Stage de Master 2

effectué au Laboratoire de Mathématiques d'Orsay
sous la tutelle de Monsieur Lagoutière et de Monsieur Rousset
frederic.lagoutiere@math.u-psud.fr, frederic.rousset@math.u-psud.fr

Étude de l'équation d'Airy sous contrainte

par

Clémentine Courtès

Université Paris Sud 11
15 rue Georges Clemenceau 91405 Orsay Cedex, France

Je tiens à remercier mes deux encadrants de stage pour leur soutien continu et leur complémentarité.
J'ai été particulièrement touchée par la grande disponibilité de Monsieur Lagoutière, par sa gentillesse et par ses compétences pédagogiques.
Je remercie également beaucoup Monsieur Rousset pour la qualité et le niveau de nos discussions. Ses interventions très pertinentes m'ont été d'une grande aide.
Je suis ravie de poursuivre en thèse avec ce binôme.

Table des matières

Introduction

Résumé

Dans ce mémoire, nous étudions, tout d'abord d'un point de vue analyse numérique puis en rajoutant une contrainte $u \geq -1$, l'équation d'Airy monodimentionnelle, $\partial_t u + \partial_x^3 u = 0$, et imposons à la solution d'être périodique en espace. Après avoir rappelé quelques propriétés (conservation de l'Hamiltonien, régularité du noyau), nous généralisons l'étude numérique à des équations dispersives à dérivées en espace d'ordre impair quelconque $\partial_t u + \partial_x^{2p+1} u = 0$ et démontrons la convergence des schémas numériques pourvu que le pas de temps soit comparé au pas d'espace à la puissance $2p+1$. Nous concluons ensuite sur le système contraint. La contrainte étant, a priori, incompatible avec la vérification de l'équation d'Airy, nous nous ramenons à considérer cette équation comme la formulation variationnelle d'un problème de minimisation. Introduire $u \geq -1$ revient donc à résoudre un problème d'optimisation sous contrainte.

Lorsqu'il s'agit de modéliser des écoulements de fluides, la théorie des systèmes hyperboliques est le plus souvent mise à contribution. Les équations d'Euler sont alors les équations privilégiées pour décrire le mouvement d'un fluide, à partir de sa vitesse, de sa pression et de sa densité entre autre. Cependant, cette théorie reste inachevée pour beaucoup d'écoulements. En effet pour les écoulements sanguins par exemple, elle ne peut, à elle seule, expliquer les phénomènes comme l'absence de chocs. Il est donc nécessaire de prendre en compte un autre aspect : le caractère dispersif des équations. La dispersion fait ainsi intervenir des dérivées spatiales d'ordre 3 et le système d'Euler-Korteweg devient donc plus pertinent :

$$(EK) \begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + \rho^2) = \varepsilon \partial_x^3 \rho, \end{cases}$$

avec un écoulement monodimentionnel, ρ étant la densité et u la vitesse du fluide. Le système d'Euler-Korteweg (dont le problème de Cauchy fut initialement étudié par S.Benzoni-Gavage, R.Danchin et S.Descombes [Benzoni-Gavage *et al.*, 2006]) ne sera pas pris en compte tel quel dans ce mémoire : une première simplification consiste à ne considérer qu'une seule équation dispersive : l'équation de Korteweg-de Vries $\partial_t u + \partial_x^3 u + u\partial_x u = 0$. Puis la linéarisation de cette équation autour de zéro donne naissance à l'équation d'Airy, équation étudiée dans ce mémoire.

$$\partial_t u + \partial_x^3 u = 0. \qquad \text{(Airy)}$$

En réalité, les équations de modélisation seules ne suffisent pas à décrire l'écoulement : bien souvent, la configuration géométrique de l'écoulement doit être prise en compte (comme

la hauteur des tuyaux par exemple). Mathématiquement, cela se traduit par un ensemble de contraintes supplémentaires que la solution doit vérifier. Dans l'exemple de l'écoulement sanguin considéré plus haut (ou de tout autre écoulement biologique d'ailleurs), la contrainte sera la hauteur maximale des vaisseaux sanguins, hauteur que l'écoulement ne peut dépasser. Au niveau des équations, cela se traduit par l'assujettissement de la solution à rester dans un domaine borné que nous modéliserons par

$$(\text{Airy avec contrainte}) \begin{cases} \partial_t u + \partial_x^3 u = 0, \\ u \geq -1. \end{cases} \tag{1}$$

Comme la norme L^2 de la solution de l'équation d'Airy se conserve, la contrainte $u \geq -1$ est à différencier de la contrainte introduite par F.Berthelin et F.Bouchut [Berthelin et Bouchut, 2003] qui ont, de leur côté, étudié les équations de modélisation lorsqu'un réservoir d'eau ou une rivière déborde, il y a donc perte de masse dans ce cas.

Dans ce mémoire, nous proposons une étude de l'équation d'Airy monodimensionnelle et imposons à la solution d'être périodique en espace. L'aspect analyse numérique est particulièrment développé, avec une généralisation aux équations dispersives à des dérivés en espace d'ordre impair : $\partial_t u + \partial_x^{2p+1} u = 0$, $p \in \mathbb{N}$, pour lesquelles un calcul de consistance et de stabilité révèle la convergence des schémas numériques à condition que le pas de temps soit comparé au pas d'espace à la puissance $2p+1$. Une réflexion sur l'ordre de la convergence en fonction de la régularité plus ou moins forte de la condition initiale clôt cette analyse numérique. Ensuite, l'étude de l'équation d'Airy se pousuit avec le système (1). La contrainte est prise en compte de deux façons possibles. Une première méthode repose sur le formalisme hamiltonien de l'équation d'Airy et notamment sur la conservation de l'Hamiltonien pour reconstruire une solution contrainte de même Hamiltonien. La deuxième méthode reformule l'équation d'Airy sous forme d'un problème de minimisation, problème auquel il est plus facile, par la suite, de rajouter la contrainte $u \geq -1$. Il suffit pour cela de faire de l'optimisation sous contrainte.

Dans un premier temps, nous nous sommes attachés à rappeler les propriétés de l'équation d'Airy : conservation de la norme L^2, conservation de l'Hamiltonien $H(u) = \int_0^L \frac{1}{2}(\partial_x u(t,x))^2 dx = \frac{1}{2}||\partial_x u(t,\cdot)||^2_{L^2([0,L])}$, noyau d'Airy exprimé en terme de fonction d'Airy $\mathrm{Ai}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{ix\xi + i\frac{\xi^3}{3}} d\xi$, régularité $\mathscr{C}^\infty(\mathbb{R})$ de la fonction d'Airy, *etc.* Une comparaison de quelques schémas numériques s'ensuit : θ-schémas aux différences finies centrées, décentrées à droite, décentrées à gauche, mais aussi un schéma symplectique qui assure numériquement la conservation de l'Hamiltonien . Pour cela, nous avons adapté, à l'équation d'Airy, le schéma présenté par H.Kanazawa, T.Matsuo et T.Yaguchi [Kanazawa *et al.*, 2012]. Dans une seconde partie, une généralisation aux équations dispersives à dérivée en espace d'ordre impair quelconque $\partial_t u + \partial_x^{2p+1} u = 0$, $p \in \mathbb{N}$ nous permet de montrer là encore, la régularité $\mathscr{C}^\infty(\mathbb{R})$ de leur noyau et d'étudier la convergence des schémas numériques. La contrainte $u \geq -1$ sera introduite en troisième partie. Une projection qui conserve l'Hamiltonien, (*i.e.* une « symétrie » par rapport à $[u_j^n, u_{j+k}^n]$ si $u_{j+1}^n, ..., u_{j+k-1}^n$ sont plus petits que -1), est la première façon de prendre en compte cette contrainte. L'utilisation de la formulation variationnelle d'un problème de minimisation sous contrainte est une deuxième façon, peut-être plus rigoureuse, de procéder.

Chapitre 1

Généralités sur l'équation d'Airy

Nous étudions la régularité de la solution de l'équation d'Airy : contrairement à l'équation de Burgers par exemple, l'effet régularisant assure ici à la solution d'être au moins dans le même espace de Sobolev que u_0. Nous comparons ensuite différents schémas numériques de résolution de cette équation.

Dans ce premier chapitre, nous étudions le problème de Cauchy pour l'équation d'Airy suivant :

$$(\mathcal{P}_{\text{init non contraint}}) \begin{cases} \partial_t u + \partial_x^3 u = 0, & \text{(1.1a)} \\ u_{|_{t=0}} = u_0, & \text{(1.1b)} \end{cases}$$

Nous recherchons une solution u périodique en espace, de période L. De plus, u_0 est supposée à moyenne nulle, de sorte que $u(t,\cdot)$ le soit également, pour tout $t \in [0,T]$. En effet, en intégrant formellement l'équation d'Airy sur $[0,t]\times[0,L]$ nous obtenons la conservation de $\int_0^L u(t,x)dx$.

Nous introduisons les espaces périodiques : $H^k_{\text{per}}([0,L])$, pour $k \in \mathbb{N}$ (dans lesquels nous pouvons voir les fonctions comme égales à leur série de Fourier).

Définition. L'espace de Sobolev $L-$périodique d'ordre k, noté $H^k_{\text{per}}([0,L])$, est l'ensemble des fonctions $L-$périodiques sur $\mathbb{R}$, appartenant à $L^2([0,L])$ et telles que leurs k premières dérivées au sens des distributions appartiennent aussi à $L^2([0,L])$. Nous le munissons de la norme

$$||u||_{H^k_{\text{per}}} = (\sum_{l\leq k} ||u^{(l)}||^2_{L^2([0,L])})^{\frac{1}{2}}.$$

Les fonctions appartenant aux espaces $H^k_{\text{per}}([0,L])$ sont caractérisées par le comportement de leur série de Fourier.

1 Régularité de la solution

Tout d'abord, familiarisons-nous avec des solutions non forcément périodiques.
Par la suite, nous utiliserons la transformée de Fourier en espace définie comme suit :

- Si $u \in L^2(\mathbb{R})$ nous notons sa transformée de Fourier $\widehat{u}(\xi) = \int_{\mathbb{R}} u(x)e^{-i\xi x}dx$.
- Si u est une distribution tempérée ($u \in \mathscr{S}'(\mathbb{R})$), nous notons $\mathcal{F}(u)$ sa transformée de Fourier, définie par $\forall \varphi \in \mathscr{S}(\mathbb{R}) < \mathcal{F}(u), \varphi > = < u, \widehat{\varphi} >$.

L'utilisation de la transformée de Fourier assure l'existence et l'unicité d'une solution au problème de Cauchy ($\mathcal{P}_{\text{init non contraint}}$). En effet, en « appliquant » la transformée de Fourier à l'équation d'Airy, nous obtenons

$$\partial_t \widehat{u}(t,\xi) + (i\xi)^3 \widehat{u}(t,\xi) = 0,\ \forall (t,\xi) \in [0,T]\times\mathbb{R}.$$

Soit, en résolvant cette équation :

$$\widehat{u}(t,\xi) = \widehat{u}_0(\xi) e^{i\xi^3 t}, \forall (t,\xi) \in [0,T]\times\mathbb{R}.$$

D'où, $\forall (t,x) \in [0,T]\times\mathbb{R}, u(t,x) = [u_0 \star \mathcal{F}^{-1}(e^{i\xi^3 t})](x)$, où $\star$ est le produit de convolution en la variable d'espace.

Définition 1. *Nous définissons la fonction d'Airy comme suit,*

$$\mathrm{Ai}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\frac{\xi^3}{3}+i\xi x} d\xi, \forall x \in \mathbb{R},$$

ainsi que le noyau d'Airy par

$$\mathrm{K}_t(x) = \frac{1}{|\sqrt[3]{3t}|}\mathrm{Ai}(\frac{x}{\sqrt[3]{3t}}), \forall x \in \mathbb{R}.$$

La solution au problème de Cauchy s'écrit :

$$u(t,x) = [u_0 \star \mathrm{K}_t](x), \forall (t,x) \in [0,T]\times\mathbb{R}.$$

Proposition 1 (Régularité du noyau)**.** *Le noyau d'Airy est de classe* $\mathscr{C}^\infty(\mathbb{R})$.

Démonstration. La preuve se trouve par exemple dans [Zuily et Queffélec, 2007]. ■

Remarque. *Si nous considérons une équation générale* $\partial_t u + \partial_x^{2p+1} u = 0$*, avec p quelconque dans* $\mathbb{N}$*. Nous définissons de même l'équivalent de la fonction d'Airy,*

$$\mathrm{Ai}^{\mathrm{gen}}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\frac{\xi^{2p+1}}{2p+1}+i\xi x} d\xi,$$

nous pouvons alors définir la solution comme

- *si p est pair,* $u(t,x) = [u_0 \star K_t^{\mathrm{gen}}(-\cdot)](x)$,
- *si p est impair,* $u(t,x) = [u_0 \star K_t^{\mathrm{gen}}](x)$,

où le noyau K_t^{gen} *vaut*

$$K_t^{\mathrm{gen}}(x) = \frac{1}{|(2p+1)t|^{\frac{1}{2p+1}}}\mathrm{Ai}^{\mathrm{gen}}\Big(\frac{x}{((2p+1)t)^{\frac{1}{2p+1}}}\Big).$$

La fonction $\mathrm{Ai}^{\mathrm{gen}}$ *est, elle aussi,* $\mathscr{C}^\infty(\mathbb{R})$*. En effet, en adaptant la preuve fournie dans le livre de C.Zuily et H.Queffélec [Zuily et Queffélec, 2007], nous montrons que* $2\pi\mathrm{Ai}^{\mathrm{gen}}$ *est limite uniforme lorsque* ε *converge vers* 0*, de la solution de* $y_\varepsilon^{(2p)}(x) + (-1)^p x y_\varepsilon(x) + (-1)^p 2\varepsilon y_\varepsilon'(x) = 0, \forall x \in \mathbb{R}$*. Ce qui permet de montrer que* $2\pi\mathrm{Ai}^{\mathrm{gen}}$ *est solution de* $y^{(2p)}(x) + (-1)^p x y(x) = 0$*, et que* $\mathrm{Ai}^{\mathrm{gen}}$ *a bien la régularité souhaitée.*

Proposition 2. Ai *est à décroissance rapide sur* $\mathbb{R}^+$.

