation de la composante raide de la solution exacte. Dans l'expression (RR), on évalue également

$$R_k \mathcal{L}_1(h) e^m = \exp(\frac{1}{\varepsilon}\lambda_k h) R_k \exp(hB) e^m,$$

et (RR) devient également à tous les ordres dominants en h :

$$e_k^{m+1} \sim \eta_k^{m+1} = O(\varepsilon).$$

Les composantes, les erreurs, bref tout est en $O(\varepsilon)$. Ceci n'est pas relié au pas de temps h, et est plus petit que tous les ordres de h dans notre hypothèse tant que $h/\varepsilon \gg 1$.

Composantes non raides Soit x_k une composante non raide. On distingue ce qui se passe à la première itération et aux itérations suivantes.

À la première itération, on a

$$\begin{aligned} \eta_k(h) &= x_k(h) - \exp(h\mu_k) R_k(\exp(hB)x(0)) \\ &= x_k(h) - x_k(0) + R_k((\exp(h\mu_k)\exp(hB) - \mathrm{I})x(0)) \\ &= x_k(h) - x_k(0) + O(h). \end{aligned}$$

À la première itération, les composantes non raides ne sont absolument pas affectées par la raideur du système. Comme

$$\dot{x}_k(t) = R_k(D+B)x_k(t) = \mu_k x_k(t) + R_k B x(t),$$

on a également $x_k(h) - x_k(0) = O(h)$, indépendemment de la raideur ε. En conclusion, pour la première itération, $\eta_k(h) = O(h)$.

Pour les itérations suivantes,

$$\eta_k(t+h) = x_k(t+h) - \exp(h\mu_k) R_k \exp(hB) x(t).$$

On développe en tenant compte que $x(t) = O(1)$,

$$\begin{aligned}
\exp(h\mu_k) R_k \exp(hB) x(t) &= (1 + h\mu_k + O(h^2)) R_k \exp(hB) x(t) \\
&= R_k \exp(hB) x(t) + h\mu_k R_k \exp(hB) x(t) + O(h^2) \\
&= R_k (\mathrm{I} + hB) x(t) + h\mu_k R_k \mathrm{I} x(t) + O(h^2) \\
&= x_k(t) + h R_k B x(t) + h\mu_k x_k(t) + O(h^2) \\
&= x_k(t) + h\dot{x}_k(t) + O(h^2).
\end{aligned}$$

toujours indépendemment de la raideur ε. On reconnaît le début du développement limité :

$$x_k(t+h) = x_k(t) + h\dot{x}_k(t) + ...$$

Pour continuer ce développement, il faut savoir estimer la dérivée seconde

$$\ddot{x}_k(t) = R_k(D^2 + DB + BD + B^2) x(t).$$

Les lignes $R_k D^2$, $R_k DB$ et $R_k B^2$ sont indépendantes de ε. Ce n'est pas le cas de $R_k BD$. Mais les composantes de $Dx(t)$ qui correspondent aux valeurs propres raides sont de taille $O(\varepsilon)$ comme B est également borné, la dérivée seconde $\ddot{x}_k(t)$ est bornée. On a donc

$$x_k(t+h) = x_k(t) + h\dot{x}_k(t) + O(h^2)$$

et

$$\eta_k(t+h) = x_k(t+h) - \exp(h\mu_k) R_k \exp(hB) x(t) = O(h^2).$$

L'erreur locale est en $O(h^2)$.

Convergence La convergence veut que l'erreur globale décroisse avec h, au moins en $O(h)$.

On n'est pas vraiment capable de montrer la convergence des composantes raides. Ceci n'est certainement pas le cas, la valeur calculée est *a priori* très mauvaise, mais également très petite. En effet, la solution exacte ainsi que l'approximation sont en $O(\varepsilon)$. Était-il de toutes façon raisonnable de vouloir bien approcher ces composantes en prenant un pas de temps h très grand devant ε ?

Nous espérons au moins converger sur les composantes non raides, celles pour lesquelles les valeurs calculées ne sont pas exponentiellement petites. Pour les composantes non raides, on a

$$e_k^{m+1} = R_k \mathcal{L}_1(h) e^m + \eta_k(t_m + h)$$

et comme on a vu que $\eta_k(t_m + h) = O(h^2)$ et $e_k^1 = O(h)$, on a $e_k^{m+1} = O(h)$ par le même raisonnement que celui de l'étude générale dans le cas non raide.

Le splitting (L1) approche très mal les composantes raides et donne sur les composantes non raides le même ordre que s'il n'y avait pas de raideur dans le système. Comparons avec les autres types de splitting.

5.4 Le splitting de Lie (L2)

Le splitting (L2) consiste à commencer par un pas de temps raide suivi d'un pas de temps non raide :

$$\mathcal{L}_2(h) = \exp(hB) \exp(hD).$$

Composantes raides On a en général

$$\begin{aligned}\eta_k(t+h) &= x_k(t+h) - R_k \exp(hB)\exp(hD)x(t)\\ &= x_k(t+h) - R_k(\mathrm{I}+hB+O(h^2))\exp(hD)x(t)\\ &= x_k(t+h) - R_k(\mathrm{I}+hB)\exp(hD)x(t) + O(h^2)\end{aligned}$$

Si maintenant x_k est une composante raide, on a de plus

$$\begin{aligned}\eta_k(t+h) &= x_k(t+h) - \exp(\frac{1}{\varepsilon}\lambda_k h)x_k(t) - hR_kB\exp(hD)x(t) + O(h^2)\\ &= O(\varepsilon) - hR_kB\exp(hD)x(t) + O(h^2)\\ &= O(\varepsilon) + O(h).\end{aligned}$$

Cette fois-ci on a bien une erreur locale qui décroît moralement avec h, mais cette erreur est potentiellement énorme relativement à la valeur exactes des composantes raides en $O(\varepsilon)$. L'erreur relative est en $O(h/\varepsilon)$.

On verra que ces erreurs locales ne s'accumulent cependant pas au cours des itérations.

Composantes non raides La matrice $\exp(-hB)$ est régulière et on a

$$\begin{aligned}\exp(-hB)\eta(t+h) &= \exp(-hB)x(t+h) - \exp(-hB)\exp(hB)\exp(hD)\exp(hB)\exp(-hB)x(t)\\ &= \exp(-hB)x(t+h) - \exp(hD)\exp(hB)\exp(-hB)x(t)\\ &= (\mathrm{I}-hB)x(t+h) - \exp(hD)\exp(hB)(\mathrm{I}-hB)x(t) + O(h^2)\\ &= x(t+h) - \exp(hD)\exp(hB)x(t) - hBx(t+h) + h\exp(hD)Bx(t) + O(h^2).\end{aligned}$$

Les deux premiers termes sont exactement l'erreur locale $\eta_k^{\mathcal{L}_1}(t+h)$ pour le splitting de Lie (L1) calculé précédemment. Les $O(h^2)$ écrits jusqu'à maintenant ne dépendent pas de la raideur ε.

Soit maintenant x_k une composante non raide. On a

$$R_k h\exp(hD)Bx(t) = h\exp(h\mu_k)R_kBx(t) = hR_kBx(t) + O(h^2).$$

On a donc

$$\begin{aligned}R_k\exp(-hB)\eta(t+h) &= \eta_k^{\mathcal{L}_1}(t+h) - hR_kBx(t+h) + hR_kBx(t) + O(h^2)\\ &= \eta_k^{\mathcal{L}_1}(t+h) - h(\dot{x}_k(t+h) - \mu_k x_k(t+h)) + h(\dot{x}_k(t) - \mu_k x_k(t)) + O(h^2).\end{aligned}$$

Pour $t=0$, on a

$$R_k\exp(-hB)\eta(h) = \eta_k^{\mathcal{L}_1}(h) - h(\dot{x}_k(h) - \mu_k x_k(h)) + h(\dot{x}_k(0) - \mu_k x_k(0)) + O(h^2) = O(h),$$

car $\dot{x}_k(0)$ est borné. Pour $t>0$, on a vu que la fonction $\ddot{x}(t)$ est bornée, ainsi

$$R_k\exp(-hB)\eta(t+h) = \eta_k^{\mathcal{L}_1}(t+h) - h^2\ddot{x}_k(t+h) + h^2\mu_k\dot{x}_k(t+h) + O(h^2) = O(h^2).$$

Ces estimations sont indépendantes de la raideur ε. Maintenant, on a

$$\eta(t+h) = \exp(hB)\exp(-hB)\eta(t+h)$$

et il faut tenir compte à la fois des composantes raides et non raides. Pour une composante raide

$$R_k\exp(-hB)\eta(t+h) = \eta_k^{\mathcal{L}_1}(t+h) + O(h) = O(\varepsilon) + O(h).$$

Pour une composante non raide

$$\begin{aligned}\eta(t+h) &= \exp(hB)\exp(-hB)\eta(t+h) = (\mathrm{I}+hB)\exp(-hB)\eta(t+h) + O(h^2)\\ \eta_k(t+h) &= R_k\exp(-hB)\eta(t+h) + hR_kB\exp(-hB)\eta(t+h) + O(h^2).\end{aligned}$$

Ainsi

$$\eta_k(h) = R_k \exp(-hB)\eta(h) + hR_kB\exp(-hB)\eta(h) + O(h^2) = O(h)$$

et pour $t > 0$

$$\eta_k(t+h) = R_k \exp(-hB)\eta(t+h) + hR_kB\exp(-hB)\eta(t+h) + O(h^2) = O(h^2).$$

Dans cette dernière estimation, on a négligé les termes en $O(h\varepsilon)$.

On a exactement les mêmes estimations pour la partie non raide que pour le splitting (L1).

Convergence Pour étudier la convergence, on considère la récurrence

$$e^{m+1} = \exp(hB)\exp(hD)e^m + \eta(t_m + h).$$

Les erreurs locales des composantes raides et non raides sont couplées. Les erreurs e_k^m pour les composantes raides sont «tuées» à chaque itération par $\exp(h\lambda_k/\varepsilon)$ et $\exp(hB)$ est borné. Seules les erreurs des composantes non raides peuvent être accumulées. On obtient alors, comme pour le schéma de splitting (L1) et pour $mh = O(1)$,

$$e^{m+1} = O(h).$$

5.5 Les splittings de Strang (S1) et (S2)

L'ordre 2 obtenu pour les méthodes de Strang dans le cas sans raideur est dû à une certaine symétrie entre les splittings (L1) et (L2), qui crée des compensations quand elles sont utilisées successivement sur un demi-pas de temps :

$$\mathcal{S}_1(h) = \mathcal{L}_1(\frac{h}{2})\mathcal{L}_2(\frac{h}{2}), \qquad \text{et} \qquad \mathcal{S}_2(h) = \mathcal{L}_2(\frac{h}{2})\mathcal{L}_1(\frac{h}{2}).$$

Dans notre cas, cette symétrie est perdue ainsi que les compensations. Comme les splittings sont de toute façon stables, c'est la deuxième moitié de chaque itération qui guide l'ordre du schéma. Ainsi le schéma de Strang (S1) a le même ordre que le schéma de Lie (L1) ; et le schéma de Strang (S2) a le même ordre que le schéma de Lie (L2).

5.6 Récapitulation

On compare les différents schémas en comparant les erreurs et erreurs relatives globales. Pour une composante non raide, elles sont du même ordre. On a le tableau suivant.

	(L1)	(L2)	(S1)	(S2)
e_k non raide	$O(h)$	$O(h)$	$O(h)$	$O(h)$
e_k raide	$O(\varepsilon)$	$O(h)$	$O(\varepsilon)$	$O(h)$
e_k/x_k raide	$O(1)$	$O(h/\varepsilon)$	$O(1)$	$O(h/\varepsilon)$

Pour des raisons d'approximation ou plutôt d'ordre des composantes raides, on a intérêt à utiliser les schémas (S1) ou (L1). Comme il n'y a pas d'intérêt du point de vue de l'ordre à utiliser une méthode de Strang et que cela est légèrement plus coûteux, il suffit d'utiliser le schéma de Lie (L1).

5.7 Une méthode alternative : le splitting de source

Dans ce qui précède, on introduit de manière artificielle une phase transitoire à chaque itération en temps, alors que le problème initial n'a cette phase transitoire que sur un intervalle de temps initial en $O(\varepsilon)$. Une méthode alternative permet d'éviter ces transitoires artificiels. On espère alors trouver de meilleures estimations pour les composantes raides.

La méthode en question s'appelle la **méthode (ST)** pour «Source Splitting», ou aussi méthode NTS («No Time Splitting»).

5.7.1 Description heuristique

Comme pour la méthode des facteurs intégraux, les rôles des matrices A et B sont très dissymétriques. Décrivons la méthode (ST) en utilisant les description simples du début du cours avant de passer à la notation avec les semi-groupes.

$$\left| \begin{array}{ll} \dot{y} = By, & y(t) = x(t), \\ \dot{z} = Az + s \text{ avec } s = (y(t+h) - y(t))/h, & z(t) = x(t). \end{array} \right.$$

On pose alors $x(t+h) = z(t+h)$. Que fait-on ? Comme pour les méthodes qui marchaient bien précédemment, on traite d'abord la partie non raide, puis la partie raide de l'équation. La partie non raide est traitée de manière explicite. C'est en quelque sorte une partie de prédiction pour les coordonnées non raides. Ensuite, on prend comme donnée initiale pour la deuxième étape la vraie donnée initiale $x(t)$. On supprime ainsi la partie transitoire artificielle, car les données initiales sur les composantes raides seront d'ordre $O(\varepsilon)$. Ceci oblige à ajouter un terme source dans la deuxième équation qui est l'équation complète (S), dans laquelle on a remplacé le terme Bx par

$$Bx(\tau) \simeq By(\tau) = \dot{y}(\tau) \simeq \frac{y(t+h) - y(t)}{h} \quad \text{pour} \quad \tau \in [t, t+h].$$

Ceci est une approximation qui semble du premier ordre. C'est ce que l'on espère obtenir, tout en ayant une meilleure approximation des composantes raides.

5.7.2 Description par les semi-opérateurs

On a exactement $y(t+h) = \exp(hB)y(t)$ et donc $s = (\exp(hB) - \mathrm{I})y(t)/h$. Ensuite l'équation $\dot{z} = Az + s$ se réécrit

$$\frac{d}{dt}\left(\exp(-tA)z(t)\right) = \exp(-tA)s$$

qui s'intègre exactement en

$$\exp(-(t+h)A)z(t+h) - \exp(-tA)z(t) = \int_t^{t+h} \exp(-\tau A)d\tau\ s = -A^{-1}(\exp(-(t+h)A) - \exp(-tA))s.$$

Ainsi

$$\begin{aligned} z(t+h) &= \exp(hA)z(t) + A^{-1}(\exp(hA) - \mathrm{I})s \\ &= \exp(hA)z(t) + \frac{1}{h}A^{-1}(\exp(hA) - \mathrm{I})(\exp(hB) - \mathrm{I})y(t). \end{aligned}$$

Comme $y(t) = z(t) = x(t)$, l'opérateur d'évolution associé à la méthode (SP) est

$$\mathcal{N}(h) = \exp(hA) + \frac{1}{h}A^{-1}(\exp(hA) - \mathrm{I})(\exp(hB) - \mathrm{I}).$$

Ici la matrice A est diagonale, calculer son exponentielle et son inverse ne pose aucun problème.

5.7.3 Estimation de l'ordre dans le cadre non raide

Comme pour les autres méthodes, on peut déterminer l'ordre dans le cas non raide en effectuant un développement limité.

$$\begin{aligned} \mathcal{N}(h) &= \mathrm{I} + hA + \frac{1}{2}h^2A^2 + O(h^3) + \frac{1}{h}A^{-1}(hA + \frac{1}{2}h^2A^2 + O(h^3))(hB + \frac{1}{2}h^2B^2 + O(h^3)) \\ &= \mathrm{I} + hA + \frac{1}{2}h^2A^2 + O(h^3) + \frac{1}{h}(h + \frac{1}{2}h^2A + O(h^3))(hB + \frac{1}{2}h^2B^2 + O(h^3)) \\ &= \mathrm{I} + hA + \frac{1}{2}h^2A^2 + O(h^3) + \frac{1}{h}(h^2B + \frac{1}{2}h^3B^2 + \frac{1}{2}h^3AB + O(h^4)) \\ &= \mathrm{I} + hA + \frac{1}{2}h^2A^2 + O(h^3) + hB + \frac{1}{2}h^2B^2 + \frac{1}{2}h^2AB + O(h^3) \\ &= \mathrm{I} + h(A+B) + \frac{1}{2}h^2(A^2 + B^2 + AB) + O(h^3). \end{aligned}$$

La méthode est exactement d'ordre 1, puisque l'erreur locale est en $O(h^2)$:

$$\mathcal{N}(h) - \mathcal{S}(h) = -\frac{1}{2}h^2 BA + O(h^3).$$

Contrairement aux méthodes de splitting plus classiques à base de produits d'exponentielles, cette méthode ne devient pas exacte si A et B commutent.

5.7.4 Estimation de l'ordre dans le cadre raide

Revenons au contexte qui nous intéresse où la matrice $A \equiv D$ est raide et diagonale.

Écrivons également une formulation intégrale pour le système (S). Par rapport à la méthode (ST), on a simplement un terme source qui dépend du temps et vaut $Bx(t)$:

$$\exp(-(t+h)D)x(t+h) - \exp(-tD)x(t) = \int_t^{t+h} \exp(-\tau D)Bx(\tau)d\tau,$$

$$x(t+h) = \exp(hD)x(t) + \exp(hD)\int_0^h \exp(-\tau D)Bx(t+\tau)d\tau,$$

à comparer avec

$$\mathcal{N}(h)x(t) = \exp(hD)x(t) + \exp(hD)\int_0^h \exp(-\tau D)\frac{1}{h}(\exp(hB) - \mathrm{I})x(t)d\tau.$$

L'erreur locale vaut

$$\eta(t+h) = \exp(hD)\int_0^h \exp(-\tau D)\left(Bx(t+\tau) - \frac{1}{h}(\exp(hB) - \mathrm{I})x(t)\right)d\tau.$$

Or, indépendamment de la raideur ε,

$$\begin{aligned} Bx(t+\tau) - \frac{1}{h}(\exp(hB) - \mathrm{I})x(t) &= B(x(t+\tau) - x(t)) - \frac{1}{h}(\exp(hB) - \mathrm{I} - hB)x(t) \\ &= B(x(t+\tau) - x(t)) + O(h)x(t), \end{aligned}$$

donc

$$\begin{aligned} \eta(t+h) &= \exp(hD)\int_0^h \exp(-\tau D)B(x(t+\tau) - x(t))d\tau + \exp(hD)\int_0^h \exp(-\tau D)d\tau\ O(h)x(t) \\ &= \int_0^h \exp((h-\tau)D)B(x(t+\tau) - x(t))d\tau + D^{-1}(\exp(hD) - \mathrm{I})O(h)x(t). \end{aligned}$$

Composantes raides Pour une composante raide x_k, on a

$$\begin{aligned} \eta_k(t+h) &= \int_0^h \exp((h-\tau)\frac{1}{\varepsilon}\lambda_k)R_k B(x(t+\tau) - x(t))d\tau + \varepsilon\lambda_k^{-1}(\exp(\frac{1}{\varepsilon}\lambda_k h) - 1)R_k O(h)x(t) \\ &= \int_0^h \exp((h-\tau)\frac{1}{\varepsilon}\lambda_k)R_k B(x(t+\tau) - x(t))d\tau + O(h\varepsilon). \end{aligned}$$

Pour estimer l'intégrale restante, il faut distinguer $t = 0$ et $t > 0$.

Pour $t = 0$, on a au pire $B(x(t+\tau) - x(t)) = O(1)$. Comme

$$\int_0^h \exp((h-\tau)\frac{1}{\varepsilon}\lambda_k) = \varepsilon\lambda_k^{-1}(1 - \exp(h\frac{1}{\varepsilon}\lambda_k))$$

on a donc $\eta_k(h) = O(\varepsilon)$.

Pour $t > 0$, on sait que $\dot{x}(t)$ est borné, donc $B(x(t+\tau) - x(t)) = O(h)$. On en déduit que $\eta_k(t+h) = O(h\varepsilon)$.

Composantes non raides Pour une composante non raide x_k, on a

$$\begin{aligned}\eta_k(t+h) &= \int_0^h \exp((h-\tau)\mu_k) R_k B(x(t+\tau)-x(t))d\tau + \mu_k^{-1}(\exp(\mu_k h)-1)R_k O(h)x(t) \\ &= \int_0^h \exp((h-\tau)\mu_k) R_k B(x(t+\tau)-x(t))d\tau + O(h^2).\end{aligned}$$

Pour $t=0$, on utilise le fait que $\exp((h-\tau)\mu_k)$ est borné. On a ainsi $\eta_k(h)=O(h)$. Pour $t>0$, on utilise comme pour les composantes raides que $\dot{x}(t)$ est borné, donc $\eta_k(t+h)=O(h^2)$.

Convergence On a toujours la relation de récurrence

$$e^{m+1} = \mathcal{N}(h)e^m + \eta(t_m+h).$$

Pour une composante raide, on a

$$e_k^{m+1} = \exp(h\frac{1}{\varepsilon}\lambda_k)e_k^m + \frac{1}{h}\varepsilon\lambda_k^{-1}(\exp(h\frac{1}{\varepsilon}\lambda_k)-1)R_k(\exp(hB)-\mathrm{I})e^m + \eta_k(t_m+h),$$

ce qui, en négligeant les termes exponentiellement petits, donne

$$e_k^{m+1} = \frac{\varepsilon}{h}R_k O(h)e^m + \eta_k(t_m+h).$$

En première approximation, $e_k^{m+1} = \eta_k(t_m+h)$.

Quant aux composantes non raides, c'est le même calcul que pour toutes les autres méthodes et on a $e_k^{m+1}=O(h)$.

On peut donc ajouter une colonne au tableau

	(ST)
e_k non raide	$O(h)$
e_k raide	$O(h\varepsilon)$
e_k/x_k raide	$O(h)$

5.8 Illustration numérique

5.8.1 Exemple 1

On considère les données suivantes

$$A = \begin{pmatrix} -10^4 & 10^4 & 1 \\ 10^4 & -10^4 & 2 \\ 1 & 1 & -2 \end{pmatrix}, \qquad B = \begin{pmatrix} -1 & 0.5 & 0.25 \\ 0.1 & 0 & 0.1 \\ 0.2 & 0.4 & -1 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Commentons un peu ces données. La matrice A a pour valeur propre raide $\lambda=-20000$ et deux valeurs propres non raides 1 et -3. Nous regardons l'erreur relative sur chacune des composantes commise au temps $t=1$. A est une matrice raide pour des pas de temps typiquement supérieurs à 10^{-3}. La matrice B est non raide. Elle a comme principale propriété de ne pas commuter avec la matrice A. Enfin, la donnée initiale a des composantes selon les trois directions propres de A.

La figure 1 représente l'erreur relative en norme l^2 due aux cinq schémas de splitting discutés, (L1), (L2), (S1), (S2) et (ST). En abscisse se trouvent les pas de temps.

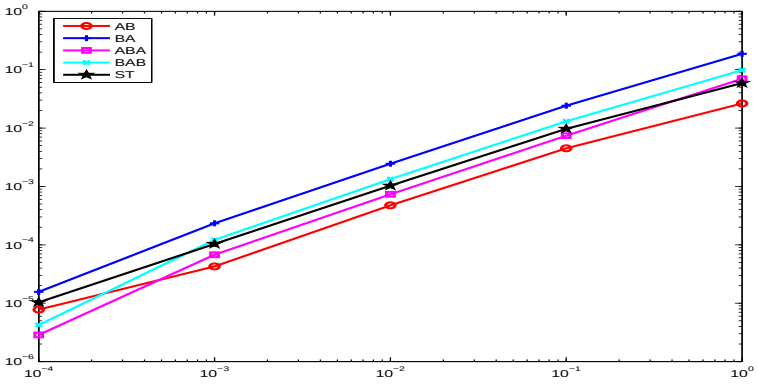

FIG. 1 – Erreurs relatives en norme l^2 pour les cinq méthodes proposées pour les pas de temps $h = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1} et 1

On remarque sur la figure que tous ces schémas sont d'ordre 1 pour des pas de temps $h \gg \varepsilon$. Lorsque le pas de temps h devient de l'ordre de grandeur de ε, ici $h = 10^{-4}$, on commence à voir une amélioration pour les schémas qui sont théoriquement d'ordre 2 pour des systèmes non raides. On remarque en particulier que du point de vue de l'ordre global, le schéma (ST) n'est meilleur que les autres.

La figure 2 représente pour chacune des composantes dans les coordonnées propres y_1, y_2 et y_3, l'erreur relative due aux cinq schémas de splitting discutés, (L1), (L2), (S1), (S2) et (ST). En abscisse se trouvent à nouveau les pas de temps.

Une représentation identique à celle de la figure 2, mais portant sur les coordonnées x_1, x_2 et x_3 dans la base d'origine donne trois courbes sensiblement identiques à celles de la figure 1 et ne présente donc pas d'intérêt. Ici, dans la base propre de A, on voit un comportement largement différent sur la composante raide y_1 et sur les deux composantes non raides y_2 et y_3.

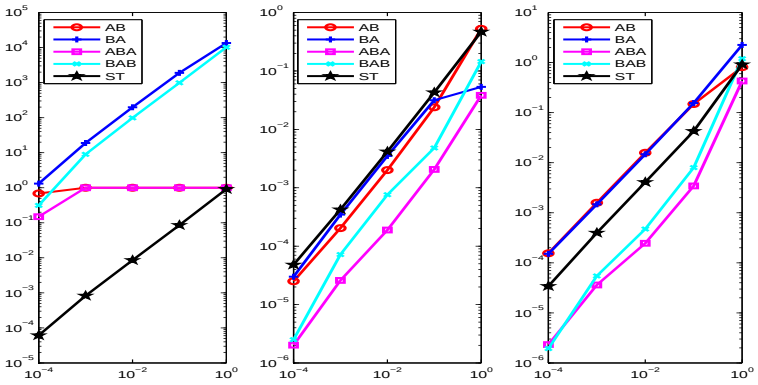

FIG. 2 – Erreurs relatives sur les coordonnées propres y_1, y_2 et y_3 pour les cinq méthodes proposées pour les pas de temps $h = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1} et 1

Pour les coordonnées non raides, on observe un comportement d'ordre 1 sans plus. On observe entre autres la perte d'ordre des méthodes de type Strang, due à l'interaction avec la composante raide.

Pour la coordonnée raide, on remarque un comportement très différents des méthodes. Les méthodes (BA) et (BAB) qui finissent par un pas non raide, sont d'ordre 1, mais avec une constante très grande. L'erreur relative est en $O(1)$ pour les méthodes qui finissent par un pas de temps raide, (AB) et (ABA), du moins tant que $h \gg \varepsilon$. Dès que $h \simeq \varepsilon$, la méthode (ABA) commence à devenir meilleure.

La méthode (ST) est d'ordre 1 avec une constante d'ordre 1. C'est elle qui donne les meilleurs résultats, comme prévu, pour cette composante raide, alors que ses performances sont moyennes sur les

composantes non raides. Le comportement est meilleur sur la composante non raide y_3 correspondant à la valeur propre négative -3.

5.8.2 Exemple 2

On considère les données suivantes

$$A = \begin{pmatrix} -10^6 & -10^6 & 1 \\ 10^6 & -10^6 & 2 \\ 1 & 1 & -2 \end{pmatrix}, \qquad B = \begin{pmatrix} -1 & 0.5 & 0.25 \\ 0.1 & 0 & 0.1 \\ 0.2 & 0.4 & -1 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Seule la matrice A diffère par rapport à l'exemple précédent. Son spectre est composé de deux valeurs propres raides $\lambda_\pm = -10^6(1 \pm i)$ complexes conjuguées et d'une valeur propre non raide $\mu = -2$.

La figure 3 représente l'erreur relative en norme l^2 due aux cinq schémas de splitting discutés, (L1), (L2), (S1), (S2) et (ST).

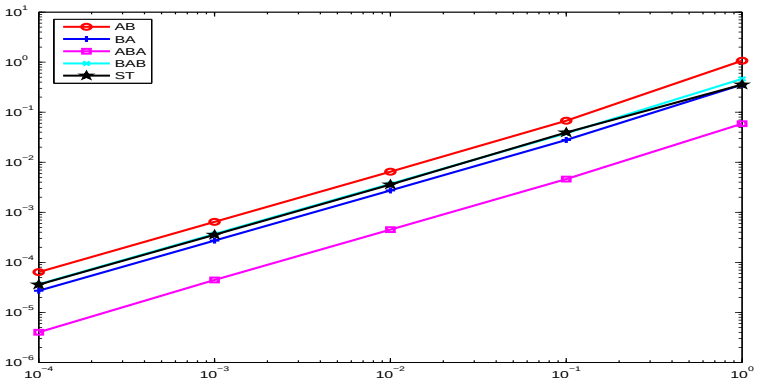

FIG. 3 – Erreurs relatives en norme l^2 pour les cinq méthodes proposées pour les pas de temps $h = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1} et 1

On retrouve l'ordre 1, mis en évidence dans l'exemple précédent. Cependant le comportement coordonnée par coordonnée est très différent de celui de l'exemple précédent. Représentons donc sur la figure 4 l'erreur relative sur chacune des composantes dans les coordonnées d'origine x_1, x_2 et x_3, pour les cinq schémas de splitting discutés, (L1), (L2), (S1), (S2) et (ST).

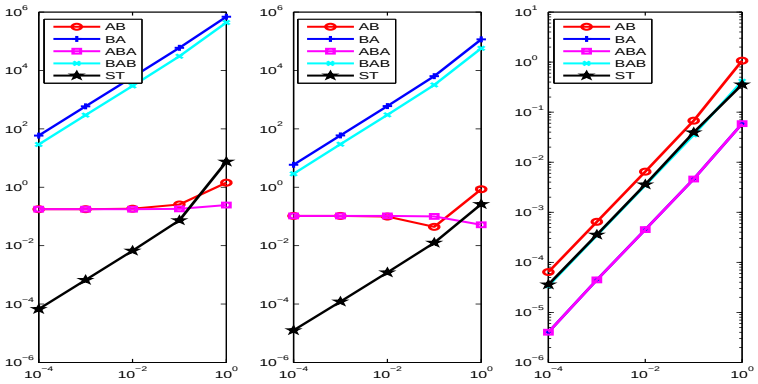

FIG. 4 – Erreurs relatives sur les coordonnées d'origine x_1, x_2 et x_3 pour les cinq méthodes proposées pour les pas de temps $h = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1} et 1

Le bloc très négatif correspondant aux deux premières composantes fait que ces coordonnées décroissent très rapidement. Au temps final $t = 1$, on a $x_1(1) \simeq -2.1\ 10^{-8}$, $x_2(1) \simeq 3.8\ 10^{-8}$ et $x_3 \simeq 5.0\ 10^{-2}$. À titre de comparaison, on avait $x_1(1) \simeq 2.8$, $x_2(1) \simeq 2.8$ et $x_3(1) \simeq 1.8$ pour l'exemple

1. Le comportement de l'erreur relative pour ces variables est ainsi comparable à celle pour des composantes raides, alors que c'est une superposition de composantes raides et non raides.

La figure 5 représente pour chacune des composantes dans les coordonnées propres y_1, y_2 et y_3, l'erreur relative due aux cinq schémas de splitting discutés, (L1), (L2), (S1), (S2) et (ST).