Démonstration. Cette preuve se fait en deux étapes. Tout d'abord, montrons par récurrence l'égalité suivante : $\forall N \in \mathbb{N}$, il existe $\mathcal{P}(y)$, polynôme en y de degré N, tel que

$$\mathrm{Ai}(x) = \frac{(-1)^N}{2\pi i^N x^{\frac{3N-1}{2}}} \int_{\mathbb{R}} e^{ix^{\frac{3}{2}}(y+\frac{y^3}{3})} \frac{\mathcal{P}(y)}{(1+y^2)^{2N}} dy.$$

POUR $N = 0$: En faisant le changement de variable $\xi = \sqrt{x}y$ dans l'intégrale, l'équation est vérifiée pour $\mathcal{P}(y) = 1$.

SUPPOSONS LA RELATION VRAIE POUR N : En remarquant que $\frac{\mathrm{d}}{\mathrm{dy}}\left(e^{ix^{\frac{3}{2}}(y+\frac{y^3}{3})}\right) = ix^{\frac{3}{2}}(1+y^2)e^{ix^{\frac{3}{2}}(y+\frac{y^3}{3})}$, nous avons, d'après l'hypothèse de récurrence

$$\begin{aligned}\mathrm{Ai}(x) &= \frac{(-1)^N}{2\pi i^N x^{\frac{3N-1}{2}}} \int_{\mathbb{R}} \frac{\left(e^{ix^{\frac{3}{2}}(y+\frac{y^3}{3})}\right)'}{ix^{\frac{3}{2}}(1+y^2)} \frac{\mathcal{P}(y)}{(1+y^2)^{2N}} dy \\ &= \frac{(-1)^{N+1}}{2\pi i^{N+1} x^{\frac{3(N+1)-1}{2}}} \int_{\mathbb{R}} e^{ix^{\frac{3}{2}}(y+\frac{y^3}{3})} \left(\frac{\mathcal{P}(y)}{(1+y^2)^{2N+1}}\right)' dy.\end{aligned}$$

Dans une seconde étape, nous remarquons que l'égalité précédente implique que, pour tout N dans $\mathbb{N}$, il existe une constante C_N telle que

$$\forall x > 0 \quad |\mathrm{Ai}(x)| \leq \frac{C_N}{2\pi x^{\frac{3N-1}{2}}}.$$

■

Proposition 3. *Pour tout $k \in \mathbb{R}$, la norme $H^k([0,L])$ de la solution d'Airy est conservée.*

Démonstration. La norme $L^2([0,L])$ est conservée (en multipliant formellement par u puis en intégrant sur $[0,T]\times[0,L]$). Comme $||\cdot||_{H^k([0,L])} = ||(1+|\xi|^2)^{\frac{k}{2}}\widehat{\cdot}||_{L^2([0,L])}$, la norme $H^k([0,L])$ est bien conservée. ■

Corollaire 1. *L'Hamiltonien $H(u) = \int_0^L \frac{|\partial_x u|^2}{2}(t,x)dx$ est conservé au cours du temps.*

Remarque. *Pour les solutions périodiques, nous avons aussi existence et unicité de la solution puisque, si $u_0(x) = \sum_{k\in\mathbb{Z}} c_k(0) e^{\frac{2ik\pi}{L}x}$, alors la solution u vérifie $u(t,x) = \sum_{k\in\mathbb{Z}} c_k(0) e^{i(\frac{2k\pi}{L})^3 t + i\frac{2k\pi}{L}x}$, $\forall (t,x) \in [0,T]\times[0,L]$.*

2 Schémas numériques

La première étape de résolution numérique du problème de Cauchy $(\mathcal{P}_{\text{init non contraint}})$ est de semi-discrétiser en temps, avec un θ-schéma ici, pour ne manipuler que des fonctions d'une seule variable : la variable spatiale. La fonction u sera donc connue seulement en les temps $t^n = n\Delta t$ pour $n \in \mathbb{N}$, Δt étant le pas de temps. Nous approchons donc $u(t^n,\cdot)$ par $u^n(\cdot)$, solution de l'équation d'une seule variable

$$\frac{u^{n+1}-u^n}{\Delta t} + (\theta u^{n+1} + (1-\theta)u^n)''' = 0, \text{ avec } \theta \in [0,1].$$

Il faut ensuite déterminer une approximation de la dérivée troisième en espace : utilisation des différences finies (décentrées à droite, centrées, décentrées à gauche), choix d'un schéma symplectique... Leur comparaison fera l'objet des paragraphes suivants.

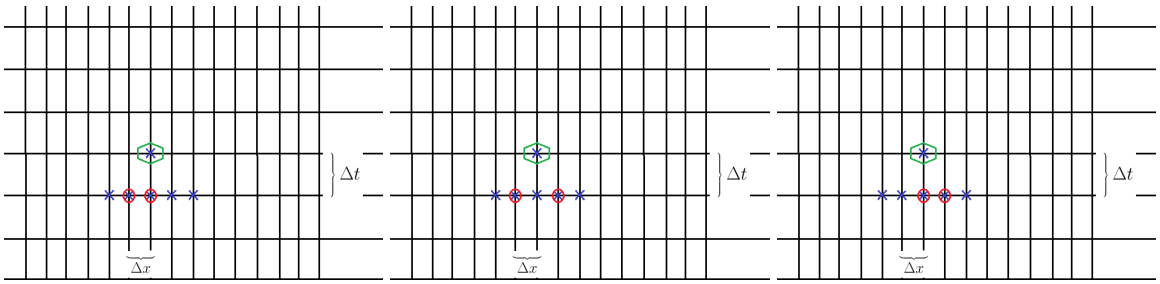
Approximation de la dérivée troisième en espace par différences finies

Nous présentons dans le tableau suivant les schémas aux différences finies décentré à gauche, centré et décentré à droite.

	décentré à gauche	centré	décentré à droite
$u'(x_j) \approx$	$\frac{u_j - u_{j-1}}{\Delta x}$	$\frac{u_{j+1} - u_{j-1}}{2\Delta x}$	$\frac{u_{j+1} - u_j}{\Delta x}$
$u''(x_j) \approx$		$\frac{u_{j+1} - 2u_j + u_{j-1}}{(\Delta x)^2}$	
$u'''(x_j) \approx$	$\frac{u_{j+1} - 3u_j + 3u_{j-1} - u_{j-2}}{(\Delta x)^3}$	$\frac{u_{j+2} - 2u_{j+1} + 2u_{j-1} - u_{j-2}}{2(\Delta x)^3}$	$\frac{u_{j+2} - 3u_{j+1} + 3u_j - u_{j-1}}{(\Delta x)^3}$
$u^{(4)}(x_j) \approx$		$\frac{u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2}}{(\Delta x)^4}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$u^{(2p)}(x_j) \approx$		$\frac{\sum_{k=0}^{2p} \binom{2p}{k} (-1)^k u_{p-k+j}}{(\Delta x)^{2p}}$	
$u^{(2p+1)}(x_j) \approx$	$\frac{\sum_{k=0}^{2p+1} \binom{2p+1}{k} (-1)^k u_{p-k+j}}{(\Delta x)^{2p+1}}$	$\frac{u_{p+j+1} + \left(\sum_{k=0}^{2p} \left[\binom{2p+1}{k} - \binom{2p+1}{k+1} \right] (-1)^k u_{p-k+j} \right) - u_{j-p-1}}{2(\Delta x)^{2p+1}}$	$\frac{\sum_{k=0}^{2p+1} \binom{2p+1}{k} (-1)^k u_{p-k+j+1}}{(\Delta x)^{2p+1}}$

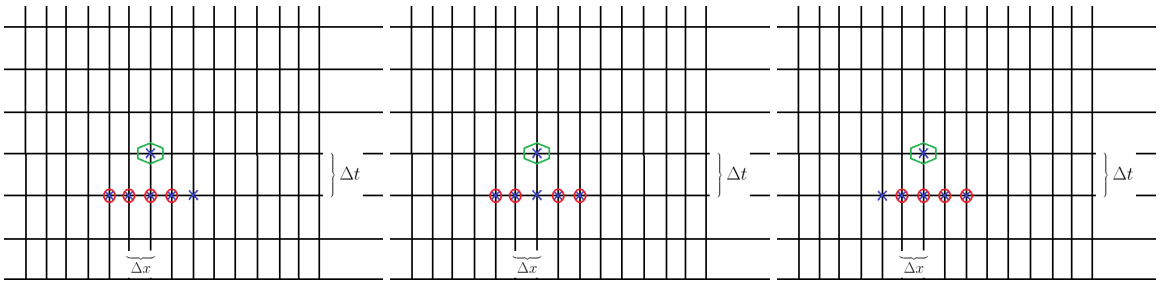
TABLE 1.1: Différences finies décentrées à gauche, centrées et décentrées à droite

Pour mieux visualiser la différence de chaque schéma, nous avons affiché la dépendance en j dans les figures suivantes. Les abscisses représentent l'espace ($x_j = j\Delta x,\ j \in \mathbb{Z}$), les ordonnées, le temps ($t^n = n\Delta t,\ n \in \lfloor \frac{T}{\Delta t} \rfloor$). Dans chaque figure, le point entouré en vert (u_j^{n+1}) est le point à construire à partir des points entourés en rouge ($u_{j-p-1}, ..., u_j, ..., u_{j+p+1}$).



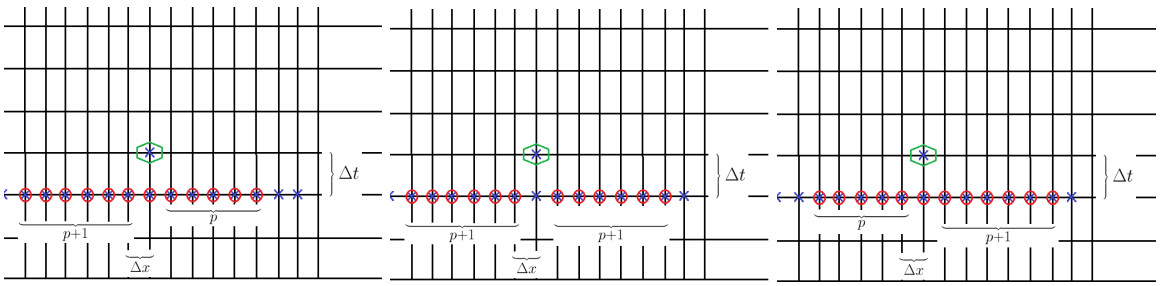
(a) schéma décentré à gauche (b) schéma centré (c) schéma décentré à droite

FIGURE 1.1: Approximation de $u'(x_j)$



(a) décentré à gauche (b) centré (c) décentré à droite

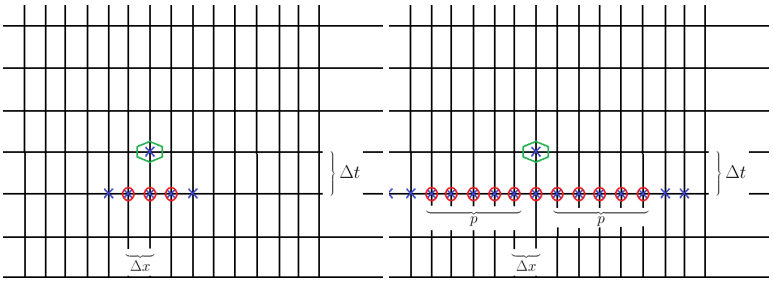
FIGURE 1.2: Approximation de $u'''(x_j)$



(a) décentré à gauche (b) centré (c) décentré à droite

FIGURE 1.3: Approximation de $u^{(2p+1)}(x_j)$

Pour les dérivées d'ordre pair, nous n'utiliserons qu'une seule formule : celle centrée.



(a) Approximation de $u''(x_j)$ (b) Approximation de $u^{(2p)}(x_j)$

FIGURE 1.4: Approximation des dérivées d'ordre pair

L'équation d'Airy totalement discrétisée devient donc, pour le schéma aux différences finies décentré à droite par exemple :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\theta\left(\frac{u_{j+2}^{n+1}-3u_{j+1}^{n+1}+3u_j^{n+1}-u_{j-1}^{n+1}}{(\Delta x)^3}\right)+(1-\theta)\left(\frac{u_{j+2}^{n}-3u_{j+1}^{n}+3u_j^{n}-u_{j-1}^{n}}{(\Delta x)^3}\right)=0,$$

$\forall (n,j)\in[\![0,N]\!]\times\mathbb{Z}$ avec $N=\lfloor\frac{T}{\Delta t}\rfloor$.

Approximation de la dérivée troisième en espace par schéma symplectique

Les trois schémas numériques précédents ne conservent pas l'Hamiltonien de l'équation d'Airy. Par contre, le schéma présenté dans [Kanazawa *et al.*, 2012] est un schéma symplectique pour l'équation de Korteweg-de Vries ; nous l'avons modifié et adapté à l'équation étudiée ici pour qu'il conserve l'Hamiltonien d'Airy.

Il s'agit d'un schéma de Crank-Nicolson (*i.e.* $\theta=\frac{1}{2}$), dans lequel les dérivées en espace sont approchées par des différences finies compactes, notées $\delta_c^{<1>}$, ce qui fournit les approximations suivantes :

$u'(x_j)$ est approché par	$[\delta_c^{<1>}(U)]_j := \frac{3}{2}\frac{u_{j+1}-u_{j-1}}{2\Delta x}-\frac{3}{5}\frac{u_{j+2}-u_{j-2}}{4\Delta x}+\frac{1}{10}\frac{u_{j+3}-u_{j-3}}{6\Delta x}$
$u''(x_j)$ est approché par	$\delta_c^{<1>}\circ\delta_c^{<1>}(U)$
$\vdots$	$\vdots$
$u^{(p)}(x_j)$ est approché par	$\underbrace{\delta_c^{<1>}\circ\delta_c^{<1>}\circ\ldots\circ\delta_c^{<1>}}_{p \text{ fois}}(U)$

TABLE 1.2: Approximation des dérivées pour le schéma symplectique

Dans le tableau précédent, U représente le vecteur $(u_j)_{j\in\mathbb{Z}}$, et dans ce qui suit, $U^n=(u_j^n)_{j\in\mathbb{Z}}$.