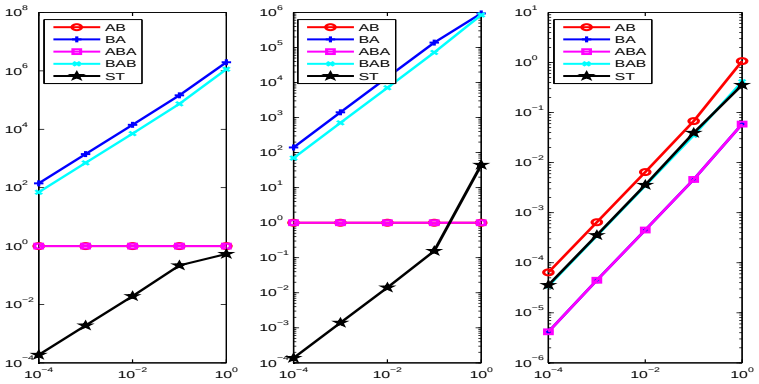

FIG. 5 – Erreurs relatives sur les coordonnées propres y_1, y_2 et y_3 pour les cinq méthodes proposées pour les pas de temps $h = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1} et 1

On a maintenant $y_1(1) \simeq 6.2\ 10^{-9}\ (1+i)$, $y_2(1) \simeq -2.6\ 10^{-9}\ (1+i)$ et $y_3(1) \simeq 4.9\ 10^{-2}$. Dans cette base propre, les conclusions sont les mêmes que pour l'exemple 1.

6 Approximation des systèmes raides

6.1 Choix du schéma d'approximation

Pour guider le choix de l'approximation de l'exponentielle nous allons traiter le cas très simple de l'équation scalaire suivante

$$\dot{x} = -\frac{1}{\varepsilon}x.$$

Rappelons qu'avec ce choix de signe, $x(t)$ décroît très rapidement vers 0. La solution exacte est $x(t) = \exp(-t/\varepsilon)x(0)$.

Si on applique le schéma d'Euler, on calcule

$$x^{n+1} = (1 - \frac{h}{\varepsilon})x^n.$$

La quantité $1 - h/\varepsilon$ a deux défauts majeurs : elle est très grande (en $O(h/\varepsilon)$) et elle est négative, ce qui fait changer x^n de signe à chaque itération alors que la solution exacte reste de signe constant. En bref, le schéma d'Euler est à proscrire absolument !

Si on applique le schéma d'Euler rétrograde, on calcule

$$x^{n+1} = \frac{1}{1 + \frac{h}{\varepsilon}}x^n.$$

La quantité $1/(1 + h/\varepsilon)$ est cette fois-ci petite, de taille $O(\varepsilon/h)$ et elle est positive. x^n va donc bien tendre vers 0 en $O((\varepsilon/h)^n)$. Ce n'est pas la bonne vitesse de convergence vers 0, c'est beaucoup trop lent, mais c'est infiniment mieux que le schéma d'Euler.

Appliquons maintenant le schéma de Crank–Nicolson. En l'absence de raideur, il est meilleur que les deux deux premiers dans la mesure où son ordre théorique est deux. Ici, on calcule

$$x^{n+1} = \frac{1 - \frac{h}{2\varepsilon}}{1 + \frac{h}{2\varepsilon}}x^n.$$

La quantité $(1-h/2\varepsilon)/(1+h/2\varepsilon)$ est de taille $O(1)$ quand ε tend vers 0, indépendemment de $h >> \varepsilon$. De plus, cette quantité est négative. On a donc $x^{n+1} \sim -x^n$. Le module est légèrement inférieur à 1, ce qui fait que l'on a une suite alternée qui tend vers 0 à une vitesse d'escargot, qui est de l'ordre $(1-\varepsilon/h)^n$:

$$\frac{1-\frac{h}{2\varepsilon}}{1+\frac{h}{2\varepsilon}} = -\frac{1-\frac{2\varepsilon}{h}}{1+\frac{2\varepsilon}{h}} = -(1-\frac{2\varepsilon}{h})(1-\frac{2\varepsilon}{h}+O((\frac{\varepsilon}{h})^2)) = -(1-\frac{\varepsilon}{h})+O((\frac{\varepsilon}{h})^2).$$

En conclusion, le seul schéma utilisable pour la partie raide parmi ceux que l'on a vus est le schéma d'Euler rétrograde.

On peut même dire un peu plus : tout schéma explicite donnera lieu à des solutions qui explosent quand $\varepsilon \to 0$. En effet, pour un schéma explicite, le propagateur sur un pas de temps h est un polynôme en h/ε. Son terme dominant domine, par définition, quand $\varepsilon \to 0$ et on a un propagateur infiniment grand (en module). Un exemple classique est le schéma de Runge–Kutta d'ordre 4, qui appliqué à l'équation $\dot{x} = Ax$ approche $\exp(tA)$ par son développement à l'ordre 4.

6.2 Approximation et splitting dans le cas scalaire raide

Regardons maintenant l'équation

$$\dot{x} = -\frac{1}{\varepsilon}x + ax.$$

Sa solution exacte décroît toujours très rapidement : $x(t) = \exp((a-1/\varepsilon)t)$. Nous allons appliquer différents schémas de splitting en tenant compte des informations obtenues précédemment, à savoir

- utiliser le schéma d'Euler rétrograde pour la partie raide,
- utiliser uniquement des méthodes de splitting qui finissent par une étape raide.

On est cependant dans le cadre commutatif (car scalaire) et on a déjà vu que pour la méthode de Lie, les deux splittings sont équivalents.

Les méthodes s'écrivent :

$$\begin{aligned}
\mathcal{L}_1\mathcal{RE}(h) &= (1+\frac{h}{\varepsilon})^{-1}(1+ha),\\
\mathcal{L}_1\mathcal{R}(h) &= (1+\frac{h}{\varepsilon})^{-1}(1-ha)^{-1},\\
\mathcal{L}_1\mathcal{RC}(h) &= (1+\frac{h}{\varepsilon})^{-1}(1+\frac{1}{2}ha)(1-\frac{1}{2}ha)^{-1},\\
\mathcal{S}_1\mathcal{RE}(h) &= (1+\frac{h}{2\varepsilon})^{-1}(1+ha)(1+\frac{h}{2\varepsilon})^{-1},\\
\mathcal{S}_1\mathcal{R}(h) &= (1+\frac{h}{2\varepsilon})^{-1}(1-ha)^{-1}(1+\frac{h}{2\varepsilon})^{-1},\\
\mathcal{S}_1\mathcal{RC}(h) &= (1+\frac{h}{2\varepsilon})^{-1}(1+\frac{1}{2}ha)(1-\frac{1}{2}ha)^{-1}(1+\frac{h}{2\varepsilon})^{-1},
\end{aligned}$$

Pour une bonne approximation du terme non raide, on choisit bien sûr h suffisamment petit pour que $1-ha > 0$. Sous cette hypothèse très raisonnable, toutes ces méthodes assurent la décroissance de l'unique variable et aucune n'assure la bonne vitesse de décroissance. Par rapport au cas matriciel, on n'a ici qu'une seule composante qui s'avère être raide et donc mal approchée. On ne peut pas étudier l'impact sur d'autres composantes non raides.

6.3 Et si on savait traiter l'exponentielle raide ...

Revenons à l'étude matricielle. Dans l'étude matricielle, on a avait supposé que la partie raide est diagonale. On sait alors toujours calculer l'exponentielle, ce qui affranchit de l'approximation de cette exponentielle qui s'avère difficile.

Pour la matrice B qui elle n'est pas diagonale (sinon les matrices commuteraient), on peut en revanche utiliser une méthode d'approximation, *a priori* au choix (Euler, Euler rétrograde, Crank–Nicolson). On peut s'amuser à refaire tous les calculs d'ordre mais cela est inutile. Comme B n'est pas raide, le pas de temps h choisit est toujours tel que $I - hB$ est inversible.

Toutes ces méthodes vont être stables et au moins d'ordre 1. Ce sont les seules propriété utilisées pour effectuer les démonstrations, qui sont du même tonneau que celle effectuées quand on supposait que l'on calculait toutes les exponentielles.

Comme il y a une perte d'ordre, rien ne sert d'utiliser une méthode d'ordre 2 comme la méthode de Crank–Nicolson.

7 Généralités sur les équations aux dérivées partielles

Souvent les méthodes de splitting sont appelées méthode des directions alternées (ADI, Alternate Directions Implicit, en anglais). Cette dénomination n'a pas vraiment de sens si on considère les exemples que l'on a vus jusqu'à maintenant. Elle prend son origine dans l'application initiale par Peacemann et Rachford d'une part et par Douglas de la méthode : la résolution de l'équation de la chaleur en décomposant le laplacien en sommes.

Il est impossible de donner un panorama complet des différentes situations pour les EDP, de part leur grande diversité. Ce sujet est par ailleurs l'objet de recherches importantes actuellement. Le sujet n'est donc pas clos.

Nous allons d'abord donner un cadre favorable dans lequel la formule de Trotter–Kato est valide. Dans ce contexte, tout marche comme dans le cas matriciel.

Par rapport au cas matriciel, des problèmes supplémentaires peuvent néanmoins se poser, liés aux estimations sur les semi-groupes d'évolution ou aux parties non linéaires. Nous verrons ces deux points sur un exemple.

Les sources de raideur sont toujours les différentes échelles en temps, mais aussi les forts gradients en espace des solutions. Ce dernier point est l'objet de recherches actuelles actives (autour de Stéphane Descombes et Marc Massot).

Un autre problème est la stabilité pour laquelle nous donnerons des définitions et que nous illustrerons par un exemple.

7.1 Formule de Trotter–Kato

Lemme 1 (Trotter–Kato)

Soient $M \geq 1$ et $\omega > 0$. On suppose que l'on a une suite de générateurs infinitésimaux A_n de semi-groupes uniformément continus $\mathcal{A}_n(t)$ qui satisfont $\|\mathcal{A}_n(t)\| \leq M \exp(\omega t)$. On suppose qu'il existe $\lambda_0 \in \mathbb{C}$ tel que $\mathrm{Re}(\lambda_0) > \omega$ et pour tout $x \in X$, $(\lambda_0 \mathrm{I} - A_n)^{-1} x \longrightarrow R(\lambda_0)x$, où l'image de $R(\lambda_0)$ est dense dans X.
Alors il existe un générateur de groupe infinitésimal A d'un semi-groupe uniformément continu $\mathcal{A}(t)$ qui satisfait $\|\mathcal{A}(t)\| \leq M \exp(\omega t)$ tel que $R(\lambda_0) = (\lambda_0 \mathrm{I} - A)^{-1}$ et $\mathcal{A}_n(t)x \longrightarrow \mathcal{A}(t)x$ pour tout $x \in X$. Cette limite es uniforme sur tout intervalle borné en temps.

L'application à notre cadre donne typiquement le résultat :

$$\lim_{n\to\infty} \left(\mathcal{A}(\frac{t}{n}) \mathcal{B}(\frac{t}{n}) \right)^n \equiv \lim_{n\to\infty} \left(\exp(\frac{t}{n} A) \exp(\frac{t}{n} B) \right)^n = \exp(t(A+B)) \equiv \mathcal{S}(t).$$

7.2 Notions de stabilité

On considère l'EDP suivante

$$\frac{du}{dt} = Au + Bu, \qquad u(t=0) = u^0 \in H,$$

Pour la résoudre on utilise un schéma donné par son semi-groupe $f(h)$. Dans la démonstration de convergence que nous avons déjà donnée dans le cadre matriciel, on avait vu qu'il fallait de la stabilité en plus de la consistance (ordre local au moins égal à 2).

Classiquement, on dira que le schéma est **fortement stable** si il existe une constante C tel que pour h petit,

$$\|f(h)\|_{\mathcal{B}(H)} \leq 1 + Ch,$$

où $\|\cdot\|_{\mathcal{B}(H)}$ est la norme des opérateurs linéaires bornés sur l'espace de Hilbert H :

$$\|A\|_{\mathcal{B}(H)} = \sup_{u\neq 0} \frac{\|Au\|_H}{\|u\|_H}.$$

Nous allons nous intéresser ci-dessous à la stabilité de l'approximation de Peaceman–Rachford :

$$\mathcal{P}(h) = (\mathrm{I} - \frac{1}{2}hA)^{-1}(\mathrm{I} + \frac{1}{2}hB)(\mathrm{I} - \frac{1}{2}hB)^{-1}(\mathrm{I} + \frac{1}{2}hA).$$

Nous allons montrer dans un premier temps que cette approximation est **faiblement stable** au sens où pour tout $u \in D(A)$, on a

$$\|\mathcal{P}(\frac{t}{n})^n u\|_H \leq \|u\|_H + \tau\|Au\|_H,$$

pour tout n tel que $t/n \leq \tau$. Il y a en quelque sens une perte de régularité.

Nous allons ensuite donner un contre-exemple à la stabilité à base de laplaciens, à savoir des opérateurs A et B pour lesquels

$$\|\mathcal{P}(h)\|_{\mathcal{B}(H)} \geq 2.$$

Le résultat général dont la démonstration dépasse de loin le cadre de ce cours est que, sous certaines hypothèses supplémentaires (voir Appendice B), on a uniquement

$$\|\mathcal{P}(h)\|_{\mathcal{B}(H)} \leq 1 + C\sqrt{h}.$$

7.3 Stabilité faible, avec perte de régularité

On suppose que A et B sont des opérateurs négatifs auto-adjoints. Pour $h > 0$, on définit sur $D(A)$ la norme

$$\|u\|_h = \|(\mathrm{I} - \frac{1}{2}hA)u\|_H.$$

Lemme 2

Pour tout $h > 0$ et tout $u \in D(A)$, $\mathcal{P}(h)A \in D(A)$ et

$$\|\mathcal{P}(h)u\|_h \leq \|u\|_h.$$

Preuve :
Si $u \in D(A)$ alors $(\mathrm{I} + hA/2)u \in H$. Par ailleurs, comme B est négatif auto-adjoint $(\mathrm{I} + hB/2)(\mathrm{I} - hB/2)^{-1}$ est une contraction dans H. Enfin $(\mathrm{I} - hA/2)^{-1}$ envoie continuement H dans $D(A)$. Ainsi $\mathcal{P}(h)$ est bien dans $D(A)$ et en outre

$$\begin{aligned} \|\mathcal{P}(h)u\|_h &\leq \|(\mathrm{I} + \frac{1}{2}hB)(\mathrm{I} - \frac{1}{2}hB)^{-1}(\mathrm{I} + \frac{1}{2}hA)(\mathrm{I} - \frac{1}{2}hA)^{-1}(\mathrm{I} - \frac{1}{2}hA)u\|_H \\ &\leq \|u\|_h. \end{aligned}$$

■

En particulier, on peut appliquer ce lemme à $h = t/n$:

$$\|\mathcal{P}(\frac{t}{n})u\|_{\frac{t}{n}} \leq \|u\|_{\frac{t}{n}}.$$

Comme A est un opérateur négatif, on a clairement

$$\|u\|_{t/n} = \|(\mathrm{I} - \frac{t}{2n}A)u\|_H \geq \|u\|_H.$$

Par ailleurs pour $t/n \leq \tau$, on a également clairement

$$\|u\|_{t/n} = \|(\mathrm{I} - \frac{t}{2n}A)u\|_H \leq \|u\|_H + \tau\|Au\|_H.$$

Ainsi

$$\|\mathcal{P}(\frac{t}{n})u\|_H \leq \|\mathcal{P}(\frac{t}{n})u\|_{\frac{t}{n}} \leq \|u\|_{\frac{t}{n}} \leq \|u\|_H + \tau\|Au\|_H.$$

Ceci est la formule de stabilité faible.

7.4 L'exemple avec des laplaciens

On se place dans l'espace fonctionnel $H = L^2(\mathbb{R}^d)^2$ et on considère les opérateurs A et B sont le forme $A = M\Delta$ et $B = N\Delta$, où M et N sont des matrices symétriques définies positives. Les opérateurs A et B sont bien négatifs auto-adjoints. C'est *a priori* un cadre sympathique de type équation de la chaleur, réputé pour être bien posé. En particulier on a $D(A) = D(B) = H^2(\mathbb{R}^d)^2$.

Néanmoins, on a le résultat suivant.

Théorème 6

Pour tout $K > 0$, il existe un choix des matrices M et N tel que

$$\|\mathcal{P}(t)\|_{\mathcal{L}(H)} \geq K.$$

Comme ce théorème est vrai pour des $K > 1$, cela implique que le schéma de Peaceman–Rachford n'est pas fortement stable. Dans ce théorème, la norme utilisée $\|\cdot\|_{\mathcal{L}(H)}$ est celle des opérateurs linéaires sur l'espace de Hilbert H, vus comme des multiplicateurs de Fourier. On utilise la norme subordonnée L^2 dans cet espace :

$$\|A\|_{\mathcal{L}(H)} = \sup_{\xi\in\mathbb{R}^d} \frac{\|\widehat{Au}(\xi)\|_H}{\|\hat{u}(\xi)\|_H}.$$

La transformée de Fourier est ici définie par

$$\hat{u}(\xi) = \int_{\mathbb{R}^d} u(x)\exp(-2i\pi x\cdot\xi)\ dx.$$

Preuve :
Pour la formule de Peaceman–Rachford, on a

$$\begin{aligned}\|\mathcal{P}(t)\|_{\mathcal{L}(H)} &= \sup_{\xi\in\mathbb{R}^d} \|(\mathrm{I} + \frac{t}{2}|\xi^2|M)^{-1}(\mathrm{I} - \frac{t}{2}|\xi^2|N)(\mathrm{I} + \frac{t}{2}|\xi^2|N)^{-1}(\mathrm{I} - \frac{t}{2}|\xi^2|M)\|_{\mathcal{B}(\mathbb{R}^2)} \\ &= \sup_{r\in\mathbb{R}} \|(\mathrm{I} + rM)^{-1}(\mathrm{I} - rN)(\mathrm{I} + rN)^{-1}(\mathrm{I} - rM)\|_{\mathcal{B}(\mathbb{R}^2)}.\end{aligned}$$

En particulier, on remarque que $\|\mathcal{P}(t)\|_{\mathcal{L}(H)}$ ne dépend pas du temps t. On choisit les formes de matrice suivantes (symétriques définies positives)

$$M = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \text{ avec } a > b > 0, \qquad N = \begin{pmatrix} r^{-2} & 0 \\ 0 & 1 \end{pmatrix}.$$

On aura alors

$$\|\mathcal{P}(t)\|_{\mathcal{L}(H)} \geq \|(\mathrm{I}+rM)^{-1}(\mathrm{I}-rN)(\mathrm{I}+rN)^{-1}(\mathrm{I}-rM)\|_{\mathcal{B}(\mathbb{R}^2)}.$$

Nous allons faire tendre r vers $+\infty$. Nous supposons donc qu'il est grand et écrivons des développements limités en r^{-1}. Calculons tout d'abord

$$\begin{aligned}(\mathrm{I}-rN)(\mathrm{I}+rN)^{-1} &= 2(\mathrm{I}+rN)^{-1}-\mathrm{I} = 2\begin{pmatrix} 1+r^{-1} & 0 \\ 0 & 1+r \end{pmatrix}^{-1} - \mathrm{I} \\ &= \begin{pmatrix} \frac{1-r^{-1}}{1+r^{-1}} & 0 \\ 0 & \frac{1-r}{1+r} \end{pmatrix} = \begin{pmatrix} \frac{1-r^{-1}}{1+r^{-1}} & 0 \\ 0 & -\frac{1-r^{-1}}{1+r^{-1}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + O(r^{-1}).\end{aligned}$$

Par ailleurs

$$\begin{aligned}(\mathrm{I}+rM)^{-1} &= \begin{pmatrix} 1+ra & rb \\ rb & 1+ra \end{pmatrix}^{-1} = \frac{1}{r}\begin{pmatrix} a+r^{-1} & b \\ b & a+r^{-1} \end{pmatrix}^{-1} \\ &= \frac{1}{r}\frac{1}{a^2-b^2+2r^{-1}+r^{-2}}\begin{pmatrix} a+r^{-1} & -b \\ -b & a+r^{-1} \end{pmatrix} \\ &= \frac{1}{r}\left\{\frac{1}{a^2-b^2}\begin{pmatrix} a & -b \\ -b & a \end{pmatrix} + O(r^{-1})\right\}, \\ (\mathrm{I}-rM) &= -r\begin{pmatrix} a-r^{-1} & b \\ b & a-r^{-1} \end{pmatrix} = -r\left\{\begin{pmatrix} a & b \\ b & a \end{pmatrix} + O(r^1)\right\}\end{aligned}$$

Ainsi

$$\begin{aligned}(\mathrm{I}+rM)^{-1}(\mathrm{I}-rN)(\mathrm{I}+rN)^{-1}(\mathrm{I}-rM) &= -\frac{1}{a^2-b^2}\begin{pmatrix} a & -b \\ -b & a \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} a & b \\ b & a \end{pmatrix} + O(r^{-1}) \\ &= -\frac{1}{a^2-b^2}\begin{pmatrix} a^2+b^2 & 2ab \\ -2ab & -(a^2+b^2) \end{pmatrix} + O(r^{-1})\end{aligned}$$

Calculer la norme $\mathcal{B}(\mathbb{R}^2)$, c'est calculer la norme subordonnée L^∞ :

$$\|A\|_\infty = \max_j \sum_k |a_{jk}|.$$

On obtient ainsi

$$\|(\mathrm{I}+rM)^{-1}(\mathrm{I}-rN)(\mathrm{I}+rN)^{-1}(\mathrm{I}-rM)\|_{\mathcal{B}(\mathbb{R}^2)} = \frac{a+b}{a-b} + O(r^{-1}).$$

On peut choisir a et b de telle façon que $(a+b)/(a-b) \geq K+1$. Pour r suffisamment grand, on aura bien $\|\mathcal{P}(t)\|_{\mathcal{L}(H)} \geq K$. ∎

7.5 Le même exemple sans laplacien

Le problème tient vraiment au fait que l'on a une EDP non scalaire. Le commutateur $[A,B]$ est un opérateur différentiel d'ordre 4. Le problème ne se pose plus si on travaille dans $L^2(\mathbb{R}^d)^2$ ou pour une EDO.

En effet, si on considère le problème

$$\dot{x} = Mx + Nx, \qquad x(0) = x^0 \in \mathbb{R}^2,$$

Il faut cette fois-ci calculer

$$\|\mathcal{P}(t)\|_{\mathcal{L}(H)} = \|(\mathrm{I}-\frac{t}{2}M)^{-1}(\mathrm{I}+\frac{t}{2}N)(\mathrm{I}-\frac{t}{2}N)^{-1}(\mathrm{I}+\frac{t}{2}M)\|_{\mathcal{B}(\mathbb{R}^2)}.$$

Maintenant, il suffit de faire des calculs valables pour $\tau = t/2$ petit. On n'utilise plus les équivalents. On a

$$(\mathrm{I}+\tau N)(\mathrm{I}-\tau N)^{-1} = \begin{pmatrix} \frac{1+\tau^{-1}}{1-\tau^{-1}} & 0 \\ 0 & -\frac{1+\tau^{-1}}{1-\tau^{-1}} \end{pmatrix} = \frac{1+\tau}{1-\tau}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Par ailleurs

$$\begin{aligned}(\mathrm{I}-\tau M)^{-1} &= -\frac{1}{\tau}\frac{1}{a^2-b^2-2\tau^{-1}+\tau^{-2}}\begin{pmatrix} a-\tau^{-1} & -b \\ -b & a-\tau^{-1}\end{pmatrix} \\ &= \frac{1}{1-2\tau+a^2\tau^2-b^2\tau^2}\begin{pmatrix} 1-a\tau & b\tau \\ b\tau & 1-a\tau\end{pmatrix},\end{aligned}$$

et

$$(\mathrm{I}+\tau M) = \begin{pmatrix} 1+a\tau & b\tau \\ b\tau & 1+a\tau\end{pmatrix}.$$

Ainsi

$$\begin{aligned}&(\mathrm{I}-\tau M)^{-1}(\mathrm{I}+\tau N)(\mathrm{I}-\tau N)^{-1}(\mathrm{I}+\tau M) \\ &\qquad = -\frac{1+\tau}{1-\tau}\frac{1}{1-2\tau+a^2\tau^2-b^2\tau^2}\begin{pmatrix} 1-(a^2+b^2)\tau^2 & -2ab\tau^2 \\ 2ab\tau^2 & -(1-(a^2+b^2)\tau^2)\end{pmatrix}.\end{aligned}$$

On calcule la norme subordonnée L^∞ pour obtenir

$$\|(\mathrm{I}-\tau M)^{-1}(\mathrm{I}+\tau N)(\mathrm{I}-\tau N)^{-1}(\mathrm{I}+\tau M)\|_{\mathcal{B}(\mathbb{R}^2)} = \frac{1+\tau}{1-\tau}\frac{1-(a+b)^2\tau^2}{1-2\tau+a^2\tau^2-b^2\tau^2}.$$

Il n'y a plus de problème de stabilité.

8 Un exemple d'EDP non linéaire

8.1 Contexte

On considère l'équation de Schrödinger non linéaire :

$$\begin{cases}\partial_t u + i\Delta u - F(u) = 0, & x\in\mathbb{R}^2,\ t>0,\\ u(x,0)=u^0(x), & x\in\mathbb{R}^2.\end{cases}$$

où F est K-lipschitizienne, $F(0)=0$ et ses quatre premières dérivées sont bornées. On notera M' et M'' le maximum de F' et F'' respectivement.

On appelle $\mathcal{S}(t)$ le flot associé à cette équation. On définit deux flots partiels, celui, $\mathcal{X}(t)$, de l'équation de Schrödinger linéaire, qui est un groupe linéaire (l'équation est bien posée en rétrograde)

$$\begin{cases}\partial_t u + i\Delta u = 0, & x\in\mathbb{R}^2,\ t>0,\\ u(x,0)=u^0(x), & x\in\mathbb{R}^2,\end{cases}$$

et celui, $\mathcal{Y}(t)$, de l'équation différentielle ordinaire

$$\begin{cases}\partial_t u - F(u) = 0, & x\in\mathbb{R}^2,\ t>0,\\ u(x,0)=u^0(x), & x\in\mathbb{R}^2.\end{cases}$$

Comme précédemment, on appelle $f(t)$ le flot associé à un splitting.

Dans ce cas précis, on peut résoudre efficacement le système splitté quelle que soit la méthode choisie. En effet, si on considère des conditions aux bords périodiques, on peut utiliser la transformée de Fourier rapide (FFT) pour résoudre le flot $\mathcal{X}(t)$. Par ailleurs, l'équation non linéaire peut être résolue point par point.

Les applications concernent en outre des non linéarités que l'on sait intégrer exactement. C'est le cas de $F(u)=i\alpha|u|^2u$ avec $\alpha=\pm1$, qui donne l'équation couramment appelée Schrödinger non linéaire cubique. En effet, $\partial_t u - i\alpha|u|^2u = 0$ implique $\partial_t|u|^2=0$. On a donc $\partial_t u - i\alpha|u^0|^2u=0$ et

$u(t) = \exp(i|\alpha|u^0|^2 t)u^0$.

Pour cette non linéarité, en dimension 2, la dimension que nous considérons, et pour $\alpha = +1$, dit cas défocalisant, l'équation est globalement bien posée dans $H^1(\mathbb{R}^2)$ par exemple. En revanche, pour $\alpha = -1$, dit cas focalisant, il existe des données initiales $u^0 \in H^1(\mathbb{R}^2)$ qui évoluent en une solution qui explose en temps fini $\|u(T)\|_{H^1} = +\infty$. Néanmoins, si on se place à une certaine distance de ce temps d'explosion, on peut supposer que la non linéarité est une fonction lipschitzienne et vérifie toutes les propriétés que nous avons supposées sur F.

La fin de ce cours est dédiée à la preuve du théorème suivant.

Pour tout u^0 dans $H^2(\mathbb{R}^2)$ et tout $T > 0$, il existe C et h_0 tels que pour tout $h \in]0, h_0]$, et pour tout n tel que $nh \leq T$ pour un splitting de Lie

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq C(\|u^0\|_{H^2})h\|u^0\|_{H^2}.$$

Si de plus $u^0 \in H^4(\mathbb{R}^2)$, alors pour un splitting de Strang

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq C(\|u^0\|_{H^4})h^2\|u^0\|_{H^4}.$$

8.2 Quelques estimations

8.2.1 Lemmes de Gronwall

On utilisera deux versions du lemme de Gronwall.

Lemme 3 (Gronwall modifié)

Soit P un polynôme à coefficients positifs et sans terme constant ($P(0) = 0$). On suppose que pour la fonction ϕ il existe une constante $C \geq 0$ telle que, pour tout temps $t \geq 0$, on a l'inéquation intégrale

$$0 \leq \phi(t) \leq \phi(0) + P(t) + C\int_0^t \phi(\tau)d\tau.$$

Alors, pour tout $\alpha > 1$, il existe $t_0 > 0$ tel que, pour tout $0 \leq t \leq t_0$, on a la majoration

$$\phi(t) \leq \phi(0)e^{Ct} + \alpha P(t).$$

Preuve :
On définit une nouvelle fonction positive

$$\psi(t) = \left(\phi(0) + P(t) + C\int_0^t \phi(\tau)d\tau\right)e^{-Ct}$$

pour laquelle on cherche à déterminer une inéquation intégrale. Tout d'abord

$$\psi'(t) = \left(P'(t) + C\phi(t) - C\left(\phi(0) + P(t) + C\int_0^t \phi(\tau)d\tau\right)\right)e^{-Ct} \leq P'(t)e^{-Ct}.$$

En intégrant, et comme $\psi(0) = \phi(0)$, on obtient la majoration

$$\psi(t) - \psi(0) = \psi(t) - \phi(0) \leq \int_0^t P'(\tau)e^{-C\tau}d\tau.$$

Comme les coefficients du polynôme P sont positifs, pour tout τ, $P'(\tau)$ est également positif et

$$\phi(t) \leq \psi(t)e^{Ct} \leq \phi(0)e^{Ct} + \int_0^t P'(\tau)e^{C(t-\tau)}d\tau \leq \phi(0)e^{Ct} + e^{Ct_0}\int_0^t P'(\tau)d\tau.$$

On peut choisir t_0 suffisamment petit pour que $e^{Ct_0} \leq \alpha$. Alors, pour tout $0 \leq t \leq t_0$,

$$\phi(t) \leq \phi(0)e^{Ct} + \alpha P(t).$$

■

Lemme 4 (Gronwall classique)

Soit p une fonction positive. On suppose que la fonction ϕ vérifie pour tout temps $t \geq 0$ l'inéquation intégrale

$$0 \leq \phi(t) \leq \phi(0) + \int_0^t p(\tau)\phi(\tau)d\tau.$$

Alors pour tout temps $t \geq 0$

$$\phi(t) \leq \phi(0)\exp(\int_0^t p(\tau)d\tau).$$

Preuve :
On pose

$$\psi(t) = [\phi(0) + \int_0^t p(\tau)\phi(\tau)d\tau]\exp(-\int_0^t p(\tau)d\tau).$$

En calculant la dérivée, et en utilisant la majoration hypothèse, on obtient que $\psi'(t) \leq 0$. On a donc $\psi(t) \leq \psi(0) = \phi(0)$ pour tout $t \geq 0$. Finalement

$$\phi(t) \leq \psi(t)\exp(\int_0^t p(\tau)d\tau) \leq \phi(0)\exp(\int_0^t p(\tau)d\tau).$$

■

8.2.2 Estimations du flot de Schrödinger $X(t)$

Lemme 5

1. *Le groupe $\mathcal{X}(t)$ est unitaire dans tous les espaces de Sobolev classiques $H^s(\mathbb{R}^2)$, $s \in \mathbb{R}$. Pour tout $w \in H^s(\mathbb{R}^2)$ et tout $t \geq 0$,*

   (X1) $\|\mathcal{X}(t)w\|_{H^s} = \|w\|_{H^s}$.