Le schéma s'écrit donc :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\left[\delta_c^{<1>}\circ\delta_c^{<1>}\circ\delta_c^{<1>}\left(\frac{U^{n+1}+U^n}{2}\right)\right]_j=0,$$
$$\forall(n,j)\in[\![0,N]\!]\times\mathbb{Z}.$$

Proposition 4. *Soit H_{num} l'Hamitonien numérique défini par*

$$H_{\text{num}}(U)=\sum_{k=1}^{J}\frac{1}{2}\left(\delta_c^{<1>}(U)\right)_k^2\Delta x$$

où $x_1,\ldots,x_J$ est une discrétisation de $[0,L]$. Alors, cet Hamiltonien est conservé par le schéma symplectique :

$$H_{\text{num}}(U^n)=H_{\text{num}}(U^{n+1}),\forall n\in[\![0,N]\!].$$

Démonstration. La preuve repose sur l'égalité suivante (voir [Kanazawa *et al.*, 2012]) :

$$\sum_{k=1}^{J}V_k\left(\delta_c^{<1>}(W)\right)_k\Delta x=-\sum_{k=1}^{J}\left(\delta_c^{<1>}(V)\right)_kW_k\Delta x.$$

Soit $n\in[\![0,N]\!]$,

$$\begin{aligned}
&\sum_{k=1}^{J}\frac{1}{2}\left[\left(\delta_c^{<1>}(U^{n+1})\right)_k^2-\left(\delta_c^{<1>}(U^n)\right)_k^2\right]\Delta x\\
&=\sum_{k=1}^{J}\frac{1}{2}\left[\left(\delta_c^{<1>}(U^{n+1})\right)_k-\left(\delta_c^{<1>}(U^n)\right)_k\right]\left[\left(\delta_c^{<1>}(U^{n+1})\right)_k+\left(\delta_c^{<1>}(U^n)\right)_k\right]\Delta x\\
&=\sum_{k=1}^{J}\left(\delta_c^{<1>}\left(U^{n+1}-U^n\right)\right)_k\left(\delta_c^{<1>}\left(\frac{U^{n+1}+U^n}{2}\right)\right)_k\Delta x\\
&=-\sum_{k=1}^{J}\left(U^{n+1}-U^n\right)_k\left(\delta_c^{<1>}\circ\delta_c^{<1>}\left(\frac{U^{n+1}+U^n}{2}\right)\right)_k\Delta x\\
&=\sum_{k=1}^{J}\left[\delta_c^{<1>}\circ\left(\delta_c^{<1>}\circ\delta_c^{<1>}\left(\frac{U^{n+1}+U^n}{2}\right)\right)\right]_k\left(\delta_c^{<1>}\circ\delta_c^{<1>}\left(\frac{U^{n+1}+U^n}{2}\right)\right)_k\Delta x\Delta t\\
&=0.
\end{aligned}$$

■

Comparaison par rapport à la « solution exacte »

Afin de vérifier le bon comportement de ces différents schémas numériques, nous avons calculé la solution exacte de l'équation d'Airy en partant d'une fonction chapeau périodisée (donnée initiale pour laquelle les coefficients de Fourier sont facilement calculables).

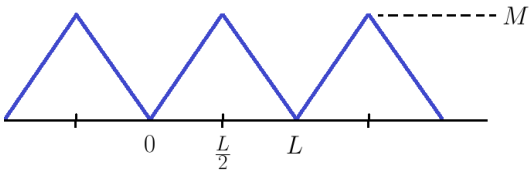


FIGURE 1.5: Donnée initiale chapeau périodisée

La série de Fourier d'une fonction u_0 chapeau périodisée de période L vaut

$$S(u_0)(x) = \frac{M}{2} - \frac{4M}{\pi^2} \sum_{m=1}^{\infty} \frac{\cos(\frac{2\pi}{L} x(2m-1))}{(2m-1)^2}.$$

La série de Fourier de la solution de l'équation d'Airy avec donnée initiale u_0 est donnée par $S(u)(t,x) = \sum_{m\in\mathbb{Z}} c_m(u)(t) e^{\frac{i2\pi mx}{L}}, \forall (t,x) \in [0,T]\times\mathbb{R}$, où les coefficients $c_m(u)$ vérifient la relation $c_m(u)'(t) + (im\frac{2\pi}{L})^3 c_m(u)(t) = 0$, soit

$$c_m(u)(t) = c_m(u_0) e^{itm^3 \left(\frac{2\pi}{L}\right)^3}.$$

D'où

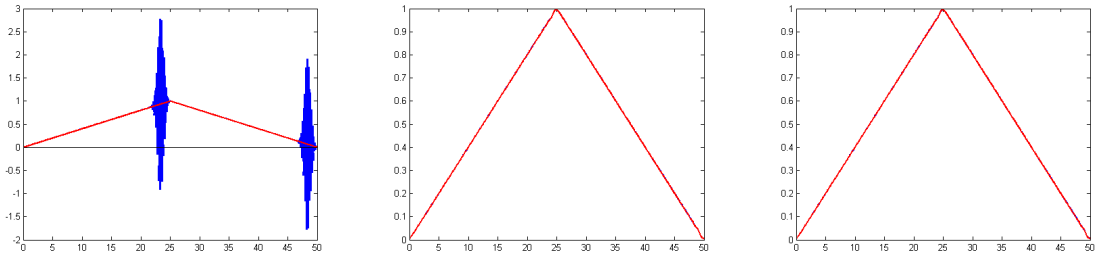
$$S(u)(t,x) = \frac{M}{2} - \frac{4M}{\pi^2} \sum_{m=1}^{\infty} \frac{\cos\left(\left(\frac{2\pi}{L}(2m-1)\right)^3 t + (2m-1)\frac{2\pi}{L}x\right)}{(2m-1)^2}.$$

Comparons le vecteur discret obtenu avec les schémas numériques à cette solution exacte (en réalité nous comparons les schémas numériques à $S_{500}(u)(t,x)$, *i.e.* la série de Fourier tronquée à la fréquence 500 de u). Les résultats font l'objet des figures suivantes. La série de Fourier $S_{500}(u)(t,x)$ est représentée en rouge, la solution discrète en bleu.

Paramètres : la période vaut $L = 50$, le nombre de maille $J = 1000$ (soit un pas d'espace $\Delta x = \frac{1}{20}$), le temps final vaut $T = 0,01$ et le pas de temps $\Delta t = \frac{1}{5000}$. Nous appelons **« nombre de Courant »** le rapport $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}}$, qui joue un rôle important dans la stabilité des schémas numériques (c.f le chapitre 2). Ici, puisque $p = 1$ pour l'équation d'Airy, ce nombre de Courant vaut

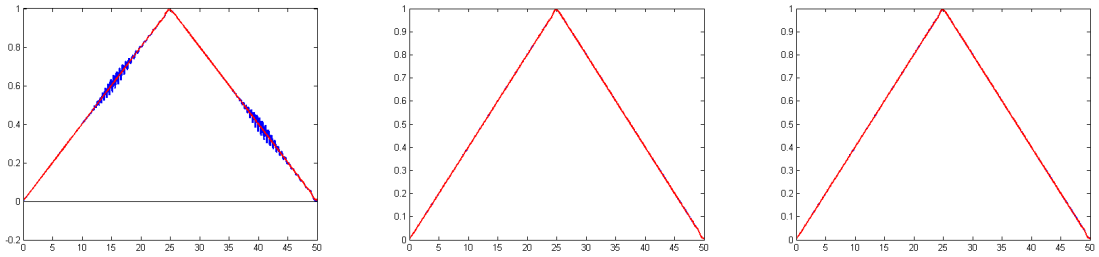
$$\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} = 6,4.$$

Conditions aux limites : Puisque nous supposons $u(t,\cdot)$ L-périodique, nous imposons numériquement : $u_0^n = u_J^n$ et $u_{J+1}^n = u_1^n$, pour tout $n \in [\![0,N]\!]$.



(a) Décentré à gauche (b) Centré (c) Décentré à droite

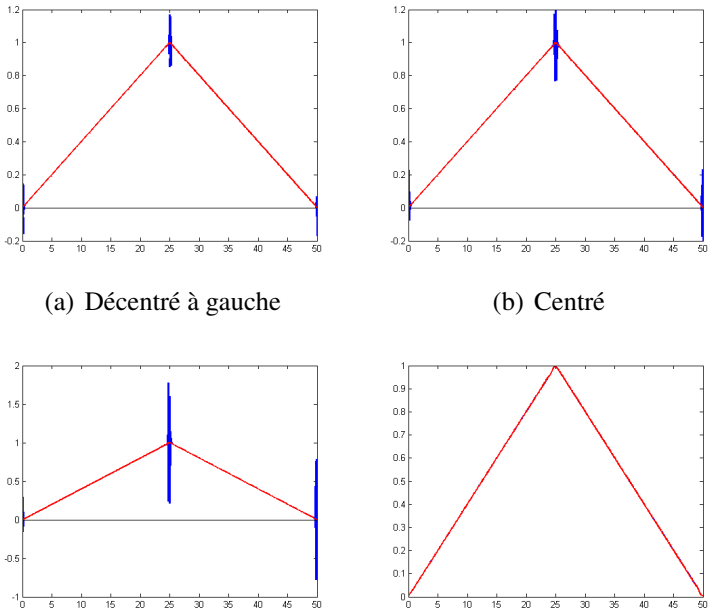
FIGURE 1.6: Schéma de Crank-Nicolson, $\theta = \frac{1}{2}$



(a) Décentré à gauche (b) Centré (c) Décentré à droite

FIGURE 1.7: Schéma implicite, $\theta = 1$

Sur la figure suivante, nous utilisons toujours les mêmes paramètres sauf pour l'image (d), pour laquelle nous changeons le nombre de mailles : $J = 538$. Pour cette image, le nombre de Courant vaut donc $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} = 0,997$.

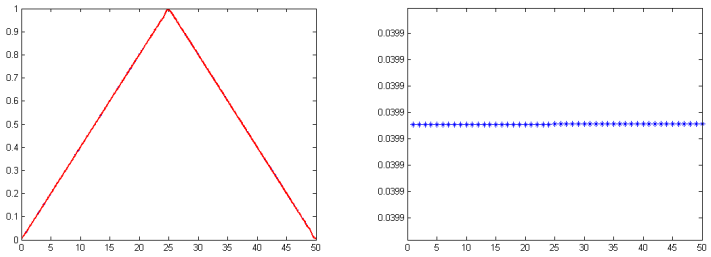


(a) Décentré à gauche (b) Centré

(c) Decentre a droite, avec $\frac{4\Delta t}{(\Delta x)^3} > 1$ (d) Decentre a droite, avec $\frac{4\Delta t}{(\Delta x)^3} < 1$

FIGURE 1.8: Schéma explicite, $\theta = 0$

Au vu de ces figures, il semblerait qu'approcher la dérivée en espace par des différences finies décentrées à gauche ne permette pas de capturer la solution. L'utilisation des différences finies décentrées à droite suppose une condition de Courant-Friedrichs-Lewy (CFL), comme l'indiquent les images (c) et (d) de la figure 1.8, du type « il est nécessaire que $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} \leq 1$ pour avoir la stabilité (avec $p = 1$ car nous étudions l'équation d'Airy) ». Nous étudierons plus en détails au chapitre 2 les conditions de stabilité L^2 pour chaque type de schéma.



(a) Schéma symplectique (b) Conservation de H_{num}

FIGURE 1.9: Schéma symplectique

L'Hamiltonien numérique est bien constant et le schéma symplectique semble précis, pour $T = 0,01$, $\Delta t = \frac{1}{5000}$, $L = 50$ et $J = 1000$ (soit $\Delta x = \frac{1}{20}$).

Comparaison des différents schémas

Nous effectuons deux types de comparaison : une comparaison de différents maillages pour le même schéma, puis une comparaison des schémas entre eux sur un même maillage. Les résultats sont représentés sur les figures suivantes. Nous utilisons une donnée initiale sous forme de créneau suivante :

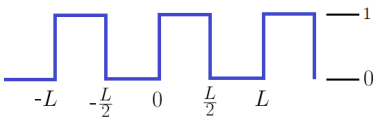
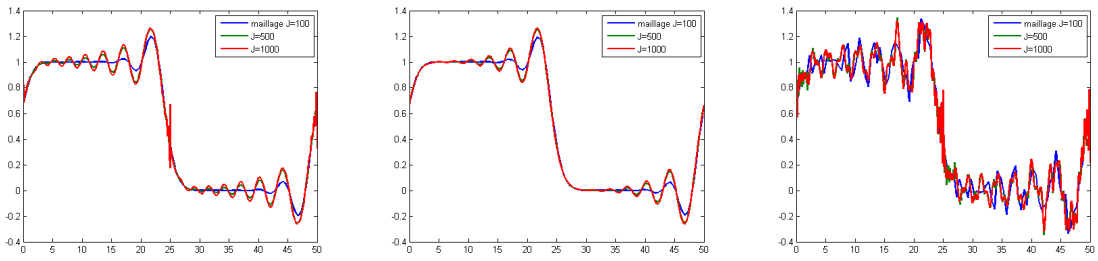


FIGURE 1.10: Donnée initiale sous forme d'un créneau périodisé



(a) Crank-Nicolson décentré à droite, $\theta = \frac{1}{2}$, (b) Implicite décentré à droite, $\theta = 1$ (c) Symplectique

FIGURE 1.11: Comparaison sur différents maillages

La figure 1.11 a été réalisée avec $T = 1$, un pas de temps $\Delta t = \frac{1}{50}$, $L = 50$ et différents J :

- $J = 100$ (donc $\Delta x = \frac{1}{2}$) pour la courbe bleue, soit un nombre de Courant égal à $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} = 0,64$,
- $J = 500$ (donc $\Delta x = \frac{1}{10}$) pour la courbe verte, soit $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} = 80$,

- $J = 1000$ (donc $\Delta x = \frac{1}{20}$) pour la courbe rouge, soit $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}} = 640$.

Nous n'avons pas affiché ici le résultat donné par le schéma explicite car ce pas de temps avec les deux derniers pas d'espace ne vérifient pas la condition de CFL. Ce schéma explicite est développé en fin de paragraphe.