2. *Pour tout $w \in H^2(\mathbb{R}^2)$ et tout $t \geq 0$,*

   (X2) $\|\mathcal{X}(t)w - w\|_{L^2} \leq t\|w\|_{H^2}$.

3. *Pour tout $w \in H^4(\mathbb{R}^2)$ et tout $t \geq 0$,*

   (X3) $\|\mathcal{X}(t)w - w\|_{H^2} \leq t\|w\|_{H^4}$.

4. *On se donne un temps $T > 0$. Il existe une constante C telle que pour tout $w \in \mathcal{C}^1([0,T];H^2) \cap L^\infty([0,T],H^4)$ et tout $0 \leq t \leq T$,*

   (X4) $\| \int_0^t (\mathcal{X}(t-s)w(s) - \mathcal{X}(\frac{t}{2})w(s))ds\|_{L^2} \leq Ct^3(\|w\|_{\mathcal{C}^1([0,T];H^2)} + \|w\|_{L^\infty([0,T],H^4)})$.

5. *Il existe une constante C telle que pour tout $w \in H^4(\mathbb{R}^2)$,*

   (X5) $\|\mathcal{X}(\frac{t}{2})w - \frac{1}{2}\mathcal{X}(t)w - \frac{1}{2}w\|_{L^2} \leq Ct^2\|w\|_{H^4}$.

Les estimations découlent du fait que $\dot{\mathcal{X}}(t) = i\Delta\mathcal{X}(t) = i\mathcal{X}(t)\Delta$.

Preuve :

1. Si u est solution de l'équation de Schrödinger non linéaire, on a dans le domaine de Fourier

   $$\partial_t \hat{u}(\xi) - i|\xi|^2 \hat{u}(\xi) = 0.$$

   Ainsi $\hat{u}(\xi)(t) = \exp(i|\xi|^2 t)\hat{u}^0(\xi)$ et en particulier $|\hat{u}(\xi)(t)| = |\hat{u}^0(\xi)|$ pour tout $\xi \in \mathbb{R}^2$. Avec la notation du flot, cela donne $|\widehat{\mathcal{X}(t)u}(\xi)| = |\hat{u}(\xi)|$. Si on calcule la norme H^s, on a

   $$\|u\|_{H^s}^2 = \int_{\mathbb{R}^2} (1+|\xi|^{2s})|\hat{u}(\xi)|^2 d\xi = \int_{\mathbb{R}^2} (1+|\xi|^{2s})|\widehat{\mathcal{X}(t)u}(\xi)|^2 d\xi = \|\mathcal{X}(t)u\|_{H^s}^2.$$

2. Pour $w \in H^2(\mathbb{R}^2)$,

   $$\|\mathcal{X}(t)w - w\|_{L^2} = \|\int_0^t \dot{\mathcal{X}}(\tau)w d\tau\|_{L^2} = \|\int_0^t \mathcal{X}(\tau)\Delta w d\tau\|_{L^2} \leq \int_0^t \|\Delta w\|_{L^2} d\tau = t\|w\|_{H^2}.$$

3. Pour $w \in H^4(\mathbb{R}^2)$, on fait la même estimation, mais dans $H^2(\mathbb{R}^2)$ au lieu de $L^2(\mathbb{R}^2)$.

4. On effectue un développement de Taylor avec reste intégral autour de $t/2$:

   $$\mathcal{X}(t-\tau) - \mathcal{X}(\frac{t}{2}) = (\frac{t}{2} - \tau)\dot{\mathcal{X}}(\frac{t}{2}) + \int_{\frac{t}{2}}^{t-\tau} (t-\tau-\sigma)\ddot{\mathcal{X}}(\sigma)d\sigma.$$

   Si on applique ceci à une fonction $w \in \mathcal{C}^1([0,T];H^2) \cap L^\infty([0,T],H^4)$

   $$(\mathcal{X}(t-\tau) - \mathcal{X}(\frac{t}{2}))w(\tau) = i(\frac{t}{2} - \tau)\mathcal{X}(\frac{t}{2})\Delta w(\tau) - \int_{\frac{t}{2}}^{t-\tau} (t-\tau-\sigma)\mathcal{X}(\sigma)\Delta^2 w(\tau)d\sigma.$$

   Pour obtenir les meilleurs estimations, on a intérêt à symétriser le plus possible les expressions. En particulier, on remarque que

   $$\int_0^t (\frac{t}{2} - \tau)\Delta w(\tau)d\tau = \int_0^{\frac{t}{2}} (\frac{t}{2} - \tau)[\Delta w(\tau) - \Delta w(t-\tau)]d\tau.$$

   Ainsi en utilisant l'unitarité (X1) de $\mathcal{X}(t/2)$ dans L^2, on a

   $$\begin{aligned}\|\int_0^t (\mathcal{X}(t-\tau)w(\tau) - \mathcal{X}(\frac{t}{2})w(\tau))d\tau\|_{L^2} \quad &\leq \quad \|\int_0^{\frac{t}{2}} (\frac{t}{2} - \tau)[\Delta w(\tau) - \Delta w(t-\tau)]d\tau\|_{L^2} \\ &\quad + \|\int_0^t \int_{\frac{t}{2}}^{t-\tau} (t-\tau-\sigma)\mathcal{X}(\sigma)\Delta^2 w(\tau)d\sigma d\tau\|_{L^2}.\end{aligned}$$

   La constante de Lipschitz de l'application $\tau \mapsto \Delta w(\tau)$ est majorée par la norme $\|w\|_{\mathcal{C}^1([0,T],H^2)}$. De plus, $\mathcal{X}(\sigma)$ est unitaire (X1) dans L^2 donc

   $$\begin{aligned}&\|\int_0^t (\mathcal{X}(t-\tau)w(\tau) - \mathcal{X}(\frac{t}{2})w(\tau))d\tau\|_{L^2} \\ &\qquad \leq 2\|w\|_{\mathcal{C}^1([0,T],H^2)} \int_0^{\frac{t}{2}} (\frac{t}{2} - \tau)^2 d\tau + \|w\|_{L^\infty([0,T],H^4)} \int_0^t \int_{\frac{t}{2}}^{t-\tau} (t-\tau-\sigma)d\sigma \; d\tau \\ &\qquad \leq Ct^3(\|w\|_{\mathcal{C}^1([0,T],H^2)} + \|w\|_{L^\infty([0,T],H^4)}).\end{aligned}$$

5. De la même façon, on peut combiner deux développements de Taylor avec reste intégral autour de $t/2$, pour obtenir

   $$\mathcal{X}(\frac{t}{2}) - \frac{1}{2}\mathcal{X}(t) - \frac{1}{2}\mathcal{X}(0) = -\frac{1}{2}\int_0^{\frac{t}{2}} \sigma(\ddot{\mathcal{X}}(\sigma) + \ddot{\mathcal{X}}(t-\sigma))d\sigma.$$

   Si on applique ceci à une fonction $L^\infty([0,T],H^4)$, on a

   $$(\mathcal{X}(\frac{t}{2}) - \frac{1}{2}\mathcal{X}(t) - \frac{1}{2}\mathcal{X}(0))w(\tau) = \frac{1}{2}\int_0^{\frac{t}{2}} \sigma(\mathcal{X}(\sigma) + \mathcal{X}(t-\sigma))\Delta^2 w(\tau)d\sigma.$$

   À nouveau, $\mathcal{X}(\sigma)$ et $\mathcal{X}(t-\sigma)$ sont unitaires (X1) dans L^2 donc il existe une constante C telle que

   $$\|\mathcal{X}(\frac{t}{2})w - \frac{1}{2}\mathcal{X}(t)w - \frac{1}{2}w\|_{L^2} \leq Ct^2\|w\|_{H^4}.$$

■

8.2.3 Estimations du flot non linéaire $\mathcal{Y}(t)$

Si u est solution de l'équation différentielle non linéaire, on a $\partial_t u = F(u)$, ce qui s'intègre en

$$u(t) = u(0) + \int_0^t F(u(\tau))d\tau.$$

Ceci permet une réécriture du flot non linéaire $\mathcal{Y}(t)$ sous forme intégrale :

$$\mathcal{Y}(t)w = w + \int_0^t F(\mathcal{Y}(\tau)w)d\tau.$$

Lemme 6

1. *Soit $w \in H^2(\mathbb{R}^2)$. Alors il existe une constante C qui ne dépend que de $\|w\|_{L^\infty}$ (cette norme est finie car on a une injection continue de $H^2(\mathbb{R}^2)$ dans $L^\infty(\mathbb{R}^2)$) telle que pour tout $0 \leq t \leq 1$*

   (Y1) $\|\mathcal{Y}(t)w\|_{L^2} \leq e^{Kt}\|w\|_{L^2}$
   (Y2) $\|\mathcal{Y}(t)w\|_{H^2} \leq C\|w\|_{H^2}$.

2. *Si de plus $w \in H^4(\mathbb{R}^2)$, alors il existe une constante C qui ne dépend que de $\|w\|_{L^\infty}$ telle que $0 \leq t \leq 1$*

   (Y3) $\|\mathcal{Y}(t)w\|_{H^4} \leq C\|w\|_{H^4}$.

3. *Pour w_1, $w_2 \in L^2(\mathbb{R}^2)$, il existe une constante C qui ne dépend que de F telle que pour tout $0 \leq t \leq 1$*

   (Y4) $\|\mathcal{Y}(t)w_1 - \mathcal{Y}(t)w_2\|_{L^2} \leq (1 + Ct)\|w_1 - w_2\|_{L^2}$.

4. *Soit $w \in L^2(\mathbb{R}^2)$. Alors, il existe une constante C qui ne dépend que de F telle que pour tout $0 \leq t \leq 1$,*

   (Y5) $\|\mathcal{Y}(t)w - w\|_{L^2} \leq Ct\|w\|_{L^2}$.

Preuve :

1. Comme F est K-lipschitzienne et $F(0) = 0$, on a, pour tout $v \in L^\infty$,

   $$\|F(v)\|_{L^\infty} = \|F(v) - F(0)\|_{L^\infty} \leq K\|v\|_{L^\infty}.$$

   La formulation intégrale fournit tout d'abord l'estimation L^∞ :

   $$\|\mathcal{Y}(t)w\|_{L^\infty} \leq \|w\|_{L^\infty} + K\int_0^t \|\mathcal{Y}(\tau)w\|_{L^\infty}\, d\tau,$$

   ce qui par le lemme de Gronwall classique donne

   $$\|\mathcal{Y}(t)w\|_{L^\infty} \leq e^{Kt}\|w\|_{L^\infty}.$$

   En partant de la même formulation intégrale, et comme on a également $\|F(v)\|_{L^2} \leq K\|v\|_{L^2}$, on obtient l'estimation L^2 :

   $$\|\mathcal{Y}(t)w\|_{L^2} \leq \|w\|_{L^2} + K\int_0^t \|\mathcal{Y}(\tau)w\|_{L^2}d\tau.$$

   Cette estimation nous donne immédiatement l'estimation (Y1) pour tout temps $t \geq 0$. Soit D un opérateur différentiel du premier ordre, on a

   $$D\mathcal{Y}(t)w = Dw + \int_0^t F'(\mathcal{Y}(\tau)w)D(\mathcal{Y}(\tau)w)d\tau.$$

Rappelons que F' est majorée, ce qui donne

$$\|D\mathcal{Y}(t)w\|_{L^2} \leq \|Dw\|_{L^2} + M' \int_0^t \|D\mathcal{Y}(\tau)w\|_{L^2} d\tau.$$

On différentie une fois de plus pour obtenir

$$\Delta\mathcal{Y}(t)w = \Delta w + \int_0^t [F''(\mathcal{Y}(\tau)w)D(\mathcal{Y}(\tau)w)^2 + F'(\mathcal{Y}(\tau)w)\Delta\mathcal{Y}(\tau)w]d\tau.$$

On a également supposé que F'' est majorée, ainsi

$$\|\Delta\mathcal{Y}(t)w\|_{L^2} \leq \|\Delta w\|_{L^2} + \int_0^t [M''\|D\mathcal{Y}(\tau)w\|_{L^2}^2 + M'\|\Delta\mathcal{Y}(\tau)w\|_{L^2}]d\tau.$$

Rappelons l'inégalité de Gagliardo–Nirenberg :

$$\|D\mathcal{Y}(\tau)w\|_{L^2}^2 \leq \|\mathcal{Y}(\tau)w\|_{H^2}\|\mathcal{Y}(\tau)w\|_{L^\infty}.$$

Ainsi

$$\|\Delta\mathcal{Y}(t)w\|_{L^2} \leq \|\Delta w\|_{L^2} + \int_0^t [M''\|\mathcal{Y}(\tau)w\|_{L^\infty} + M'] \, \|\mathcal{Y}(\tau)w\|_{H^2} d\tau.$$

Les estimations L^∞ et L^2 que nous avons obtenues prouvent qu'il existe une constante c qui dépend de $\|w\|_{L^\infty}$ telle que

$$\|\mathcal{Y}(t)w\|_{H^2} \leq \|w\|_{H^2} + c\int_0^t (1 + e^{K\tau})\|\mathcal{Y}(\tau)w\|_{H^2} d\tau.$$

Enfin, le lemme classique de Gronwall donne

$$\|\mathcal{Y}(t)w\|_{H^2} \leq \|w\|_{H^2} \exp\left(c\int_0^t (1 + e^{K\tau})d\tau\right).$$

Pour $t \leq 1$, il existe une constante C telle que

$$\exp\left(c\int_0^t \left(1 + e^{K\tau}\right) d\tau\right) \leq C.$$

On obtient ainsi l'estimation (Y2).

2. On obtient (Y3) en utilisant le même type de raisonnement que précédemment. On a alors évidement besoin des majorants de F jusqu'à sa quatrième dérivée.
3. On calcule la différence de deux formulations intégrales :

$$\mathcal{Y}(t)w_1 - \mathcal{Y}(t)w_2 = (w_1 - w_2) + \int_0^t [F(\mathcal{Y}(\tau)w_1) - F(\mathcal{Y}(\tau)w_2)]d\tau.$$

Comme F est K-lipschitzienne, on a

$$\|\mathcal{Y}(t)w_1 - \mathcal{Y}(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + K\int_0^t \|\mathcal{Y}(\tau)w_1 - \mathcal{Y}(\tau)w_2\|_{L^2} d\tau.$$

On applique le lemme de Gronwall classique avec $\phi(t) = \|\mathcal{Y}(t)w_1 - \mathcal{Y}(t)w_2\|_{L^2}$ et $p(t) = K$. On obtient

$$\|\mathcal{Y}(t)w_1 - \mathcal{Y}(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} \exp(Kt).$$

Pour $0 \leq t \leq 1$, il existe une constante C qui ne dépend que de K et donc que de la fonction F telle que $\exp(Kt) \leq (1 + Ct)$. On obtient l'estimation (Y4).

4. On utilise la formule intégrale pour trouver

$$\|\mathcal{Y}(t)w - w\|_{L^2} = \|\int_0^t F(\mathcal{Y}(\tau)w)d\tau\|_{L^2} \leq K\int_0^t \|\mathcal{Y}(\tau)w\|_{L^2} d\tau.$$

On utilise ensuite (Y1) et on borne l'exponentielle pour $0 \leq t \leq 1$, ce qui donne

$$\|\mathcal{Y}(t)w - w\|_{L^2} \leq K\exp(Kt)\int_0^t \|w\|_{L^2} d\tau \leq Ct\|w\|_{L^2}.$$

On obtient l'estimation (Y5).