Pour comparer les schémas les uns par rapport aux autres, nous fixons le nombre de mailles à $J = 1000$ (la période vaut toujours $L = 50$, le temps final $T = 1$, et le pas de temps $\Delta t = \frac{1}{50}$). Pour ces paramètres, le nombre de Courant vaut 640.

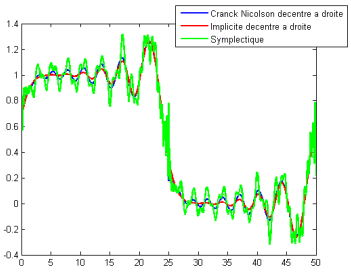


FIGURE 1.12: Comparaisons des schémas numériques entre eux

La courbe bleue représente le schéma de Crank-Nicolson avec des différences finies décentrées à droite. Elle est visuellement proche de la courbe rouge (schéma implicite avec dérivée décentrée à droite). Par contre, le schéma symplectique est beaucoup plus oscillant (courbe verte sur la figure 1.12).

Pour voir le comportement du schéma explicite décentré à droite, nous avons changé le pas de temps en fonction du pas d'espace de sorte à ce que le nombre de Courant soit toujours égal à 1. Plus précisément, nous prenons toujours $L = 50,\ T = 1$, mais pour chaque J différents ($J = 100,\ 500$ ou 1000) Δt vaut maintenant $\Delta t = \frac{(\Delta x)^3}{4}$.

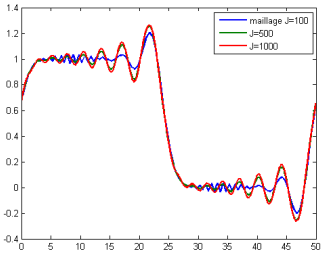


FIGURE 1.13: Schéma explicite avec dérivée décentrée à droite sur différents maillages, avec la condition de CFL respectée

Chapitre 2

Étude de la convergence des schémas numériques

Nous déterminons l'ordre de convergence des schémas numériques pour l'équation d'Airy, et généralisons ce calcul à une équation quelconque du type $\partial_t u + \partial_x^{2p+1} u = 0$. Une distinction en fonction de la régularité de la donnée initiale u_0 est effectuée.

1 Stabilité l^2

Nous travaillerons dans ce chapitre sur tout $\mathbb{R}$ et considérerons donc que la solution discrète U^n du schéma est définie pour tout $j \in \mathbb{Z}$.

Rappels :

- Nous posons toujours $U^n = (u_j^n)_{j\in\mathbb{Z}}$ dans $l^2(\mathbb{Z})$ que nous munissons de la norme suivante : $||U^n||_{l^2}^2 := \sum_{j\in\mathbb{Z}} |u_j^n|^2 \Delta x$.
- Pour $(u_j^n)_{j\in\mathbb{Z}} \in l^2(\mathbb{Z})$, nous notons $\widehat{U^n}(\xi) := \sum_{k\in\mathbb{Z}} u_k^n e^{2i\pi k\xi}$ pour $\xi \in [0,1]$. Alors $\widehat{U^n} \in L^2([0,1])$, où $L^2([0,1])$ est un espace de Hilbert que nous munissons de la base hilbertienne $(e^{2i\pi k\xi})_{k\in\mathbb{Z}}$. Nous avons de plus les égalités $u_k^n = \int_0^1 e^{-2i\pi k\xi} \widehat{U^n}(\xi) d\xi$, et $||U^n||_{l^2}^2 = \Delta x \int_0^1 |\widehat{U^n}(\xi)|^2 d\xi$.
- Enfin, nous définissons les deux opérateurs de décalage : $S^+U^n = (u_{j+1}^n)_{j\in\mathbb{Z}}$, avec $\widehat{S^+U^n}(\xi) = e^{-2i\pi\xi}\widehat{U^n}(\xi)$, et $S^-U^n = (u_{j-1}^n)_{j\in\mathbb{Z}}$, qui vérifie quant à lui $\widehat{S^-U^n}(\xi) = e^{2i\pi\xi}\widehat{U}(\xi)$, $\forall \xi \in [0,1]$. Plus généralement, $\widehat{(u_{l+j})_{j\in\mathbb{Z}}}(\xi) = e^{-2i\pi l\xi}\widehat{U^n}(\xi), \forall (l,\xi) \in \mathbb{Z}\times[0,1]$.

1.1 Le θ-schéma aux différences finies décentré à droite

D'après l'expression de l'approximation de $u^{(2p+1)}$ par les différences finies décentrées à droite (voir le tableau 1.1), nous travaillons avec le schéma suivant :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u_{p-k+j+1}^{n+1}}{(\Delta x)^{2p+1}}\right) + (1-\theta)\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u_{p-k+j+1}^{n}}{(\Delta x)^{2p+1}}\right) = 0,$$

$\forall (n,j) \in [\![0,N]\!] \times \mathbb{Z}$.

D'après le rappel précédent, nous avons $(u^n_{p-k+j+1})_{j\in\mathbb{Z}} = (S^+)^{p-k+1}U^n$, ce qui donne :

$$\widehat{U^{n+1}}(\xi)\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k+1)\xi}\right) = \widehat{U^n}(\xi)\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k+1)\xi}\right).$$

Lemme 1.

$$\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k+1)\xi} = e^{-i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}. \tag{2.1}$$

Démonstration. En effet,

$$\begin{aligned}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k+1)\xi} &= e^{2i\pi p\xi}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k\left(e^{-2i\pi\xi}\right)^{2p-k+1}\\ &= e^{2i\pi p\xi}\left(e^{-2i\pi\xi}-1\right)^{2p+1}\\ &= e^{-i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}.\end{aligned}$$

■

Donc, en notant $A_1(\xi)$ le coefficient d'amplification, nous obtenons :

$$\widehat{U^{n+1}}(\xi) = \underbrace{\frac{\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}e^{-i\pi\xi}(-i)^{2p+1}(2\sin(\pi\xi))^{2p+1}\right)}{\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}e^{-i\pi\xi}(-i)^{2p+1}(2\sin(\pi\xi))^{2p+1}\right)}}_{A_1(\xi)}\widehat{U^n}(\xi). \tag{2.2}$$

Deux cas sont à considérer, selon la parité de p.

1.1.1 Dérivée spatiale $\partial_x^{2p+1}u$ avec p pair

Dans le cas où p est pair, nous avons $(-i)^{2p+1}e^{-i\pi\xi} = -ie^{-i\pi\xi} = -i\cos(\pi\xi)-\sin(\pi\xi)$. La stabilité est donc assurée si $|A_1(\xi)|^2 \le 1$, ce qui se traduit par :

$$\begin{aligned}&\frac{-2(1-\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}(-\sin(\pi\xi)) + \frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}(-\sin(\pi\xi))^2\\ &\qquad + \frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2\\ &\le \frac{2\theta\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}(-\sin(\pi\xi)) + \frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}(-\sin(\pi\xi))^2\\ &\qquad + \frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2.\end{aligned}$$

Soit encore,

$$(1-\theta)+\frac{(1-\theta)^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p} \le -\theta+\frac{\theta^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p},$$

$$1+\frac{\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}-\frac{2\theta\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq 0,$$

$$\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}}(\sin(\pi\xi))^{2p}(1-2\theta)\leq -1.$$

Si $p\neq 0$, ceci est impossible, quel que soit $\xi\in[0,1]$,
(par exemple pour $\xi=0$, $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}}(\sin(\pi\xi))^{2p}(1-2\theta)=0$.)

Par contre, si $p=0$, $\frac{\Delta t}{(\Delta x)}(1-2\theta)\leq -1\Rightarrow\frac{1}{2}+\frac{\Delta x}{2\Delta t}\leq\theta$.

Conclusion :
- si p est pair et non nul, le θ-schéma aux différences finies décentré à droite est instable,
- si p est nul, le θ-schéma aux différences finies décentré à droite est stable sous la condition CFL : $\theta\geq\frac{1}{2}+\frac{\Delta x}{2\Delta t}$.

1.1.2 Dérivée spatiale $\partial_x^{2p+1}u$ avec p impair

Dans ce cas, $(-i)^{2p+1}e^{-i\pi\xi}=ie^{-i\pi\xi}=i\cos(\pi\xi)+\sin(\pi\xi)$. La condition $|A_1(\xi)|^2\leq 1$ se traduit donc par l'inéglité

$$\begin{aligned}\frac{-2(1-\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2&\\ +\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2&\\ \leq\frac{2\theta\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2&\\ +\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2&\end{aligned}$$

Ce qui, après simplifications, devient :

$$-(1-\theta)+\frac{(1-\theta)^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq\theta+\frac{\theta^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p},$$

$$-1+\frac{(1-2\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq 0,$$

$$\frac{\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}(1-2\theta)\leq 1.$$

La stabilité est donc assurée si $\frac{2^{2p}\Delta t}{(\Delta x)^{2p+1}}(1-2\theta)\leq 1$, soit $\theta\geq\frac{1}{2}-\frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t}$.
Conclusion : si p est impair, et si la condition CFL suivante est vérifiée

$$\theta\geq\frac{1}{2}-\frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t},$$

alors le θ-schéma aux différences finies décentré à droite est stable.

1.2 Le θ-schéma aux différences finies décentré à gauche

Dans ce schéma, la dérivée spatiale est approchée par :

$$\partial_x^{2p+1}u(t^n,x_j)=\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^n,x_{p-k+j})}{(\Delta x)^{2p+1}}+\underset{\Delta x\to 0}{O}(\Delta x)+\underset{\Delta t\to 0}{O}(\Delta t).$$

Ce qui conduit au schéma numérique suivant :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\theta\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u_{p-k+j}^{n+1}}{(\Delta x)^{2p+1}}\right)+(1-\theta)\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u_{p-k+j}^{n}}{(\Delta x)^{2p+1}}\right)=0,$$

$\forall (n,j)\in[\![0,N]\!]\times\mathbb{Z}$.

Ici encore, pour étudier la stabilité l^2 du schéma, nous utilisons les séries de Fourier en espace.

$$\widehat{U^{n+1}}(\xi)\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k)\xi}\right)$$
$$=\widehat{U^n}(\xi)\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(p-k)\xi}\right).$$

D'où

$$\widehat{U^{n+1}}(\xi)\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}e^{2i\pi\xi(p+1)}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi(2p+1-k)\xi}\right)$$
$$=\widehat{U^n}(\xi)\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}e^{2i\pi\xi(p+1)}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k e^{-2i\pi\xi(2p+1-k)}\right).$$

En utilisant la simplification détaillée au lemme 1et en notant $A_2(\xi)$ le coefficient d'amplification, nous obtenons :

$$\widehat{U^{n+1}}(\xi)=\underbrace{\frac{\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}e^{i\pi\xi}\left(-2i\sin(\pi\xi)\right)^{2p+1}\right)}{\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}e^{i\pi\xi}\left(-2i\sin(\pi\xi)\right)^{2p+1}\right)}}_{A_2(\xi)}\widehat{U^n}(\xi). \tag{2.3}$$

Remarque : La différence entre le θ-schéma décentré à droite et le θ-schéma décentré à gauche se situe au niveau du moins dans $e^{-i\pi\xi}$.

Ici encore, nous devons dissocier deux cas selon la parité de p.

1.2.1 Dérivée spatiale $\partial_x^{2p+1}u$ avec p pair

Lorsque p est pair : $(-i)^{2p+1}e^{i\pi\xi}=-ie^{i\pi\xi}=-i\cos(\pi\xi)+\sin(\pi\xi)$. La condition $|A_2(\xi)|^2\leq 1$ revient donc à :

$$\frac{-2(1-\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2$$
$$+\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2$$
$$\leq \frac{2\theta\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2$$
$$+\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2.$$

Ce qui devient :

$$-(1-\theta)+\frac{(1-\theta)^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq\theta+\frac{\theta^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p},$$
$$-1+\frac{(1-2\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq 0. \tag{2.4}$$

Si $2^{2p}\frac{(1-2\theta)\Delta t}{(\Delta x)^{2p+1}}\leq 1$, alors (2.4) est vérifié, quel que soit $\xi\in[0,1]$.

Conclusion : Si la condition de type CFL suivante $\theta\geq\frac{1}{2}-\frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t}$, et si p est pair, alors ce θ-schéma est stable en norme l^2.

1.2.2 Dérivée spatiale $\partial_x^{2p+1}u$ avec p impair

Dans ce cas, $(-i)^{2p+1}e^{i\pi\xi}=ie^{i\pi\xi}=i\cos(\pi\xi)-\sin(\pi\xi)$.
$|A_2(\xi)|^2\leq 1$ implique donc :

$$\frac{2(1-\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2$$
$$+\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2$$
$$\leq \frac{-2\theta\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p+1}\sin(\pi\xi)+\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\sin(\pi\xi)^2$$
$$+\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}(2\sin(\pi\xi))^{4p+2}\cos(\pi\xi)^2.$$

Ce qui revient à :

$$(1-\theta)+\frac{(1-\theta)^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq-\theta+\frac{\theta^2\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p},$$
$$1+\frac{(1-2\theta)\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}\leq 0. \tag{2.5}$$

Or l'inégalité $\frac{\Delta t}{(\Delta x)^{2p+1}}(2\sin(\pi\xi))^{2p}(1-2\theta)\leq -1$ est impossible pour tout $\xi\in[0,1]$ et p impair.

Conclusion : le θ-schéma aux différences finies décentré à gauche avec p impair est inconditionnellement instable.