■

8.2.4 Estimations du flot total $\mathcal{S}(t)$

Nous pouvons écrire la solution su système total sous forme intégrale en utilisant le propagateur $\mathcal{X}(t)$. Si u est solution de l'équation de Schrödinger non linéaire, on a

$$u(t) = \mathcal{X}(t)u^0 + \int_0^t \mathcal{X}(t-\tau)F(u(\tau))d\tau.$$

Nous allons faire le raisonnement en supposant que la donnée initiale est dans $H^2(\mathbb{R}^2)$. Un raisonnement identique permet de traiter le cas $u^0 \in H^4(\mathbb{R}^2)$. On se donne une temps T et on considère une fonction $w \in E \equiv L^\infty(0,T;H^2(\mathbb{R}^2))$. On peut alors définir l'application Z qui à $w \in E$ associe

$$(Zw)(t) = \mathcal{X}(t)u^0 + \int_0^t \mathcal{X}(t-\tau)F(w(\tau))d\tau.$$

On considère la boule B de centre 0 et de rayon $R = 2\|u^0\|_{H^2}$ dans E. Nous allons montrer que si T est suffisamment petit, Z définit une contraction dans cette boule B. Comme par ailleurs Z est clairement continue sur cet espace, Z admettra un unique point fixe, qui sera nécessairement la solution de l'équation de Schrödinger non linéaire.

Grâce à l'unitarité (X1) de $\mathcal{X}(t)$, nous pouvons écrire

$$\|(Zw)(t)\|_{H^2} \leq \|u^0\|_{H^2} + \int_0^t \|F(w(\tau))\|_{H^2} d\tau.$$

L'espace $H^2(\mathbb{R}^2)$ s'injecte dans les fonctions continues et donc *a fortiori* bornées sur $[0,T]$. Comme par ailleurs la fonction F est bornée ainsi que ses deux premières dérivées, la quantité $\|F(w(\tau))\|_{H^2}$ peut-être majorée par $C\|w\|_{L^\infty(0,T;H^2)}$. Cette constante C ne dépend que de F et pas de w. Comme $w \in B$, on a

$$\|(Zw)(t)\|_{H^2} \leq \frac{1}{2}R + CRt.$$

On choisit T tel que $CT \leq 1/2$ et on a alors

$$\|Zw\|_{L^\infty(0,T;H^2)} \leq R.$$

L'image par Z de w reste bien dans la boule B.

Pour montrer que l'on a une contraction, on prend w_1 et w_2 dans cette boule. On a

$$(Zw_1)(t) - (Zw_2)(t) = \int_0^t \mathcal{X}(t-\tau)[F(w_1(\tau)) - F(w_2(\tau))]d\tau.$$

On utilise les mêmes ingrédients que pour montrer que l'on reste dans la boule.

$$\begin{aligned} \|(Zw_1)(t) - (Zw_2)(t)\|_{H^2} &\leq \int_0^t \|F(w_1(\tau)) - F(w_2(\tau))\|_{H^2} d\tau \leq C\int_0^t \|w_1(\tau) - w_2(\tau)\|_{H^2} d\tau \\ &\leq Ct\|w_1 - w_2\|_{L^\infty(0,T;H^2)}. \end{aligned}$$

Pour un temps T suffisamment petit, on peut assurer que

$$\|(Zw_1)(t) - (Zw_2)(t)\|_{H^2} < \|w_1 - w_2\|_{L^\infty(0,T;H^2)}.$$

On a donc contraction.

Ceci permet d'assurer l'existence d'un temps T tel que la solution de l'équation de Schrödinger non linéaire est bien posé. Le fait que la constante C ne dépende pas de w permet d'appliquer à nouveau ce raisonnement à la donnée initiale $u(T)$, puis $u(2T)$, ... (bootstrap) et d'obtenir ainsi l'existence globale.

En conclusion, pour tout $u^0 \in H^2(\mathbb{R}^2)$ et tout temps T, $\mathcal{S}(t)u^0$ est uniformément borné dans $H^2(\mathbb{R}^2)$ pour $t \leq T$.

De même, on peut monter que pour tout $u^0 \in H^4(\mathbb{R}^2)$ et tout temps T, $\mathcal{S}(t)u^0$ est uniformément borné dans $H^4(\mathbb{R}^2)$ pour $t \leq T$.

8.3 Régularité lipschitzienne de $f(t)$

Pour montrer la convergence et calculer l'ordre des schémas, on peut se ramener comme dans le cas linéaire à étudier une erreur locale. En effet,

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq \sum_{j=0}^{n-1} \|f(h)^{n-j-1} f(h)\mathcal{S}(jh)u^0 - f(h)^{n-j-1}\mathcal{S}((j+1)h)u^0\|_{L^2}.$$

Si on sait montrer, que f a une régularité lipschitzienne, à savoir qu'il existe une constante C_0 telle que pour w_1 et $w_2 \in L^2$ et tout temps $t \in [0,1]$

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + C_0 t)\|w_1 - w_2\|_{L^2},$$

alors on aura

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq \sum_{j=0}^{n-1} (1 + C_0 h)^{n-j-1} \|(f(h) - \mathcal{S}(h))\mathcal{S}(jh)u^0\|_{L^2}.$$

On se ramènera ensuite à nouveau à l'étude de l'erreur locale $\|(f(h) - \mathcal{S}(h))v^0\|_{L^2}$ que l'on appliquera à $v^0 = S^{jh}u^0$. Commençons par montrer la régularité lipschizienne pour les quatre schémas.

8.3.1 Schéma de Lie (L1) : $f(t) = \mathcal{X}(t)\mathcal{Y}(t)$

On prend une donnée initiale $v^0 \in L^2(\mathbb{R}^2)$. On peut remplacer $\mathcal{Y}(t)v^0$ par sa formulation intégrale, ce qui permet par composition avec $\mathcal{X}(t)$ d'exprimer également la solution par le schéma de splitting sous forme intégrale :

$$f(t)v^0 = \mathcal{X}(t)v^0 + \int_0^t \mathcal{X}(t)F(\mathcal{Y}(\tau)v^0)\ d\tau.$$

Si on prend deux données initiales w_1 et $w_2 \in L^2(\mathbb{R}^2)$, comme $\mathcal{X}(t)$ est linéaire, la différence des deux solutions est donnée par

$$f(t)w_1 - f(t)w_2 = \mathcal{X}(t)(w_1 - w_2) + \int_0^t \mathcal{X}(t)(F(\mathcal{Y}(\tau)w_1) - F(\mathcal{Y}(\tau)w_2)d\tau.$$

Comme $\mathcal{X}(t)$ est unitaire (X1) dans L^2,

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + \int_0^t \|F(\mathcal{Y}(\tau)w_1) - F(\mathcal{Y}(\tau)w_2)\|_{L^2} d\tau.$$

On utilise ensuite le fait que F est lipschitzienne et l'estimation (Y4). Ainsi, il existe une constante C qui dépend uniquement de F telle que pour tout $0 \leq t \leq 1$

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + Ct)\|w_1 - w_2\|_{L^2}.$$

8.3.2 Schéma de Lie (L2) : $f(t) = \mathcal{Y}(t)\mathcal{X}(t)$

On exprime cette fois-ci la formulation intégrale pour $\mathcal{Y}(t)$ appliqué à $\mathcal{X}(t)v^0$ pour obtenir

$$f(t)v^0 = \mathcal{X}(t)v^0 + \int_0^t F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0)\ d\tau.$$

La différence entre deux telles solutions est

$$f(t)w_1 - f(t)w_2 = \mathcal{X}(t)(w_1 - w_2) + \int_0^t [F(\mathcal{Y}(\tau)\mathcal{X}(t)w_1) - F(\mathcal{Y}(\tau)\mathcal{X}(t)w_2)]d\tau.$$

On calcule la norme L^2 en utilisant une première fois que $\mathcal{X}(t)$ est unitaire (X1) sur L^2 :

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + \int_0^t \|F(\mathcal{Y}(\tau)\mathcal{X}(t)w_1) - F(\mathcal{Y}(\tau)\mathcal{X}(t)w_2)\|_{L^2} d\tau.$$

Comme F est lipschitzienne

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + K \int_0^t \|\mathcal{Y}(\tau)\mathcal{X}(t)w_1 - \mathcal{Y}(\tau)\mathcal{X}(t)w_2\|_{L^2} d\tau.$$

On utilise ensuite l'estimation (Y4) pour obtenir qu'il existe une constante C qui ne dépend que de F telle que pour tout $0 \leq t \leq 1$

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + K \int_0^t (1 + C\tau)\|\mathcal{X}(t)w_1 - \mathcal{X}(t)w_2\|_{L^2} d\tau.$$

Comme $\mathcal{X}(t)$ est linéaire et unitaire (X1) sur L^2, en changeant de constante C qui ne dépend toujours que de F, on obtient pour $0 \leq t \leq 1$

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + Ct)\|w_1 - w_2\|_{L^2}.$$

8.3.3 Schéma de Strang (S1) : $f(t) = \mathcal{X}(t/2)\mathcal{Y}(t)\mathcal{X}(t/2)$

Cette fois-ci, on applique la formulation intégrale de $\mathcal{Y}(t)$ en $\mathcal{X}(t/2)$ puis on recompose à nouveau par $\mathcal{X}(t/2)$ et cela donne

$$f(t)v^0 = \mathcal{X}(t)v^0 + \int_0^t \mathcal{X}(\frac{t}{2})F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0)\ d\tau.$$

La différence entre deux solutions est

$$f(t)w_1 - f(t)w_2 = \mathcal{X}(t)(w_1 - w_2) + \int_0^t \mathcal{X}(\frac{t}{2})(F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})w_1) - F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})w_2))d\tau,$$

On utilise l'unitarité (X1) de $\mathcal{X}(t)$ et $\mathcal{X}(t/2)$. Ainsi

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq \|w_1 - w_2\|_{L^2} + \int_0^t \|F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})w_1) - F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})w_2)\|_{L^2} d\tau.$$

L'intégrale à estimer est la même que celle du cas précédent en remplaçant $\mathcal{X}(t)$ par $\mathcal{X}(t/2)$ ce qui ne change rien aux estimations. On a donc

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + Ct)\|w_1 - w_2\|_{L^2}.$$

8.3.4 Schéma de Strang (S2) : $f(t) = \mathcal{Y}(t/2)\mathcal{X}(t)\mathcal{Y}(t/2)$

On applique la formulation intégrale de $\mathcal{Y}(t/2)$ à $\mathcal{X}(t)\mathcal{Y}(t/2)v^0$. Attention : comme ceci se passe sur un demi-pas de temps, l'intégrale est naturellement à prendre sur l'intervalle $[0, t/2]$:

$$f(t)v^0 = \mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 + \int_0^{\frac{t}{2}} F(\mathcal{Y}(\tau)\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0)\ d\tau.$$

La différence entre deux solutions est

$$\begin{aligned} f(t)w_1 - f(t)w_2 &= \mathcal{X}(t)\mathcal{Y}(\frac{t}{2})w_1 - \mathcal{X}(t)\mathcal{Y}(\frac{t}{2})w_2 \\ &\quad + \int_0^{\frac{t}{2}} [F(\mathcal{Y}(\tau)\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})w_1) - F(\mathcal{Y}(\tau)\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})w_2)]d\tau, \end{aligned}$$

On majore d'une part $\mathcal{X}(t)\mathcal{Y}(t/2)w_1 - \mathcal{X}(t)\mathcal{Y}(t/2)w_2$ en norme L^2 en utilisant successivement que $\mathcal{X}(t)$ est unitaire dans L^2 puis (Y4). Pour le terme intégral, on utilise que F est lipschitzienne, puis (Y4), puis que $\mathcal{X}(t)$ est unitaire et à nouveau (Y4). On finalement

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + Ct)\|w_1 - w_2\|_{L^2}.$$

Comme on cherche le même type d'estimation pour les quatre schémas, les démonstration sont de même difficulté dans les quatre cas. Ce sera sensiblement différent pour les calculs d'erreurs locales qui suivent !

8.4 Erreurs locales

Pour déterminer l'ordre du schéma, on écrit la formulation intégrale que nous avons déjà utilisée pour étudier le problème de Cauchy :

$$\mathcal{S}(t)v^0 = \mathcal{X}(t)v^0 + \int_0^t \mathcal{X}(t-\tau)F(\mathcal{S}(\tau)v^0)d\tau.$$

On écrit la différence avec la solution splittée sous la forme

$$\mathcal{S}(t)v^0 - f(t)v^0 = \int_0^t \mathcal{X}(t-\tau)\left[F(S(\tau)v^0) - F(f(\tau)v^0)\right]d\tau + R(t),$$

où $R(t)$ est un reste. Comme $\mathcal{X}(t-\tau)$ est unitaire (X1) dans L^2 et F est K-lipschitzienne, on a une estimation de l'erreur locale

$$\|\mathcal{S}(t)v^0 - f(t)v^0\|_{L^2} \leq K\int_0^t \|\mathcal{S}(\tau)v^0 - f(\tau)v^0\|_{L^2}d\tau + \|R(t)\|_{L^2}.$$

Il reste uniquement à montrer que $\|R(t)\|_{L^2} = O(t^{p+1})$ pour t petit et utiliser le lemme de Gronwall modifié pour conclure que le schéma est d'ordre p.