1.3 Le θ-schéma aux différences finies centré

La discrétisation de l'équation $\partial_t u + \partial_x^{2p+1} u = 0$ par ce schéma puis l'introduction de la série de Fourier en espace conduisent à :

$$\widehat{U^{n+1}}(\xi)\left(1+\frac{\theta\Delta t}{2(\Delta x)^{2p+1}}e^{i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}+\frac{\theta\Delta t}{2(\Delta x)^{2p+1}}e^{-i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}\right)$$
$$=\widehat{U^{n}}(\xi)\left(1-\frac{(1-\theta)\Delta t}{2(\Delta x)^{2p+1}}e^{i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}-\frac{(1-\theta)\Delta t}{2(\Delta x)^{2p+1}}e^{-i\pi\xi}(-2i\sin(\pi\xi))^{2p+1}\right).$$

Soit

$$\widehat{U^{n+1}}(\xi)\left(1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}\cos(\pi\xi)(-2i\sin(\pi\xi))^{2p+1}\right)=\widehat{U^{n}}(\xi)\left(1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}\cos(\pi\xi)(-2i\sin(\pi\xi))^{2p+1}\right).$$

$$\widehat{U^{n+1}}(\xi)=\underbrace{\frac{1-\frac{(1-\theta)\Delta t}{(\Delta x)^{2p+1}}\cos(\pi\xi)(-2i\sin(\pi\xi))^{2p+1}}{1+\frac{\theta\Delta t}{(\Delta x)^{2p+1}}\cos(\pi\xi)(-2i\sin(\pi\xi))^{2p+1}}}_{A_3(\xi)}\widehat{U^{n}}(\xi).$$

$|A_3(\xi)|^2 \leq 1$ implique :

$$\frac{(1-\theta)^2(\Delta t)^2}{(\Delta x)^{4p+2}}\cos(\pi\xi)^2(2\sin(\pi\xi))^{4p+2}\leq\frac{\theta^2(\Delta t)^2}{(\Delta x)^{4p+2}}\cos(\pi\xi)^2(2\sin(\pi\xi))^{4p+2}$$

D'où, $1-2\theta\leq 0\Rightarrow\theta\geq\frac{1}{2}$.

Conclusion : si $\theta\geq\frac{1}{2}$ le θ-schéma centré est stable pour tout p (impair ou pair).

1.4 Tableau récapitulatif

	θ-schéma			Implicite ($\theta = 1$)			Explicite ($\theta = 0$)		
	décentré à gauche	centré	décentré à droite	décentré à gauche	centré	décentré à droite	décentré à gauche	centré	décentré à droite
$m = 1$ (transport)	oui si $\theta \geq \frac{1}{2} - \frac{\Delta x}{2\Delta t}$	oui si $\theta \geq \frac{1}{2}$	oui si $\theta \geq \frac{1}{2} + \frac{\Delta x}{2\Delta t}$	oui	oui	oui si $1 \geq \frac{\Delta x}{\Delta t}$	oui si $1 \leq \frac{\Delta x}{\Delta t}$	non	non
$m = 3$ (Airy)	non	oui si $\theta \geq \frac{1}{2}$	oui si $\theta \geq \frac{1}{2} - \frac{(\Delta x)^3}{8\Delta t}$	non	oui	oui	non	non	oui si $\frac{(\Delta x)^3}{\Delta t} \geq 4$
$m = 2p+1$, p pair $p \neq 0$	oui si $\theta \geq \frac{1}{2} - \frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t}$	oui si $\theta \geq \frac{1}{2}$	non	oui	oui	non	oui si $\frac{(\Delta x)^{2p+1}}{\Delta t} \geq 2^{2p}$	non	non
$m = 2p+1$, p impair	non	oui si $\theta \geq \frac{1}{2}$	oui si $\theta \geq \frac{1}{2} - \frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t}$	non	oui	oui	non	non	oui si $\frac{(\Delta x)^{2p+1}}{\Delta t} \geq 2^{2p}$

2 Étude de la consistance pour une équation générale

Afin d'étudier la consistance pour une équation générale du type $\partial_t u(t,x) + \partial_x^{2p+1} u(t,x) = 0$, pour tout $x \in \mathbb{R}$, et $t \in [0,T]$, nous utilisons la notion des différences divisées. En effet, l'erreur de consistance se calcule à partir du schéma numérique dans lequel le vecteur discret u_j^n est remplacé par la solution continue aux points t^n et x_j (*i.e.* $u(t^n,x_j)$). Une somme de termes en $u(t^n,x_k)$ apparaît donc, qu'il nous faut simplifier pour faciliter sa manipulation : c'est ce que permet justement l'utilisation des différences divisées.

2.1 Résultats préliminaires sur les différences divisées

Notation et rappels : La différence divisée d'ordre k de la fonction f en les points $x_0,...,x_k$ sera notée $f[x_0,...,x_k]$. Elle peut se définir par la récurrence suivante :

$$\begin{cases} f[x_j] = f(x_j), \forall j \in \mathbb{Z}, \\ f[x_0,x_1,...,x_k] = \dfrac{f[x_1,...,x_k] - f[x_0,...,x_{k-1}]}{x_k - x_0}, \forall k \geq 1. \end{cases}$$

Les lemmes suivants permettent de simplifier l'écriture du schéma numérique quand on applique ce dernier à la solution continue u.

Lemme 2. *Pour tous $(J,l) \in \mathbb{N}^2$, et $x_l,...,x_{l+J}$, $J+1$ points régulièrement espacés d'une distance Δx, on a*

$$f[x_l,...,x_{l+J}] = \frac{\displaystyle\sum_{k=0}^{J} \binom{J}{k} (-1)^{k+J} f(x_{k+l})}{J!(\Delta x)^J}.$$

Démonstration. Nous procédons par récurrence sur J.

- POUR $J = 0$. Soit $l \in \mathbb{N}$, $f[x_l] = f(x_l)$, par définition des différences divisées. De plus, $\dfrac{\displaystyle\sum_{k=0}^{0} \binom{0}{k} (-1)^k f(x_{k+l})}{0!} = f(x_l)$. La relation est donc bien initalisée.
- SUPPOSONS LA RELATION VRAIE POUR J, ET POUR TOUT $l \in \mathbb{N}$. Par définition des différences divisées,

$$f[x_l,...,x_{l+J+1}] = \frac{f[x_{l+1},...,x_{l+J+1}] - f[x_l,...,x_{l+J}]}{x_{l+J+1} - x_l}.$$

En utilisant l'hypothèse de récurrence, nous obtenons :

$$\begin{aligned} f[x_l,...,x_{l+J+1}] &= \frac{\sum_{k=0}^{J}\binom{J}{k}(-1)^{k+J}f(x_{k+l+1}) - \sum_{k=0}^{J}\binom{J}{k}(-1)^{J+k}f(x_{k+l})}{J!(\Delta x)^J(x_{l+J+1}-x_l)} \\ &= \frac{\sum_{m=1}^{J+1}\binom{J}{m-1}(-1)^{m-1+J}f(x_{m+l}) - \sum_{k=0}^{J}\binom{J}{k}(-1)^{J+k}f(x_{k+l})}{(J+1)!(\Delta x)^{J+1}} \\ &= \frac{\left[\sum_{k=1}^{J}[\binom{J}{k-1}+\binom{J}{k}](-1)^{k+J+1}f(x_{k+l})\right] + f(x_{J+1+l}) - (-1)^J f(x_l)}{(J+1)!(\Delta x)^{J+1}} \\ &= \frac{\sum_{k=0}^{J+1}\binom{J+1}{k}(-1)^{J+1+k}f(x_{k+l})}{(J+1)!(\Delta x)^{J+1}}. \end{aligned}$$

■

Lemme 3. *Soient $j\in\mathbb{Z}$, $p\in\mathbb{N}$, et $x_{j-p-1},...,x_{j+p+1}$ $2p+3$ points régulièrement espacés d'une distance de Δx, alors les trois relations suivantes sont vérifiées.*

$$f[x_{j-p},...,x_{j+p}] = \frac{\sum_{k=0}^{2p}\binom{2p}{k}(-1)^k f(x_{k+j-p})}{(2p)!(\Delta x)^{2p}} = \frac{\sum_{k=0}^{2p}\binom{2p}{k}(-1)^k f(x_{p-k+j})}{(2p)!(\Delta x)^{2p}}, \tag{2.6}$$

$$f[x_{j-p},...,x_{j+p},x_{j+p+1}] = \frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^{k+1} f(x_{k+j-p})}{(2p+1)!(\Delta x)^{2p+1}} = \frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k f(x_{p-k+j+1})}{(2p+1)!(\Delta x)^{2p+1}}, \tag{2.7}$$

$$f[x_{j-p-1},x_{j-p},...,x_{j+p}] = \frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^{k+1} f(x_{k+j-p-1})}{(2p+1)!(\Delta x)^{2p+1}} = \frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k f(x_{p-k+j})}{(2p+1)!(\Delta x)^{2p+1}}. \tag{2.8}$$

Démonstration. Tout se base sur le lemme précédent.

- PREUVE DE LA RELATION (2.6). Il suffit d'appliquer le lemme 2, mais en remplaçant J par $2p$ et l par $j-p$. Le changement de variable $l=2p-k$ dans la somme prouve la deuxième églité de (2.6). En effet, une fois le changement de variable effectué, l'égalité devient :
$$f[x_{j-p},...,x_{j+p}] = \frac{\sum_{l=0}^{2p}\binom{2p}{2p-l}(-1)^{2p-l} f(x_{p-l+j})}{(2p)!(\Delta x)^{2p}}.$$
Nous concluons avec $\binom{2p}{2p-l} = \binom{2p}{l}$.
- PREUVE DE LA RELATION (2.7). Nous appliquons le lemme 2, avec $J=2p+1$ et $l=j-p$. Là encore, pour obtenir la deuxième égalité de la relation (2.7), nous faisons le

changement de variable $l = 2p+1-k$ dans la somme, ce qui nous donne le résultat escompté.

$$f[x_{j-p},...,x_{j+p},x_{j+p+1}] = \frac{\sum_{l=0}^{2p+1} \binom{2p+1}{2p+1-l} (-1)^{2p+2-l} f(x_{p-l+j+1})}{(2p+1)!(\Delta x)^{2p+1}}.$$

- PREUVE DE LA RELATION (2.8). La première égalité s'obtient en remplaçant J par $2p+1$ et l par $j-p-1$ dans le lemme 2. Quant à la deuxième, elle s'obtient grâce au changement de variable $l = 2p+1-k$ dans la somme.

■

2.2 Approximation de la dérivée spatiale

Nous allons appliquer la proposition suivante à la fonction $u(t^n,\cdot)$, où u est la solution de l'équation de départ $\partial_t u(t,x) + \partial_x^{2p+1} u(t,x) = 0, \forall t \in [0,T]$ et $\forall x \in \mathbb{R}$.

Proposition 5. *Soit $x_0,...,x_J$ une suite de $J+1$ points distincts deux à deux. Soit f une fonction de classe $\mathscr{C}^J(\mathbb{R})$ alors, il existe $\xi \in]\min(x_0,...,x_J), \max(x_0,...,x_J)[$ tel que*

$$f[x_0,...,x_J] = \frac{f^{(J)}(\xi)}{J!}.$$

Démonstration. La preuve repose sur l'utilisation répétée du théorème de Rolle, voir par exemple J.P. Demailly [Demailly, 2006]. ■

Notation : Nous notons $u(t^n,[x_0,...,x_k])$ la différence divisée d'ordre k associée à la variable d'espace, *i.e.*

$$\begin{cases} u(t^n,[x_j]) = u(t^n,x_j), \forall j \in \mathbb{Z}, \\ u(t^n,[x_0,...,x_{m+1}]) = \dfrac{u(t^n,[x_1,...,x_{m+1}]) - u(t^n,[x_0,...,x_m])}{x_{m+1}-x_0}, \quad \forall m \geq 1. \end{cases}$$

Nous obtenons donc l'égalité suivante :

Pout tous t^n, p et j, et pourvu que $u \in \mathscr{C}([0,T],\mathscr{C}^{2p+2}(\mathbb{R}))$, il existe $\xi_j^n \in]x_{j-p},x_{j+p+1}[$ tel que

$$\begin{aligned} u(t^n,[x_{j-p},...,x_{j+p+1}]) &= \frac{1}{(2p+1)!}\partial_x^{2p+1} u(t^n,\xi_j^n) \\ &= \frac{1}{(2p+1)!}\partial_x^{2p+1} u(t^n,x_j) + \frac{1}{(2p+1)!}\partial_x^{2p+2} u(t^n,y_j^n)(\xi_j^n - x_j), \end{aligned}$$

pour un certain $y_j^n \in [\min(x_j,\xi_j^n), \max(x_j,\xi_j^n)]$.

Différences finies décentrées à droite :

La partie spatiale du schéma aux différences finies décentrées à droite donne donc :

$$\frac{\sum_{k=0}^{2p+1} \binom{2p+1}{k} (-1)^k u(t^n,x_{p-k+j+1})}{(\Delta x)^{2p+1}} = \partial_x^{2p+1} u(t^n,x_j) + \partial_x^{2p+2} u(t^n,y_j^n)(\xi_j^n - x_j),$$

avec $\xi_j^n \in]x_{j-p}, x_{j+p+1}[$ et $y_j^n \in [\min(x_j,\xi_j^n), \max(x_j,\xi_j^n)]$, et

$$\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^{n+1},x_{p-k+j+1})}{(\Delta x)^{2p+1}} = \partial_x^{2p+1}u(t^{n+1},x_j) + \partial_x^{2p+2}u(t^{n+1},y_j^{n+1})(\xi_j^{n+1}-x_j)$$
$$= \partial_x^{2p+1}u(t^n,x_j) + \partial_t\partial_x^{2p+1}u(\tau_j^n,x_j)\Delta t +$$
$$\partial_x^{2p+2}u(t^{n+1},y_j^{n+1})(\xi_j^{n+1}-x_j),$$

avec $\xi_j^{n+1} \in]x_{j-p}, x_{j+p+1}[$, $y_j^{n+1} \in [\min(x_j,\xi_j^{n+1}), \max(x_j,\xi_j^{n+1})]$, $\tau_j^n \in [t^n, t^{n+1}]$.

Différences finies décentrées à gauche :

Pour le schéma aux différences finies décentré à gauche, nous obtenons :

$$\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^n,x_{p-k+j})}{(\Delta x)^{2p+1}} = \partial_x^{2p+1}u(t^n,x_j) + \partial_x^{2p+2}u(t^n,\bar{y}_j^n)(\bar{\xi}_j^n - x_j),$$

avec $\bar{\xi}_j^n \in]x_{j-p-1}, x_{j+p}[$ et $\bar{y}_j^n \in [\min(x_j,\bar{\xi}_j^n), \max(x_j,\bar{\xi}_j^n)]$, et

$$\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^{n+1},x_{p-k+j})}{(\Delta x)^{2p+1}} = \partial_x^{2p+1}u(t^{n+1},x_j) + \partial_x^{2p+2}u(t^{n+1},\bar{y}_j^{n+1})(\bar{\xi}_j^{n+1} - x_j),$$

avec $\bar{\xi}_j^{n+1} \in]x_{j-p-1}, x_{j+p}[$, $\bar{y}_j^{n+1} \in [\min(x_j,\bar{\xi}_j^{n+1}), \max(x_j,\bar{\xi}_j^{n+1})]$, ce qui redonne lieu au même développement que précédemment.

Différences finies centrées :

Pour le schéma aux différences finies centré, nous avons la relation

$$\frac{u(t^n,x_{p+j+1}) + \left[\sum_{k=0}^{2p}\left[\binom{2p+1}{k} - \binom{2p+1}{k+1}\right](-1)^k u(t^n,x_{p-k+j})\right] - u(t^n,x_{-p-1+j})}{2(\Delta x)^{2p+1}} =$$
$$\frac{1}{2}\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^n,x_{p-k+j+1})}{(\Delta x)^{2p+1}} + \frac{1}{2}\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^n,x_{p-k+j})}{(\Delta x)^{2p+1}}.$$

D'où,

$$\frac{u(t^n,x_{p+j+1}) + \left[\sum_{k=0}^{2p}\left[\binom{2p+1}{k} - \binom{2p+1}{k+1}\right](-1)^k u(t^n,x_{p-k+j})\right] - u(t^n,x_{-p-1+j})}{2(\Delta x)^{2p+1}} =$$
$$\partial_x^{2p+1}u(t^n,x_j) + \frac{1}{2}\partial_x^{2p+2}u(t^n,y_j^n)(\xi_j^n - x_j) + \frac{1}{2}\partial_x^{2p+2}u(t^n,\bar{y}_j^n)(\bar{\xi}_j^n - x_j).$$

Pour t^{n+1}, nous obtenons de même, la demi-somme des développements des schémas décentré à droite et décentré à gauche.

2.3 Calcul de l'erreur de consistance

Dans ce paragraphe, nous avons besoin d'une certaine régularité pour la solution u puisque nous lui appliquons des développements de Taylor successivement. Nous reviendrons sur ces notions de régularité de u au paragraphe sur l'ordre de convergence (section 3).

L'erreur de consistance au temps t^n et en x_j sera notée ε_j^n. Nous remplaçons le vecteur $(u_j^n)_{n\leq \frac{T}{\Delta t}, j\in\mathbb{Z}}$, par la solution continue prise en les points (t^n, x_j) pour obtenir la relation suivante :

θ-schéma aux différences finies décentré à droite :

$$\frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t}+(1-\theta)\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^n,x_{p-k+1+j})}{(\Delta x)^{2p+1}}\right) + \theta\left(\frac{\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k u(t^{n+1},x_{p-k+1+j})}{(\Delta x)^{2p+1}}\right)=\varepsilon_j^n.$$

Soit encore, grâce aux développements de Taylor précédents :

$$\partial_t u(t^n,x_j)+\partial_t^2 u(\mu_j^n,x_j)\frac{\Delta t}{2}+(1-\theta)\left(\partial_x^{2p+1}u(t^n,x_j)+\partial_x^{2p+2}u(t^n,y_j^n)(\xi_j^n-x_j)\right) + \theta\left(\partial_x^{2p+1}u(t^n,x_j)+\partial_t\partial_x^{2p+1}u(\tau_j^n,x_j)\Delta t+\partial_x^{2p+2}u(t^{n+1},y_j^{n+1})(\xi_j^{n+1}-x_j)\right)=\varepsilon_j^n.$$

D'où :

$$\overbrace{\partial_t u(t^n,x_j)+\partial_x^{2p+1}u(t^n,x_j)}^{=0}+\Delta t\left(\frac{\partial_t^2 u(\mu_j^n,x_j)}{2}+\theta\partial_t\partial_x^{2p+1}u(\tau_j^n,x_j)\right)+ (\xi_j^n-x_j)(1-\theta)\partial_x^{2p+2}u(t^n,y_j^n)+(\xi_j^{n+1}-x_j)\theta\partial_x^{2p+2}u(t^{n+1},y_j^{n+1})=\varepsilon_j^n.$$

Or, $|\xi_j^s-x_j|\leq|x_{j+p+1}-x_{j-p}|=(2p+1)\Delta x$, avec $s\in\{n,n+1\}$.

Donc

$$\varepsilon_j^n\leq\Delta t\left(\frac{|\partial_t^2 u(\mu_j^n,x_j)|}{2}+\theta|\partial_t\partial_x^{2p+1}u(\tau_j^n,x_j)|\right)+\Delta x(2p+1)\left((1-\theta)|\partial_x^{2p+2}u(t^n,y_j^n)|+\theta|\partial_x^{2p+2}u(t^{n+1},y_j^{n+1})|\right),$$

à condition que $\partial_t^2 u$, $\partial_t\partial_x^{2p+1}u$ et $\partial_x^{2p+2}u$ existent.

Conclusion : L'erreur de consistance pour le schéma aux différences finies décentré à droite et pour une solution $u\in\mathscr{C}^2([0,T],\mathscr{C}(\mathbb{R}))\cap\mathscr{C}^1([0,T],\mathscr{C}^{2p+1}(\mathbb{R}))\cap\mathscr{C}([0,T],\mathscr{C}^{2p+2}(\mathbb{R}))$ est donc d'ordre 1 en espace et (au moins) 1 en temps. Plus particulièrement, nous avons la majoration :

$$\varepsilon_j^n\leq\Delta t\left(\frac{|\partial_t^2 u(\mu_j^n,x_j)|}{2}+\theta|\partial_t\partial_x^{2p+1}u(\tau_j^n,x_j)|\right)+\Delta x(2p+1)\left((1-\theta)|\partial_x^{2p+2}u(t^n,y_j^n)|+\theta|\partial_x^{2p+2}u(t^{n+1},y_j^{n+1})|\right),$$

avec $\mu_j^n \in [t^n, t^{n+1}]$, $\tau_j^n \in [t^n, t^{n+1}]$, puis y_j^{n+1} et y_j^n tous deux dans l'intervalle $]x_{j-p-1}, x_{j+p+1}[$.

Remarque. Dans le cas particulier du schéma de Crank-Nicolson, $\theta = \frac{1}{2}$, l'erreur de convergence est d'ordre 1 en espace mais 2 en temps. En effet, en écrivant les développements de Taylor à un ordre plus élevé en temps, nous obtenons une majoration de ε_j^n dont le coefficient devant Δt vaut $|\frac{\partial_t^2 u(t^n,x_j)}{2} + \theta \partial_t \partial_x^{2p+1} u(t^n,x_j)|$. Or, ce terme correspond à l'équation de départ dérivée en temps, $\frac{\mathrm{d}}{\mathrm{dt}}(\underbrace{\partial_t u + \partial_x^{2p+1} u}_{=0}) = 0$, ce terme est nul par définition de u.

θ-schéma aux différences finies décentré à gauche :

Comme vu précédemment, les développements de Taylor pour ce schéma sont les mêmes que pour le schéma décentré à droite, nous obtenons donc la même erreur de consistance : ordre 1 en espace et (au moins) 1 en temps.

θ-schéma aux différences finies centré :
Nous savons que l'erreur de consistance pour ce schéma est la demi-somme des erreurs de consistance des schémas précédents. Cependant, si nous faisons la demi-somme des erreurs de consistance obtenues sur les schémas décentré à droite et décentré à gauche, nous obtenons, ici encore, une erreur de consistance d'ordre 1 en espace et (au moins) 1 en temps, ce qui n'est pas optimal.
Nous pouvons en fait montrer que le θ-schéma centré est d'ordre 2 en espace et (au moins) 1 en temps (avec le cas particulier du schéma de Crank-Nicolson, qui est d'ordre 2 en temps). Pour cela, nous faisons la demi-somme des schémas décentré à droite et décentré à gauche :

$$\frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} + \frac{1}{2}(2p+1)!u(t^n,[x_{j-p},...,x_{j+p+1}]) + \frac{1}{2}(2p+1)!u(t^n,[x_{j-p-1},...,x_{j+p}]) = \varepsilon_j^n.$$

En utilisant la définition des différences divisées, nous avons donc

$$\begin{aligned}\varepsilon_j^n &= \frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} + \frac{(2p+1)!}{2}\Big(\frac{u(t^n,[x_{j-p+1},...,x_{j+p+1}]) - u(t^n,[x_{j-p},...,x_{j+p}])}{x_{j+p+1}-x_{j-p}} \\ &\qquad + \frac{u(t^n,[x_{j-p},...,x_{j+p}]) - u(t^n,[x_{j-p-1},...,x_{j+p-1}])}{x_{j+p}-x_{j-p-1}}\Big) \\ &= \frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} + \frac{(2p)!}{2}(u(t^n,[x_{j-p+1},...,x_{j+p+1}]) - u(t^n,[x_{j-p-1},...,x_{j+p-1}]))\end{aligned}$$

En utilisant la proposition 5, nous transformons $u(t^n,[x_{j-p+1},...,x_{j+p+1}]) - u(t^n,[x_{j-p-1},...,x_{j+p-1}])$ en soustraction de deux dérivées en espace. Nous obtenons un ordre 2 en espace en utilisant les développements de Taylor.

Dans la suite du mémoire, nous noterons d'ordre (au moins) 1 en temps et (au moins) 1 en espace pour englober les différents cas résumés ici :

	décentré à gauche	centré	décentré à droite
$\theta \neq \frac{1}{2}$	1 en temps 1 en espace	1 en temps 2 en espace	1 en temps 1 en espace
$\theta = \frac{1}{2}$ Crank-Nicolson	2 en temps 1 en espace	2 en temps 2 en espace	2 en temps 1 en espace

TABLE 2.1: Résumé des différents ordres pour les erreurs de consistance

3 Ordre de convergence

3.1 Équation générale : $\partial_t u + \partial_x^{2p+1} u = 0$ pour une donnée initiale $u^0 \in H^{4p+3}(\mathbb{R})$

Tout d'abord, la régularité de u_0 se transmet à la solution u pour tout temps, puisque la norme $H^k(\mathbb{R})$ se conserve comme vu précédemment au chapitre 1. Donc $\forall t \in [0,T]$, $u(t,\cdot) \in H^{4p+3}(\mathbb{R})$.
Dans un premier temps, la régularité $H^{4p+3}(\mathbb{R})$ est utile pour pouvoir utiliser le calcul de consistance fait à la section précédente. Comme $H^s(\mathbb{R}) \subset \mathscr{C}^k(\mathbb{R})$ pour $s > \frac{1}{2} + k$, ces développements sont corrects ($u \in \mathscr{C}^{4p+2}(\mathbb{R})$). Nous verrons dans un second paragraphe, comment s'affranchir de cette contrainte de régularité.

Afin de pouvoir travailler avec n'importe quel schéma (différences finies décentrées à droite, centrées ou décentrées à gauche), nous ne précisons pas l'ordre de consistance optimal, et écrivons (au moins) d'ordre 1 en temps et (au moins) d'ordre 1 en espace. Nous supposons également que la condition de CFL (si elle existe) est satisfaite.
Nous introduisons l'erreur globale $e_j^n = u_j^n - u(t^n, x_j)$. D'après la définition de l'erreur de consistance ε_j^n et les équations du schéma numérique, l'erreur globale vérifie la relation :

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} + \theta \left(\frac{\sum_{k=0}^{2p+1} \binom{2p+1}{k} (-1)^k e_{p-k+j+1}^{n+1}}{(\Delta x)^{2p+1}} \right) + (1-\theta) \left(\frac{\sum_{k=0}^{2p+1} \binom{2p+1}{k} (-1)^k e_{p-k+j+1}^{n}}{(\Delta x)^{2p+1}} \right) = -\varepsilon_j^n.$$

Nous avons ici explicité cette relation dans le cas des différences finies décentrées à droite, une relation similaire existe pour les deux autres schémas, il suffit d'appliquer ces schémas à e_j^n.

Le θ-schéma aux différences finies décentré à droite
En utilisant ensuite la définition de $\widehat{U^n}$ rappelée dans le début du chapitre 2, nous aboutissons à :

$$\widehat{e^{n+1}}(\xi) = A_1(\xi)\widehat{e^n}(\xi) - \frac{\Delta t}{1 + \theta \frac{\Delta t}{(\Delta x)^{2p+1}} e^{-i\pi\xi} (-2i\sin(\pi\xi))^{2p+1}} \widehat{\varepsilon^n}(\xi).$$

En utilisant la stabilité l^2 et en supposant que la condition de CFL est satisfaite, nous obtenons :

$$\begin{aligned}
||e^{n+1}||_{l^2} &= \sqrt{\Delta x} ||\widehat{e^{n+1}}||_{L^2([0,1])} \\
&\leq \sqrt{\Delta x} ||A_1(\xi)\widehat{e^n}(\xi)||_{L^2([0,1])} + \sqrt{\Delta x} || \frac{\Delta t}{1 + \theta \frac{\Delta t}{(\Delta x)^{2p+1}} e^{-i\pi\xi} (-2i\sin(\pi\xi))^{2p+1}} \widehat{\varepsilon^n} ||_{L^2([0,1])} \\
&\leq \sqrt{\Delta x} ||\widehat{e^n}(\xi)||_{L^2([0,1])} + \sqrt{\Delta x}(\Delta t) || \frac{1}{1 + \theta \frac{\Delta t}{(\Delta x)^{2p+1}} e^{-i\pi\xi} (-2i\sin(\pi\xi))^{2p+1}} ||_{L^\infty([0,1])} ||\widehat{\varepsilon^n}||_{L^2([0,1])}
\end{aligned}$$

$|| \frac{1}{1+\theta \frac{\Delta t}{(\Delta x)^{2p+1}} e^{-i\pi\xi} (-2i\sin(\pi\xi))^{2p+1}} ||_{L^\infty([0,1])}$ est bien plus petit que 1 car nous nous sommes placés dans le cas des schémas aux différences finies décentré à droite stables (avec une contrainte de type CFL à vérifier), *i.e.* nous étudions les cas p impair, ou $p = 0$.
Dans le cas $p = 0$, qui correspond au cas de l'équation de transport discrétisée avec les différences finies décentrées à droite, il est nécessaire de restreindre un peu la condition de CFL de stabilité et de prendre $\theta \geq \max(\frac{\Delta x}{\Delta t}, \frac{1}{2} + \frac{\Delta x}{2\Delta t})$, pour avoir la stabilité l^2 et la norme $|| \frac{1}{1+\theta \frac{\Delta t}{(\Delta x)^{2p+1}} e^{-i\pi\xi} (-2i\sin(\pi\xi))^{2p+1}} ||_{L^\infty([0,1])}$ inférieure à 1.