8.4.1 Schéma de Lie (L1) : $f(t) = \mathcal{X}(t)\mathcal{Y}(t)$

Pour $v^0 \in H^2(\mathbb{R}^2)$ et $0 \leq t \leq 1$, le reste $R(t)$ s'écrit

$$R(t) = \int_0^t \mathcal{X}(t-\tau)F(\mathcal{X}(\tau)\mathcal{Y}(\tau)v^0)d\tau - \int_0^t \mathcal{X}(t)F(\mathcal{Y}(\tau)v^0)d\tau.$$

On peut écrire ce reste sous la forme $R(t) = \int_0^t \mathcal{X}(t-\tau)R_1(\tau)d\tau$ où

$$R_1(\tau) = F(\mathcal{X}(\tau)\mathcal{Y}(\tau)v^0) - \mathcal{X}(\tau)F(\mathcal{Y}(\tau)v^0).$$

On additionne et on soustrait la quantité $F(\mathcal{Y}(\tau)v^0)$ pour obtenir

$$R_1(\tau) = F(\mathcal{X}(\tau)\mathcal{Y}(\tau)v^0) - F(\mathcal{Y}(\tau)v^0) + F(\mathcal{Y}(\tau)v^0) - \mathcal{X}(\tau)F(\mathcal{Y}(\tau)v^0).$$

Comme F est lipschitzienne et grâce à (X2) et (Y2), on a d'une part

$$\|F(\mathcal{X}(\tau)\mathcal{Y}(\tau)v^0) - F(\mathcal{Y}(\tau)v^0)\|_{L^2} \leq CK\tau\|v^0\|_{H^2},$$

et d'autre part, en utilisant (X2),

$$\|F(\mathcal{Y}(\tau)v^0) - \mathcal{X}(\tau)F(\mathcal{Y}(\tau)v^0)\|_{L^2} \leq \tau\|F(\mathcal{Y}(\tau)v^0\|_{H^2}.$$

Comme F lipschitzienne, ce qui majore $F(w)$ nul en 0, et ses deux premières dérivées sont bornées, on a

$$\|F(\mathcal{Y}(\tau)v^0) - \mathcal{X}(\tau)F(\mathcal{Y}(\tau)v^0)\|_{L^2} \leq C\tau\|\mathcal{Y}(\tau)v^0\|_{H^2}.$$

On utilise enfin (Y2), pour obtenir

$$\|F(\mathcal{Y}(\tau)v^0) - \mathcal{X}(\tau)F(\mathcal{Y}(\tau)v^0)\|_{L^2} \leq C\tau\|v^0\|_{H^2}.$$

Finalement, il existe une constante qui ne dépend que de F telle que pour $0 \leq t \leq 1$

$$\|R_1(\tau)\|_{L^2} \leq C\tau\|v^0\|_{H^2}.$$

En revenant à l'intégrale, et comme $\mathcal{X}(t-\tau)$ est unitaire (X1) sur L^2,

$$\|R(t)\|_{L^2} \leq C\|v^0\|_{H^2}\int_0^t \tau d\tau = \frac{Ct^2}{2}\|v^0\|_{H^2}.$$

8.4.2 Schéma de Lie (L2) : $f(t) = \mathcal{Y}(t)\mathcal{X}(t)$

Pour $v^0 \in H^2(\mathbb{R}^2)$ et $0 \leq t \leq 1$, le reste $R(t)$ s'écrit

$$R(t) = \int_0^t \mathcal{X}(t-\tau)F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0)d\tau - \int_0^t F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0)d\tau.$$

On peut cette fois-ci écrire $R(t) = \int_0^t R_1(\tau)d\tau$ avec

$$R_1 = \mathcal{X}(t-\tau)F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0),$$

et de même que précédemment on ajoute et on retranche $F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0)$ pour obtenir

$$\begin{aligned} R_1(\tau) &= \mathcal{X}(t-\tau)F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) \\ &\quad + F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0). \end{aligned}$$

De même que pour le deuxième bout à estimer du splitting précédent, on a

$$\|\mathcal{X}(t-\tau)F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0)\|_{L^2} \leq C(t-\tau)\|\mathcal{X}(\tau)v^0\|_{H^2}$$

et comme $\mathcal{X}(\tau)$ est unitaire (X1) sur H^2, on peut en fait majorer par $C(t-\tau)\|v^0\|_{H^2}$. Pour la deuxième quantité, on utilise le fait que F est lipschitzienne, puis (Y4) pour obtenir

$$\|F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0)\|_{L^2} \leq K(1+Ct)\|\mathcal{X}(t)v^0 - \mathcal{X}(\tau)v^0\|_{L^2}.$$

Comme $\mathcal{X}(t-\tau)$ est unitaire (X1) dans L^2 et par linéarité

$$\|\mathcal{X}(t)v^0 - \mathcal{X}(\tau)v^0\|_{L^2} = \|\mathcal{X}(t-\tau)v^0 - v^0\|_{L^2}.$$

On utilise enfin (X3) et

$$\begin{aligned} \|F(\mathcal{Y}(\tau)\mathcal{X}(\tau)v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(t)v^0)\|_{L^2} &\leq K(1+Ct)(t-\tau)\|v^0\|_{H^2} \\ &\leq C(t-\tau)\|v^0\|_{H^2} \end{aligned}$$

pour $0 \leq t \leq 1$. Finalement

$$\begin{aligned} \|R_1(\tau)\|_{L^2} &\leq C(t-\tau)\|v^0\|_{H^2}, \\ \|R(t)\|_{L^2} &\leq C\|v^0\|_{H^2}\int_0^t (t-\tau)d\tau = \frac{Ct^2}{2}\|v^0\|_{H^2}. \end{aligned}$$

8.4.3 Schéma de Strang (S1) : $f(t) = \mathcal{X}(t/2)\mathcal{Y}(t)\mathcal{X}(t/2)$

Pour $v^0 \in H^4(\mathbb{R}^2)$ et $0 \leq t \leq 1$, le reste $R(t)$ s'écrit

$$R(t) = \int_0^t \mathcal{X}(t-\tau)F(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0)d\tau - \int_0^t \mathcal{X}(\frac{t}{2})F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0)d\tau.$$

On définit cette fois-ci deux intégrandes que l'on estime séparément :

$$R(t) = \int_0^t R_1(\tau)d\tau + \mathcal{X}(\frac{t}{2})\int_0^t R_2(\tau)d\tau,$$

avec

$$\begin{aligned} R_1(\tau) &= \mathcal{X}(t-\tau)w(\tau) - \mathcal{X}(\frac{t}{2})w(\tau), \qquad \text{et } w(\tau) = F(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0), \\ R_2(\tau) &= F(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0) - F(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0). \end{aligned}$$

Pour le premier reste R_1, on utilise tout d'abord (X4) pour obtenir

$$\|\int_0^t R_1(\tau)d\tau\|_{L^2} \leq Ct^3(\|w\|_{\mathcal{C}^1([0,T];H^2)} + \|w\|_{L^\infty([0,T],H^4)}).$$

Les estimations sur les flots $\mathcal{X}$ et $\mathcal{Y}$ assurent alors que

$$\begin{aligned}\|w\|_{\mathcal{C}^1([0,T];H^2)} &\leq C\|v^0\|_{H^4},\\ \|w\|_{L^\infty([0,T],H^4)} &\leq C\|v^0\|_{H^4}.\end{aligned}$$

Ainsi

$$\|\int_0^t R_1(\tau)d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^4}.$$

Pour le second reste R_2, un développement de Taylor avec reste intégral donne

$$\begin{aligned}R_2(\tau) &= F'(v^0)\cdot(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - \mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0)\\ &\quad + \int_0^1 (1-\theta)\Big[F''(v^0+\theta(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - v^0))\cdot\ (\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - v^0)^2\\ &\qquad -F''(v^0+\theta(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0 - v^0))\cdot(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0 - v^0)^2\Big]d\theta.\end{aligned}$$

On utilise une inégalité triangulaire, puis (X2) et (Y5) et enfin (Y2) et (X1), pour obtenir

$$\begin{aligned}&\|\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - v^0\|_{L^2}\\ &\quad\leq \|\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - \mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0\|_{L^2} + \|\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - \mathcal{X}(\frac{\tau}{2})v^0\|_{L^2} + \|\mathcal{X}(\frac{\tau}{2})v^0 - v^0\|_{L^2}\\ &\quad\leq \frac{\tau}{2}\|\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0\|_{H^2} + C\tau\|\mathcal{X}(\frac{\tau}{2})v^0\|_{L^2} + \frac{\tau}{2}\|v^0\|_{H^2}\\ &\quad\leq C\tau\|v^0\|_{H^2}.\end{aligned}$$

On voit aisément que de même

$$\|\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0 - v^0\|_{L^2} \leq Ct\|v^0\|_{H^2}.$$

On utilise ensuite que $H^2(\mathbb{R}^2)$ est une algèbre

$$\begin{aligned}&\Big\|\int_0^1 (1-\theta)\Big[F''(v^0+\theta(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - v^0))\cdot(\mathcal{X}(\frac{\tau}{2})\mathcal{Y}(\tau)\mathcal{X}(\frac{\tau}{2})v^0 - v^0)^2\\ &\qquad - F''(v^0+\theta(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0 - v^0))\cdot(\mathcal{Y}(\tau)\mathcal{X}(\frac{t}{2})v^0 - v^0)^2\Big]d\theta\Big\|_{L^2}\\ &\qquad\leq Ct^2\|v^0\|_{H^2}^2.\end{aligned}$$

Posons $R_3(\tau) = \mathcal{X}(\tau/2)\mathcal{Y}(\tau)\mathcal{X}(\tau/2)v^0 - \mathcal{Y}(\tau)\mathcal{X}(t/2)v^0$. La formulation intégrale de $\mathcal{Y}(t)$ donne

$$\begin{aligned}R_3(\tau) &= \mathcal{X}(\tau)v^0 - \mathcal{X}(\frac{t}{2})v^0 + \int_0^\tau [\mathcal{X}(\frac{\tau}{2})F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0) - F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0)]d\sigma\\ &\quad + \int_0^\tau [F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0) - F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{t}{2})v^0)]d\sigma.\end{aligned}$$

On utilise l'estimation (X4) appliquée à une fonction $w = v^0$ constante en temps pour obtenir

$$\|\int_0^t (\mathcal{X}(\tau)v^0 - \mathcal{X}(\frac{t}{2})v^0)d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^4}.$$

Pour la deuxième partie, on utilise l'estimation (X2), puis le fait que F est lipschitzienne, puis les estimations (Y2) et (X1) et enfin deux intégrations en temps, ce qui donne

$$\|\int_0^t\int_0^\tau (\mathcal{X}(\frac{\tau}{2})F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0) - F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0))d\sigma d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^2}.$$

De même, on utilise le fait que F est lipschitzienne, les estimations (Y4), (X2) et (X1) et enfin deux intégrations en temps pour obtenir, pour tout $0 \leq t \leq 1$,

$$\|\int_0^t\int_0^\tau (F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{\tau}{2})v^0) - F(\mathcal{Y}(\sigma)\mathcal{X}(\frac{t}{2})v^0))d\sigma d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^2}.$$

Finalement, comme $\mathcal{X}(t/2)$ est unitaire (X1)

$$\|\mathcal{X}(\frac{t}{2})\int_0^t R_2(\tau)d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^4},$$

et en conclusion

$$\|R(t)\|_{L^2} \leq C(1+\|v^0\|_{H^4})t^3\|v^0\|_{H^4}.$$

8.4.4 Schéma de Strang (S2) : $f(t) = \mathcal{Y}(t/2)\mathcal{X}(t)\mathcal{Y}(t/2)$

Pour $v^0 \in H^4(\mathbb{R}^2)$ et $0 \leq t \leq 1$, le reste $R(t)$ s'écrit

$$\begin{aligned} R(t) &= \int_0^t \mathcal{X}(t-\tau)F(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0)d\tau - \frac{1}{2}\int_0^t \mathcal{X}(t)F(\mathcal{Y}(\frac{\tau}{2})v^0)d\tau \\ &\quad -\frac{1}{2}\int_0^t F(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0)d\tau. \end{aligned}$$

Grâce à des développements de Taylor

$$\begin{aligned} F(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0) &= F(v^0) + F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) \\ &\quad + \int_0^1 (1-\theta)F''(v^0+\theta(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0))\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0)^2 d\theta, \\ F(\mathcal{Y}(\frac{\tau}{2})v^0) &= F(v^0) + F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) \\ &\quad + \int_0^1 (1-\theta)F''(v^0+\theta(\mathcal{Y}(\frac{\tau}{2})v^0 - v^0))\cdot(\mathcal{Y}(\frac{\tau}{2})v^0 - v^0)^2 d\theta, \\ F(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0) &= F(v^0) + F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 - v^0) \\ &\quad + \int_0^1 (1-\theta)F''(v^0+\theta(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 - v^0))\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 - v^0)^2 d\theta, \end{aligned}$$

On utilise à nouveau les estimations différentes estimations et

$$\begin{aligned} \|\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0\|_{L^2} &\leq C\tau\|v^0\|_{H^2}, \\ \|\mathcal{Y}(\frac{\tau}{2})v^0 - v^0\|_{L^2} &\leq C\tau\|v^0\|_{H^2}, \\ \|\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 - v^0\|_{L^2} &\leq Ct\|v^0\|_{H^2}. \end{aligned}$$

On peut majorer l'intégrale des restes intégraux sur l'intervalle $[0,t]$ par $Ct^3\|v^0\|_{H^2}^2$. Il reste ensuite à estimer $\int_0^t R_1(\tau)d\tau$ où

$$\begin{aligned} R_1(\tau) &= \left(\mathcal{X}(t-\tau) - \frac{1}{2}\mathcal{X}(t) - \frac{1}{2}\mathrm{I}\right)F(v^0) + (\mathcal{X}(t-\tau)-\mathrm{I})F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) \\ &\quad -\frac{1}{2}(\mathcal{X}(t)-\mathrm{I})F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) + F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) \\ &\quad -\frac{1}{2}F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})v^0 - v^0) - \frac{1}{2}F'(v^0)\cdot(\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 - v^0). \end{aligned}$$

On majore le premier terme par $Ct^3\|v^0\|_{H^4}$, en combinant les estimations (X4) et (X5). Les deux termes suivants sont majorés respectivement par $CM'(t-\tau)\tau\|v^0\|_{H^4}$ et $CM't\tau\|v^0\|_{H^2}$. Enfin, comme $F'(v^0)$ est un opérateur linéaire, il nous faut étudier la quantité $\int_0^t F'(v^0)R_2(\tau)d\tau$ avec

$$\begin{aligned} R_2(\tau) &= \mathcal{Y}(\frac{\tau}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0 - \frac{1}{2}\mathcal{Y}(\frac{\tau}{2})v^0 - \frac{1}{2}\mathcal{Y}(\frac{\tau}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0 \\ &= \mathcal{X}(\tau)v^0 + \frac{1}{2}\int_0^\tau \mathcal{X}(\tau)F(\mathcal{Y}(\frac{\sigma}{2})v^0)d\sigma + \frac{1}{2}\int_0^\tau F(\mathcal{Y}(\frac{\sigma}{2})\mathcal{X}(\tau)\mathcal{Y}(\frac{\tau}{2})v^0)d\sigma \\ &\quad -\frac{1}{2}v^0 - \frac{1}{4}\int_0^\tau F(\mathcal{Y}(\frac{\sigma}{2})v^0)d\sigma \\ &\quad -\frac{1}{2}\mathcal{X}(t)v^0 - \frac{1}{4}\int_0^t \mathcal{X}(t)F(\mathcal{Y}(\frac{\sigma}{2})v^0)d\sigma - \frac{1}{4}\int_0^t F(\mathcal{Y}(\frac{\sigma}{2})\mathcal{X}(t)\mathcal{Y}(\frac{t}{2})v^0)d\sigma, \end{aligned}$$

où on a utilisé la formulation intégrale sur le flot $\mathcal{Y}$ de manière intensive. Partout où F intervient, on additionne et on soustrait la quantité $F(v^0)$. On peut estimer les termes avec cette différence par $Ct^2\|v^0\|_{H^2}$, et leur intégrale en temps par $Ct^3\|v^0\|_{H^2}$. Il ne reste que les termes

$$\begin{aligned} R_3(\tau) &= \mathcal{X}(\tau)v^0 - \frac{1}{2}v^0 - \frac{1}{2}\mathcal{X}(t)v^0, \\ R_4(\tau) &= \frac{1}{2}\int_0^\tau \mathcal{X}(\tau)F(v^0)d\sigma - \frac{1}{4}\int_0^t \mathcal{X}(t)F(v^0)d\sigma. \end{aligned}$$

On a d'après (X4) et (X5)

$$\begin{aligned} R_3(\tau) &= (\mathcal{X}(\tau)v^0 - \mathcal{X}(\frac{t}{2})v^0) + (\mathcal{X}(\frac{t}{2})v^0 - \frac{1}{2}v^0 - \frac{1}{2}\mathcal{X}(t)v^0), \\ \|\int_0^t R_3(\tau)d\tau\|_{L^2} &\leq Ct^3\|v^0\|_{H^4}; \\ R_4(\tau) &= \frac{1}{2}\left(\tau(\mathcal{X}(\tau)-\mathcal{X}(t))F(v^0) + (\tau - \frac{t}{2})\mathcal{X}(t)F(v^0)\right), \\ \|\int_0^t R_4(\tau)d\tau\|_{L^2} &= \|\int_0^t \tau(\mathcal{X}(\tau)-\mathcal{X}(t))F(v^0)d\tau\|_{L^2} \leq Ct^3\|v^0\|_{H^2}, \end{aligned}$$

le dernier terme de $R_4(\tau)$ étant d'intégrale nulle. Finalement, on obtient que

$$\|R(t)\|_{L^2} \leq C(1+\|v^0\|_{H^4})t^3\|v^0\|_{H^4}.$$

Maintenant que l'on a calculé le reste pour chacun des quatre schémas, le lemme de Gronwall permet de conclure immédiatement quant à l'erreur locale des schémas.