Le θ-schéma aux différences finies décentré à gauche
Le calcul est identique en remplaçant A_1 par A_2, $e^{-i\pi\xi}$ par $e^{i\pi\xi}$ et en ne considérant que les cas p pair puisque nous supposons notre schéma stable.

Le θ-schéma aux différences finies centré

Dans le cas du θ-schéma aux différences finies centré, $e^{-i\pi\xi}$ est remplacé par $\cos(\pi\xi)$, et le coefficient d'amplification A_1 par A_3, mais les conclusions restent identiques.

Soit encore pour n'improte quel schéma (différences finies décentrées à gauche, centrées ou décentrées à droite) :

$$\begin{aligned} ||e^{n+1}||_{l^2} &\leq ||e^n||_{l^2} + \Delta t||\varepsilon^n||_{l^2} \\ &\leq \ \vdots \\ &\leq ||e^0||_{l^2} + \sum_{k=0}^{n} \Delta t||\varepsilon^k||_{l^2}. \end{aligned}$$

Or, $u(0,x_j) = u_0(x_j) = u_j^0$, donc $e_j^0 = 0, \forall j \in \mathbb{Z}$.
D'où

$$\sup_{n\in[\![0,N]\!]} ||e^n||_{l^2} \leq \sup_{n\in[\![0,N]\!]} \sum_{k=0}^{n-1} \Delta t||\varepsilon^k||_{l^2} = T \sup_{k\in[\![0,N]\!]} ||\varepsilon^k||_{l^2}. \tag{2.9}$$

Il nous faut donc déterminer $||\varepsilon^k||_{l^2}, \forall k \in [\![0,N]\!]$.

D'après l'expression de l'erreur de consistance ε_j^n, nous avons :

$$\begin{aligned} ||\varepsilon^n||_{l^2}^2 \leq \sum_{j\in\mathbb{Z}} (\Delta x) \Big\{ &(\Delta t)^2 ||\partial_t^2 u(\mu_j^n,\cdot)||_{L^\infty[x_j,x_{j+1}]}^2 + 4\theta^2(\Delta t)^2 ||\partial_x^{2p+1}\partial_t u(\tau_j^n,\cdot)||_{L^\infty[x_j,x_{j+1}]}^2 \\ &+4(2p+1)^2(\Delta x)^2(1-\theta)^2 ||\partial_x^{2p+2}u(t^n,\cdot)||_{L^\infty([x_{j-p-1},x_{j+p+1}])}^2 + 4(2p+1)^2(\Delta x)^2\theta^2 ||\partial_x^{2p+2}u(t^{n+1},\cdot)||_{L^\infty([x_{j-p-1},x_{j+p+1}])}^2 \Big\}. \end{aligned}$$

Nous aurons ensuite besoin de la majoration 2.11 (que nous démontrons grâce au lemme suivant). Nous appliquerons (2.11) successivement à $\partial_x^{2p+1}\partial_t u$, $\partial_x^{2p+2}u$ ou encore $\partial_t^2 u$.

Lemme 4. *Soit $v \in H^1(\mathbb{R})$, alors*

$$||v||_{L^\infty(\mathbb{R})} \leq \sqrt{2}||v||_{L^2(\mathbb{R})}^{\frac{1}{2}} ||v'||_{L^2(\mathbb{R})}^{\frac{1}{2}}. \tag{2.10}$$

Démonstration. Il s'agit d'un cas particulier de l'inégalité de Gagliardo-Nirenberg (c.f [Brezis, 2005]). ■

L'inégalité (2.10) n'est valable que sur $\mathbb{R}$ entier, or nous aimerions majorer la norme $L^\infty([x_j,x_{j+1}])$. Nous allons démontrer le lemme suivant :

Lemme 5. *Soit $v \in H^1(\mathbb{R})$, et j un entier quelconque dans $\mathbb{Z}$, alors il existe une constante C_1 indépendante de j telle que*

$$||v||_{L^\infty([x_j,x_{j+1}])} \leq \frac{C_1}{(\Delta x)^{\frac{1}{2}}} ||v||_{H^1([x_{j-1},x_{j+2}])}, \quad \forall j \in \mathbb{Z}. \tag{2.11}$$

Démonstration. Afin de restreindre le domaine d'étude de $\mathbb{R}$ à $[x_j,x_{j+1}]$, nous introduisons une fonction cutt-off χ.

Définition 2. *Dans toute la suite, χ sera une fonction cutt-off définie comme suit :*

- $\chi \in \mathscr{C}^\infty(\mathbb{R})$,
- $0 \leq \chi \leq 1$,
- $\chi \equiv 1$ *sur* $[0,1]$,
- $\chi \equiv 0$ *sur* $[-1,2]^c$.

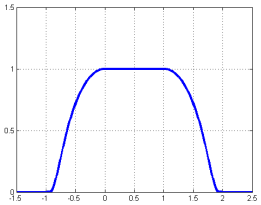


FIGURE 2.1: Fonction cutt-off χ

À partir de cette fonction χ, nous construisons χ_j définie par $\chi_j(x) = \chi(\frac{x-x_j}{\Delta x})$. χ_j a donc la même régularité que celle de χ mais vérifie $\chi_j \equiv 1$ sur $[x_j, x_{j+1}]$, et $\chi_j \equiv 0$ sur $[x_{j-1}, x_{j+2}]^c$.
Nous étudions $\chi_j v$. La définition de χ_j permet d'établir l'égalité (a) et l'utilisation de l'inégalité (2.10) prouve l'inégalité (b).

$$||v||_{L^\infty([x_j,x_{j+1}])} \underset{(a)}{=} ||\chi_j v||_{L^\infty([x_j,x_{j+1}])} \leq ||\chi_j v||_{L^\infty(\mathbb{R})} \underset{(b)}{\leq} \sqrt{2}||\chi_j v||^{\frac{1}{2}}_{L^2(\mathbb{R})}||(\chi_j v)'||^{\frac{1}{2}}_{L^2(\mathbb{R})}$$

Or, $||\chi_j v||^{\frac{1}{2}}_{L^2(\mathbb{R})} = ||\chi_j v||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])} \leq ||\chi_j||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}||v||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])} = ||\chi||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}||v||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])}$.
D'autre part, en remarquant que $\chi_j' = \frac{1}{\Delta x}\chi'(\frac{\cdot - x_j}{\Delta x})$, nous avons :

$$\begin{aligned}
||(\chi_j v)'||^{\frac{1}{2}}_{L^2(\mathbb{R})} &\leq \left(||\chi_j v'||_{L^2(\mathbb{R})} + ||\chi_j' v||_{L^2(\mathbb{R})}\right)^{\frac{1}{2}} \\
&= \left(||\chi_j v'||_{L^2([x_{j-1},x_{j+2}])} + ||\frac{1}{\Delta x}\chi'(\frac{\cdot - x_j}{\Delta x})v||_{L^2([x_{j-1},x_{j+2}])}\right)^{\frac{1}{2}} \\
&\leq \left(||\chi_j||_{L^\infty(\mathbb{R})}||v'||_{L^2([x_{j-1},x_{j+2}])} + \frac{1}{\Delta x}||\chi'||_{L^\infty(\mathbb{R})}||v||_{L^2([x_{j-1},x_{j+2}])}\right)^{\frac{1}{2}}.
\end{aligned}$$

Puis, nous utilisons le fait que $||\chi_j||_{L^\infty(\mathbb{R})} = ||\chi||_{L^\infty(\mathbb{R})} = 1$, ainsi que l'identité $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, (identité vraie dès que a et b sont positifs) :

$$\begin{aligned}
\sqrt{2}||\chi_j v||^{\frac{1}{2}}_{L^2(\mathbb{R})}||(\chi_j v)'||^{\frac{1}{2}}_{L^2(\mathbb{R})} &\leq \sqrt{2}||v||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])}\left(\frac{1}{(\Delta x)^{\frac{1}{2}}}||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}||v||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])} + ||v'||^{\frac{1}{2}}_{L^2([x_{j-1},x_{j+2}])}\right) \\
&\leq \sqrt{2}||v||_{L^2([x_{j-1},x_{j+2}])}\frac{||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}}{(\Delta x)^{\frac{1}{2}}} + \frac{\sqrt{2}}{2}\left(||v||_{L^2([x_{j-1},x_{j+2}])} + ||v'||_{L^2([x_{j-1},x_{j+2}])}\right) \\
&\leq \max\left(\sqrt{2}\frac{||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}}{(\Delta x)^{\frac{1}{2}}} + \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)\underbrace{\left(||v||_{L^2([x_{j-1},x_{j+2}])} + ||v'||_{L^2([x_{j-1},x_{j+2}])}\right)}_{\leq\sqrt{2}||v||_{H^1([x_{j-1},x_{j+2}])}}.
\end{aligned}$$

Puisque Δx sera par la suite amené à converger vers 0, nous pouvons considérer $\Delta x \leq 1$. Donc $\max\left(\sqrt{2}\frac{||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}}{(\Delta x)^{\frac{1}{2}}} + \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \frac{2||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})} + (\Delta x)^{\frac{1}{2}}}{\sqrt{2}(\Delta x)^{\frac{1}{2}}}$.
Soit en combinant les inégalités précédentes :

$$||v||_{L^\infty([x_j,x_{j+1}])} \leq \frac{2||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})} + (\Delta x)^{\frac{1}{2}}}{(\Delta x)^{\frac{1}{2}}}||v||_{H^1([x_{j-1},x_{j+2}])}.$$

Il ne nous reste plus qu'à poser

$$C_1 = 2||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})} + 1.$$

C_1 est bien indépendante de j, et l'inégalité (2.11) est bien vérifiée.

■

Remarque. *En suivant la même preuve, nous montrons que*

$$||v||_{L^\infty([x_{j-p-1},x_{j+p+1}])} \leq \frac{C_2}{(\Delta x)^{\frac{1}{2}}}||v||_{H^1([x_{j-3p-3},x_{j+3p+3}])},$$

avec

$$C_2 = \frac{\sqrt{2}||\chi'||^{\frac{1}{2}}_{L^\infty(\mathbb{R})}}{\sqrt{p+1}} + 1,$$

et ce, pour tous j dans $\mathbb{Z}$. C_2 est bien une constante indépendante de j.

En revenant à la majoration de l'erreur de consistance ε^n_j, nous obtenons :

$$\begin{aligned}||\varepsilon^n||^2_{l^2} \leq \sum_{j\in\mathbb{Z}}(\Delta x)\Bigg\{&(\Delta t)^2\frac{C_1^2}{\Delta x}||\partial_t^2 u(\mu_j^n,\cdot)||^2_{H^1([x_{j-1},x_{j+2}])} + 4\theta^2(\Delta t)^2\frac{C_1^2}{\Delta x}||\partial_x^{2p+1}\partial_t u(\tau_j^n,\cdot)||^2_{H^1([x_{j-1},x_{j+2}])}\\ &+4(2p+1)^2(\Delta x)^2(1-\theta)^2\frac{C_2^2}{\Delta x}||\partial_x^{2p+2}u(t^n,\cdot)||^2_{H^1([x_{j-3p-3},x_{j+3p+3}])}\\ &+4(2p+1)^2(\Delta x)^2\theta^2\frac{C_2^2}{\Delta x}||\partial_x^{2p+2}u(t^{n+1},\cdot)||^2_{H^1([x_{j-3p-3},x_{j+3p+3}])}\Bigg\}.\end{aligned}$$

Soit encore :

$$\begin{aligned}&||\varepsilon^n||^2_{l^2} \leq 3(\Delta t)^2C_1^2||\partial_t^2u(\mu_j^n,\cdot)||^2_{H^1(\mathbb{R})} + 12\theta^2(\Delta t)^2C_1^2||\partial_x^{2p+1}\partial_t u(\tau_j^n,\cdot)||^2_{H^1(\mathbb{R})}\\ &+4(2p+1)^2(\Delta x)^2(1-\theta)^2C_2^2(6p+6)||\partial_x^{2p+2}u(t^n,\cdot)||^2_{H^1(\mathbb{R})} + 4(2p+1)^2(\Delta x)^2\theta^2C_2^2(6p+6)||\partial_x^{2p+2}u(t^{n+1},\cdot)||^2_{H^1(\mathbb{R})}.\end{aligned}$$

Nous obtenons donc la majoration suivante pour l'erreur globale (en remarquant que $\partial_t\partial_x^{2p+1}u = -\partial_x^{4p+2}u$ ainsi que $\partial_t^2u = \partial_x^{4p+2}u$ et que $T = N\Delta t$) :

$$\begin{aligned}&\sup_{n\in[0,N]}||e^n||_{l^2}\\ &\leq T\Big(3(\Delta t)^2C_1^2||\partial_x^{4p+2}u_0||^2_{H^1(\mathbb{R})} + 12\theta^2(\Delta t)^2C_1^2||\partial_x^{4p+2}u_0||^2_{H^1(\mathbb{R})}\\ &+24(p+1)(2p+1)^2C_2^2(1-2\theta+2\theta^2)(\Delta x)^2||\partial_x^{2p+2}u_0||^2_{H^1(\mathbb{R})}\Big)^{\frac{1}{2}}\\ &\leq \underbrace{T\max\left(\sqrt{3+12\theta^2}C_1||\partial_x^{4p+2}u_0||_{H^1(\mathbb{R})}, \sqrt{24(p+1)(1-2\theta+2\theta^2)}(2p+1)C_2||\partial_x^{2p+2}u_0||_{H^1(\mathbb{R})}\right)}_{=C_3(||\partial_x^{4p+2}u_0||_{H^1(\mathbb{R})},||\partial_x^{2p+2}u_0||_{H^1(\mathbb{R})})}(\Delta t+\Delta x).\end{aligned}$$

Donc à condition que $u_0 \in H^{4p+3}(\mathbb{R})$, nous avons une convergence d'ordre (au moins) 1 en temps et (au moins) 1 en espace.