Lemme 7

Soit $v^0 \in H^2(\mathbb{R}^2)$, alors il existe $t_0 > 0$ tel que pour tout $0 \leq t \leq t_0$ et une constante C qui ne dépend de $\|v^0\|_{H^2}$, tels que pour un schéma de Lie

$$\|f(t)v^0 - S(t)v^0\|_{L^2} \leq Ct^2\|v^0\|_{H^2}.$$

Si de plus $v^0 \in H^4(\mathbb{R}^2)$, alors il existe $t_1 > 0$ tel que pour tout $0 \leq t \leq t_1$ et une constante C qui ne dépend de $\|v^0\|_{H^4}$, tels que pour un schéma de Strang

$$\|f(t)v^0 - S(t)v^0\|_{L^2} \leq Ct^3\|v^0\|_{H^4}.$$

8.5 Estimation d'ordre

Théorème 7

Pour tout u^0 dans $H^2(\mathbb{R}^2)$ et tout $T > 0$, il existe C et h_0 tels que pour tout $h \in]0, h_0]$, et pour tout n tel que $nh \leq T$ pour un splitting de Lie

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq C(\|u^0\|_{H^2})h\|u^0\|_{H^2}.$$

Si de plus $u^0 \in H^4(\mathbb{R}^2)$, alors pour un splitting de Strang

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq C(\|u^0\|_{H^4})h^2\|u^0\|_{H^4}.$$

Preuve :
Nous avons déjà remarqué que

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq \sum_{j=0}^{n-1} \|f(h)^{n-j-1}f(h)\mathcal{S}(jh)u^0 - f(h)^{n-j-1}\mathcal{S}((j+1)h)u^0\|_{L^2}.$$

Pour tous les schémas étudiés, nous avons montré l'existence d'une constante C_0 telle que pour tout w_1 et $w_2 \in L^2(\mathbb{R}^2)$ et tout temps $t \in [0,1]$

$$\|f(t)w_1 - f(t)w_2\|_{L^2} \leq (1 + C_0 t)\|w_1 - w_2\|_{L^2},$$

et ainsi

$$\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} \leq \sum_{j=0}^{n-1} (1 + C_0 h)^{n-j-1} \|(f(h) - \mathcal{S}(h))\mathcal{S}(jh)u^0\|_{L^2}.$$

Pour une méthode de Lie et $u^0 \in H^2(\mathbb{R}^2)$, pour tout j tel que $jh \leq T$, $\mathcal{S}(jh)u^0 \in H^2(\mathbb{R}^2)$ et est uniformément borné dans cet espace. Ainsi

$$\|(f(h) - \mathcal{S}(h))\mathcal{S}(jh)u^0\|_{L^2} \leq C(\|u^0\|_{H^2})h^2\|u^0\|_{H^2},$$

et on déduit que

$$\begin{aligned}
\|f(h)^n u^0 - \mathcal{S}(nh)u^0\|_{L^2} &\leq C(\|u^0\|_{H^2})\|u^0\|_{H^2} \sum_{j=0}^{n-1} \exp(C_0 h)^{n-j-1} h^2 \\
&\leq C(\|u^0\|_{H^2})\|u^0\|_{H^2} \exp(C_0 T) n h^2 \leq C(\|u^0\|_{H^2})\|u^0\|_{H^2} h.
\end{aligned}$$

Pour une méthode de Strang et $u^0 \in H^4(\mathbb{R}^2)$, pour tout j tel que $jh \leq T$, $\mathcal{S}(jh)u^0 \in H^4(\mathbb{R}^2)$ et est uniformément borné dans cet espace. On a

$$\begin{aligned}
\|f(h)^n u^0 - \mathcal{S}(nh)u^0\| &\leq C(\|u^0\|_{H^4})\|u^0\|_{H^4} \sum_{j=0}^{n-1} \exp(C_0 h)^{n-j-1} h^3 \\
&\leq C(\|u^0\|_{H^4})\|u^0\|_{H^4} \exp(C_0 T) n h^3 \leq C(\|u^0\|_{H^4})\|u^0\|_{H^4} h^2.
\end{aligned}$$

Ceci achève la preuve de notre théorème ■

8.6 Illustration numérique

Nous allons essayer de corroborer numériquement les ordres théoriques obtenus.

En pratique, la régularité des données numériques utilisées n'est pas exactement celle voulue. Par exemple, la donnée L^2 est bien plus régulière une fois discrétisée.

On calcule la solution aux temps $t^n = nh$ sur le domaine spatial $\Omega = [-10, 10] \times [-10, 10]$. La formule pour calculer l'ordre numérique est

$$p_{\text{num}} = \max_{t_n \in [0,T]} \ln \left(\frac{\|u_2 - u_1\|_{L^2(\Omega)}}{\|u_3 - u_2\|_{L^2(\Omega)}} \right) / \ln 2,$$

où u_1 est calculé avec un pas de temps h, et u_2 et u_3 avec des pas de temps $h/2$ et $h/4$ respectivement.

Les données initiales choisies sont tracées sur la figure 6. Pour éviter les réflexions sur les bords, on choisit des conditions aux bords périodiques et on utilise une transformée de Fourier rapide (FFT) pour inverser le laplacien.

L'ordre numérique obtenu pour chaque donnée initiale et chaque méthode est donné par le tableau 1. On utilise le pas de temps $h = 10^{-3}$ et $N = 128$ points de discrétisation spatiale dans chacune des deux directions.

Les résultats ne dépendent pas énormément des pas de temps et d'espace. Nous donnons cette dépendance pour la donnée initiale H^2 et la formulation de Strang dans le tableau 2.

8.7 Conclusion

Nous n'avons pas vraiment bien traité la raideur dans le cas de l'application à Schrödinger cubique. En effet, dans le cas focalisant où il y a explosion en temps fini, les propriétés ne sont démontrées que pour un temps strictement inférieur au temps d'explosion. Les constantes obtenues dépendent de normes H^2 et H^4 qui explosent (puisque c'est la norme H^1 qui explose. Il serait intéressant d'obtenir des estimations valides jusqu'au temps d'explosion, ce qui permettrait d'évaluer l'efficacité de telles méthodes pour capturer le profil d'explosion justement. C'est une source de raideur, et il faut s'attendre, comme dans le cas de termes raides pour les systèmes linéaires, à avoir des pertes d'ordre et des estimations différentes suivant l'ordre choisi pour les méthodes de splitting.

TAB. 1 – Calcul de l'ordre numérique p_{num} pour les différentes données initiales

	Lie	Strang
H^2	1.000685	2.000072
H^1	1.001721	2.006374
L^2	1.014480	2.010045

TAB. 2 – Calcul de p_{num} pour différents pas de temps et d'espace.

	$N = 64$	$N = 128$	$N = 256$
$h = 10^{-3}$	2.000016	2.000072	2.000289
$h = 10^{-2}$	2.001637	2.007160	2.030023

Références

[1] C. Besse, B. Bidégaray et S. Descombes, *Order estimates in time of splitting methods for the nonlinear Schrödinger equation,* SIAM Journal on Numerical Analysis, **40**, 26–40, 2002.

[2] H. Brezis, *Analyse fonctionnelle. Théorie et applications,* Masson, 1987.

[3] T. Cazenave et A. Haraux, *Introduction aux problèmes d'évolutionsemi-linéaires,* Mathématiques et Applications 1, Springer, 1990.

[4] J. Douglas Jr., *On the numerical integration of $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = \partial u/\partial t$ by implicit methods,* Journal of the Society of Industrial and Applied Mathematics, **3**, 42–65, 1955.

[5] G.I. Marchuk, *Metody vychislitel'noi' matematiki,* Nauka, 1989.

[6] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Applied Mathematical Sciences 44, Springer, 1983.

[7] D.W. Peaceman et H.H. Rachford Jr., *The numerical solution of parabolic and elliptic differential equations,* Journal of the Society of Industrial and Applied Mathematics, **3**, 28–41, 1955.

[8] M. Schatzman, *Higher order alternate direction methods,* Computational Methods in Applied Mechanical Engineering, **116**, 219–225, 1994.

[9] M. Schatzman, *Stability of the Peaceman–Rachford approximation,* Journal of Functional Analysis, **162**, 219–255, 1999.

[10] B. Sportisse, *An analysis of operator splitting techniques in the stiff case,* Journal of Computational Physics, **161**, 140–168, 2000.

[11] G. Strang, *On the construction and comparison of difference schemes,* SIAM Journal on Numerical Analysis, **5**(3), 506–517, 1968.

[12] J.G. Verwer et B. Sportisse, *A note on operator splitting in a stiff linear case,* Rapport technique MAS–R9830, CWI, Amsterdam, 1998.

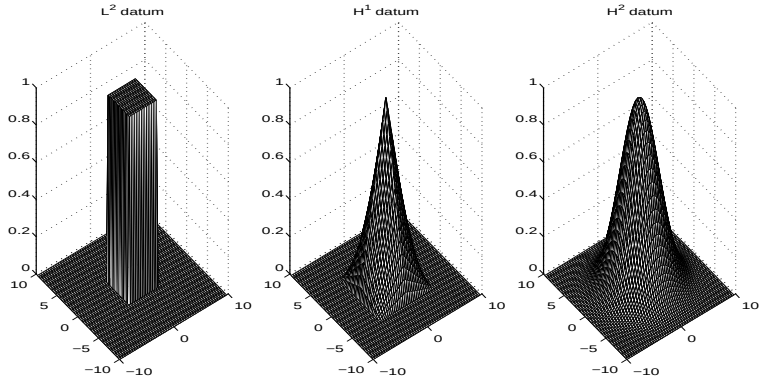

FIG. 6 – Données initiales des cas tests

A Le code des exemples

```
clear;

% Exemple 1
% A = [-1.E4 1.E4 1.;1.E4 -1.E4 2;1. 1. -2.];
% B = [-1. .5 .25;.1 0. .1; .2 .4 -1.];
% x = [1.;1.;1.];
% listedt = [1.E-4 1.E-3 1.E-2 1.E-1 1.];
% tmax = 1.;

% Exemple 2
A = [-1.E6 -1.E6 1.;1.E6 -1.E6 2;1. 1. -2.];
B = [-1. .5 .25;.1 0. .1; .2 .4 -1.];
x = [1.;1.;1.];
listedt = [1.E-4 1.E-3 1.E-2 1.E-1 1.];
tmax = 1.;

vardiag = true;

nAB = [];
nBA = [];
nABA = [];
nBAB = [];
nN = [];
EAB = [];
EBA = [];
EABA = [];
EBAB = [];
EN = [];

if vardiag
  [VA,DA]=eig(A);
  x = VA*x;
  B = VA*B*inv(VA);
  A = DA;
end

for dt=listedt
  t = dt:dt:tmax;
  expA = expm(dt*A);
```

```
expA2 = expm(dt/2*A);

expB = expm(dt*B);
expB2 = expm(dt/2*B);

expC = expm(dt*(A+B));

S = expC;
Sx = x;
for tt=t
  Sx = S*Sx;
end

AB = expA*expB;
ABx = x;
for tt=t
  ABx = AB*ABx;
end

BA = expB*expA;
BAx = x;
for tt=t
  BAx = BA*BAx;
end

ABA = expA2*expB*expA2;
ABAx = x;
for tt=t
  ABAx = ABA*ABAx;
end

BAB = expB2*expA*expB2;
BABx = x;
for tt=t
  BABx = BAB*BABx;
end

Id = [1. 0. 0.;0. 1. 0.;0. 0. 1.];
N = expA+inv(A)*(expA-Id)*(expB-Id)/dt;
Nx = x;
for tt=t
  Nx = N*Nx;
end

% normes

nSx =  sqrt(sum(abs(Sx).^2,1));
diffAB =  Sx-ABx;
diffBA =  Sx-BAx;
diffABA = Sx-ABAx;
diffBAB = Sx-BABx;
diffN =   Sx-Nx;

nAB =  [nAB  sqrt(sum(abs(diffAB).^2,1))/nSx];
nBA =  [nBA  sqrt(sum(abs(diffBA).^2,1))/nSx];
```

```
    nABA = [nABA sqrt(sum(abs(diffABA).^2,1))/nSx];
    nBAB = [nBAB sqrt(sum(abs(diffBAB).^2,1))/nSx];
    nN =   [nN   sqrt(sum(abs(diffN).^2,1))/nSx];

    % erreurs

    EAB =  [EAB  abs(diffAB./Sx)];
    EBA =  [EBA  abs(diffBA./Sx)];
    EABA = [EABA abs(diffABA./Sx)];
    EBAB = [EBAB abs(diffBAB./Sx)];
    EN =   [EN   abs(diffN./Sx)];
end

figure(1)

loglog(listedt,nAB,'r-o', ...
       listedt,nBA,'b-+', ...
       listedt,nABA,'m-s', ...
       listedt,nBAB,'c-x', ...
       listedt,nN,'k-p', 'LineWidth',2),
  legend('AB','BA','ABA','BAB','ST', ..
         'Location','NorthWest');

figure(2)

subplot(1,3,1),
  loglog(listedt,EAB(1,:),'r-o', ...
         listedt,EBA(1,:),'b-+', ...
         listedt,EABA(1,:),'m-s', ...
         listedt,EBAB(1,:),'c-x', ...
         listedt,EN(1,:),'k-p', 'LineWidth',2),
  legend('AB','BA','ABA','BAB','ST', ...
         'Location','NorthWest');

subplot(1,3,2),
  loglog(listedt,EAB(2,:),'r-o', ...
         listedt,EBA(2,:),'b-+', ...
         listedt,EABA(2,:),'m-s', ...
         listedt,EBAB(2,:),'c-x', ...
         listedt,EN(2,:),'k-p', 'LineWidth',2),
  legend('AB','BA','ABA','BAB','ST', ...
         'Location','NorthWest');

subplot(1,3,3),
  loglog(listedt,EAB(3,:),'r-o', ...
         listedt,EBA(3,:),'b-+', ...
         listedt,EABA(3,:),'m-s', ...
         listedt,EBAB(3,:),'c-x', ...
         listedt,EN(3,:),'k-p', 'LineWidth',2),
  legend('AB','BA','ABA','BAB','ST', ...
         'Location','NorthWest');
```

B Hypothèses pour la stabilité en $\sqrt{t}$

On donne ici les hypothèses développées par Michelle Schatzman pour obtenir $\|\mathcal{P}(t)\|_{\mathcal{B}(H)} \leq 1 + C\sqrt{t}$.

On suppose que A et B sont deux opérateurs auto-adjoints sur H et que D est un sous-espace dense de H tel que

(H1) $D \subset D(A) \cap D(B), \qquad AD \subset D, \qquad BD \subset D.$

Soit $\mathcal{Q}$ l'ensemble des polynômes Q tels que $Q(X) = \mathrm{I} + \sum_{j=1}^{k} Q_j X^j$ est positif pour tout X positif. Une analyse spectrale assure que $Q(A)^{-1}$ et $Q(B)^{-1}$ sont des opérateurs bornés. On suppose en outre que pour tout $Q \in \mathcal{Q}$

(H2) $Q(A)^{-1}D \subset D, \qquad Q(B)^{-1}D \subset D.$

Toute méthode produit est alors un élément de $\mathcal{L}(D)$. Si de plus il est borné dans $\mathcal{B}(H)$ alors il sera borné dans H.

On suppose qu'il existe une algèbre $\mathcal{M}$ d'opérateurs bornés sur H, telle que pour tout $m \in \mathcal{M}$

(H3) $mD \subset D : \qquad \mathcal{M} \in \mathcal{L}(D) \cap \mathcal{B}(H).$

On note $a = \sqrt{A}$ et $b = \sqrt{B}$. On suppose que pour tout $m \in \mathcal{M}$

(H4) $[m, a] \in \mathcal{M}, \qquad [m, b] \in \mathcal{M}.$

On suppose enfon qu'il existe m_1, m_2 et $m_3 \in \mathcal{M}$ tel que

(H5) $[a, b] = am_1 + bm_2 + m_3.$