Remarque. *Au lieu d'utiliser le lemme 5, nous aurions pu utiliser l'injection continue de Sobolev $H^s(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R})$ dès que $s > \frac{1}{2}$ (i.e. il existe une constante C indépendante de u telle que $||u||_{L^\infty(\mathbb{R})} \leq C||u||_{H^s(\mathbb{R})}$), (pour une démonstration, voir le livre de C.Zuily [Zuily, 2002]).*

Par contre, l'ordre de convergence obtenu avec cette injection est moins bon qu'avec le lemme 5 : nous obtenons $\sup_{n\in[0,N]}||e^n||_{l^2} \leq TC_4(\frac{\Delta t}{\sqrt{\Delta x}} + \Delta x)$, où C_4 est une constante ne dépendant que de p, de θ, de $||\partial_x^{2p+2}u_0||_{H^1(\mathbb{R})}$ et de $||\partial_x^{4p+2}u_0||_{H^1(\mathbb{R})}$.

3.2 Équation générale pour une donnée initiale moins régulière $u^0 \in H^2(\mathbb{R})$

Si u_0 est moins régulière, le calcul fait précédemment pour déterminer l'erreur de consistance n'est plus correct : nous ne pouvons effectuer ces développements de Taylor. L'astuce est ici d'introduire une suite régularisante $(\varphi^\varepsilon)_\varepsilon$ que l'on convoluera en espace avec u. Comme dans le calcul d'ordre précédent, nous étudions ici indifféremment les trois schémas aux différences finies, pourvu que la condition de CFL (si elle existe) soit vérifiée.

Définition 3. *Soit* φ *une fonction telle que*
- $\varphi \in \mathscr{C}^\infty(\mathbb{R})$, *à support compact dans* $[-1,1]$,
- $\int_{\mathbb{R}} \varphi(x)dx = 1$,
- $\varphi \geq 0$,

Nous définissons une suite régularisante en posant $\varphi^\varepsilon(\cdot) = \frac{1}{\varepsilon}\varphi(\frac{\cdot}{\varepsilon})$.

Un exemple d'une telle fonction φ est $\frac{e^{\frac{-1}{(1-||\cdot||^2)}}\mathbb{1}_{\{||\cdot||<1\}}}{\int_{\{||x||<1\}} e^{\frac{-1}{(1-||x||^2)}}dx}$.

Notation : Nous notons
- u la solution de l'équation générale $\partial_t u + \partial_x^{2p+1} u = 0$, avec donnée initiale u_0,
- u^ε la solution de l'équation générale, avec donnée initiale $u_0 \star \varphi^\varepsilon$, où $\star$ représente le produit de convolution,
- $u_\Delta = (u_j^n)_{n\leq N, j\in\mathbb{Z}}$ le vecteur discret fourni par le schéma numérique, avec donnée initiale $u_j^0 = u_0(x_j), \forall j \in \mathbb{Z}$,
- $u_\Delta^\varepsilon = ((u^\varepsilon)_j^n)_{n\leq N, j\in\mathbb{Z}}$ le vecteur discret fourni par le schéma numérique, avec donnée initiale $(u^\varepsilon)_j^0 = (u_0 \star \varphi^\varepsilon)(x_j), \forall j \in \mathbb{Z}$.

Nous notons toujours $e_j^n = u(t^n, x_j) - u_j^n$ l'erreur globale.

$$\begin{aligned}||e^n||_{l^2} &= \sqrt{\sum_{k\in\mathbb{Z}} |u(t^n,x_k) - u_k^n|^2 \Delta x} \\ &\leq \underbrace{\sqrt{\sum_{k\in\mathbb{Z}} |u(t^n,x_k) - u^\varepsilon(t^n,x_k)|^2 \Delta x}}_{[a]} + \underbrace{\sqrt{\sum_{k\in\mathbb{Z}} |u^\varepsilon(t^n,x_k) - (u^\varepsilon)_k^n)|^2 \Delta x}}_{[b]} + \underbrace{\sqrt{\sum_{k\in\mathbb{Z}} |(u^\varepsilon)_k^n - u_k^n|^2 \Delta x}}_{[c]}.\end{aligned}$$

Il nous reste donc à déterminer une majoration de $[a], [b]$, et $[c]$.
POUR $[a]$:

$$[a]^2 \leq \sum_{k\in\mathbb{Z}} \Delta x ||u(t^n,\cdot) - u^\varepsilon(t^n,\cdot)||^2_{L^\infty([x_k,x_{k+1}])}$$

ce qui devient en utilisant l'inégalité (2.11) :

$$\begin{aligned}[a]^2 &\leq \sum_{k\in\mathbb{Z}} \Delta x \frac{C_1^2}{\Delta x} ||u(t^n,\cdot) - u^\varepsilon(t^n,\cdot)||^2_{H^1([x_{k-1},x_{k+2}])} \\ &\leq 3C_1^2 ||u(t^n,\cdot) - u^\varepsilon(t^n,\cdot)||^2_{H^1(\mathbb{R})}.\end{aligned}$$

Par conservation de la norme $H^1(\mathbb{R})$, nous aboutissons donc à

$$[a] \leq \sqrt{3}C_1 ||u_0 - u_0 \star \varphi^\varepsilon||_{H^1(\mathbb{R})}.$$

Lemme 6. *La différence entre* u_0 *et* $u_0 \star \varphi^\varepsilon$ *en norme* $H^1(\mathbb{R})$ *est de l'ordre de* ε.
Plus précisemment,

$$||u_0 - u_0 \star \varphi^\varepsilon||_{H^1(\mathbb{R})} \leq \sqrt{\frac{2}{3}}\varepsilon ||(u_0)'||_{H^1(\mathbb{R})} ||\varphi||_{L^2(\mathbb{R})}. \tag{2.12}$$

Démonstration. Par définition de $\star$ et de φ^ε, nous avons

$$\int_{\mathbb{R}} |u_0(x) - (u_0 \star \varphi^\varepsilon)(x)|^2 dx = \int_{\mathbb{R}} |\int_{-\varepsilon}^{\varepsilon} [u_0(x) - u_0(x-y)]\varphi^\varepsilon(y)dy|^2 dx.$$

En utilisant l'inégalité de Cauchy-Schwarz, nous obtenons :

$$\begin{aligned}\int_{\mathbb{R}} |u_0(x) - (u_0 \star \varphi^\varepsilon)(x)|^2 dx &\leq \int_{\mathbb{R}} \int_{-\varepsilon}^{\varepsilon} [u_0(x) - u_0(x-y)]^2 dy \int_{-\varepsilon}^{\varepsilon} (\varphi^\varepsilon(y))^2 dy dx \\ &= \left(\int_{-\varepsilon}^{\varepsilon} \int_{\mathbb{R}} |u_0(x) - u_0(x-y)|^2 dx dy\right) \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi^2(z) dz.\end{aligned}$$

La deuxième égalité se justifie d'après le théorème de Tonelli et le fait que φ soit nulle en dehors de $[-1,1]$.
Il nous faut donc calculer $\int_{\mathbb{R}} |u_0(x)-u_0(x-y)|^2 dx$.
Puisque $u_0 \in H^1(\mathbb{R})$, nous avons grâce au théorème de Cauchy-Schwarz :

$$\begin{aligned}\int_{\mathbb{R}} |u_0(x)-u_0(x-y)|^2 dx &\leq \int_{\mathbb{R}} \left(\int_{x-y}^{x} (u_0)'(z)dz\right)^2 dx \\ &\leq \int_{x\in\mathbb{R}} y \int_{x-y}^{x} [(u_0)'(z)]^2 dz dx \\ &\leq y \int_{z\in\mathbb{R}} [(u_0)'(z)]^2 \underbrace{\int_{z}^{z+y} dx}_{=y} dz.\end{aligned}$$

Là encore, le théorème de Tonelli permet de justifier la dernière inégalité. D'où

$$\int_{-\varepsilon}^{\varepsilon} ||u_0 - u_0(\cdot - y)||^2_{L^2(\mathbb{R})} dy \leq \frac{2}{3}\varepsilon^3 ||(u_0)'||^2_{L^2(\mathbb{R})},$$

ce qui implique que :

$$\int_{\mathbb{R}} |u_0(x) - (u_0 \star \varphi^\varepsilon)(x)|^2 dx \leq \frac{2}{3}\varepsilon^2 ||(u_0)'||^2_{L^2(\mathbb{R})} ||\varphi||^2_{L^2(\mathbb{R})}.$$

De même, en suivant la même démarche, nous prouvons que :

$$\int_{\mathbb{R}} |(u_0)'(x) - (u_0 \star \varphi^\varepsilon)'(x)|^2 dx \leq \frac{2}{3}\varepsilon^2 ||(u_0)''||^2_{L^2(\mathbb{R})} ||\varphi||^2_{L^2(\mathbb{R})}.$$

Ceci termine la démonstration du lemme 6. ■

POUR $[b]$: Puisque $u_0 \star \varphi^\varepsilon$ a la même régularité que φ^ε, nous pouvons appliquer à u^ε, le calcul fait au paragraphe précédent, lorsque la donnée initiale est suffisamment régulière. Il existe donc une constante C_3 ne dépendant que de p, de θ, de $||\partial_x^{4p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})}$ et de $||\partial_x^{2p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})}$ telle que

$$[b] \leq T C_3(||\partial_x^{4p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})}, ||\partial_x^{2p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})})(\Delta t + \Delta x).$$

Or

$$||\partial_x^{4p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})} \leq ||u_0||_{L^2(\mathbb{R})} \left(||\,(\varphi^\varepsilon)^{(4p+2)}\,||_{L^2(\mathbb{R})} + ||\,(\varphi^\varepsilon)^{(4p+3)}\,||_{L^2(\mathbb{R})} \right),$$

ce qui par définition de φ^ε, fournit la majoration suivante :

$$||\partial_x^{4p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})} \leq ||u_0||_{L^2(\mathbb{R})} \left(\frac{1}{\varepsilon^{4p+\frac{5}{2}}} ||\varphi^{(4p+2)}||_{L^2(\mathbb{R})} + \frac{1}{\varepsilon^{4p+\frac{7}{2}}} ||\varphi^{(4p+3)}||_{L^2(\mathbb{R})} \right).$$

De même, nous avons

$$||\partial_x^{2p+2}(u_0 \star \varphi^\varepsilon)||_{H^1(\mathbb{R})} \leq ||u_0||_{L^2(\mathbb{R})} \left(\frac{1}{\varepsilon^{2p+\frac{5}{2}}} ||\varphi^{(2p+2)}||_{L^2(\mathbb{R})} + \frac{1}{\varepsilon^{2p+\frac{7}{2}}} ||\varphi^{(2p+3)}||_{L^2(\mathbb{R})} \right).$$

Les membres de droite de ces deux inégalités sont bien finis car $\varphi \in \mathscr{C}_c^\infty([-1,1])$. Puisque $\varepsilon \leq 1$, nous obtenons finalement pour le terme $[b]$:

$$[b] \leq T \frac{C_5}{\varepsilon^{4p+\frac{7}{2}}} ||u_0||_{L^2(\mathbb{R})} (\Delta t + \Delta x),$$

avec $C_5 = \max\left(\sqrt{3+12\theta^2} C_1 \left(||\varphi^{(4p+2)}||_{L^2(\mathbb{R})} + ||\varphi^{(4p+3)}||_{L^2(\mathbb{R})}\right), \sqrt{24(p+1)(1-2\theta+2\theta^2)}(2p+1) C_2 \left(||\varphi^{(2p+2)}||_{L^2(\mathbb{R})} + ||\varphi^{(2p+3)}||_{L^2(\mathbb{R})}\right)\right)$.
POUR $[c]$ En utilisant la stabilité l^2 du schéma, nous obtenons pour $s \in \{1,2,3\}$:

$$\begin{aligned}[c] &\leq ||A_s(\xi)||_{L^\infty(\mathbb{R})} ||(u^\varepsilon)^0 - u^0||_{l^2} \\ &\leq ||(u^\varepsilon)^0 - u^0||_{l^2} \\ &\leq \sqrt{\sum_{k\in\mathbb{Z}} (\Delta x) ||(u_0 \star \varphi^\varepsilon) - u_0||^2_{L^\infty([x_k, x_{k+1}])}},\end{aligned}$$

car $u_j^0 = u_0(x_j)$ et $(u^\varepsilon)_j^0 = (u_0 \star \varphi^\varepsilon)(x_j)$, $\forall j \in \mathbb{Z}$.
Nous concluons grâce aux lemmes 5 et 6 :

$$[c] \leq \varepsilon\sqrt{2}C_1||(u_0)'||_{H^1(\mathbb{R})}||\varphi||_{L^2(\mathbb{R})}.$$

Conclusion Nous avons donc la majoration suivante pour l'erreur globale :

$$\sup_{n\in[0,N]}||e^n||_{l^2} \leq \varepsilon 2\sqrt{2}C_1||(u_0)'||_{H^1(\mathbb{R})}||\varphi||_{L^2(\mathbb{R})} + T\frac{C_5}{\varepsilon^{4p+\frac{7}{2}}}||u_0||_{L^2(\mathbb{R})}(\Delta t + \Delta x).$$

Ce majorant est minimal quand ε est de l'ordre de $(\Delta t + \Delta x)^{\frac{1}{4p+\frac{9}{2}}}$. Le schéma est donc d'ordre (au moins) $\frac{2}{8p+9}$ en temps et en espace.

Remarque. *À cause de la majoration* (2.12)*, nous avons besoin d'imposer $u_0 \in H^2(\mathbb{R})$. Cependant ceci n'est pas optimal, nous n'avons en réalité pas besoin d'une aussi grande régularité pour u_0.*
Tout d'abord, nous pouvons optimiser le lemme 6. En effet, nous avons besoin de connaître de taux de convergence de $||u_0 - u_0 \star \varphi^\varepsilon||_{H^1(\mathbb{R})}$. L'inégalité démontrée dans ce lemme 6 revient à majorer$||u_0 - u_0 \star \varphi^\varepsilon||_{H^1(\mathbb{R})}$ par $\bar{C}\varepsilon||u_0||_{H^2(\mathbb{R})}$, où $\bar{C}$ est une constante indépendante de ε (mais qui dépend de φ). En réalité, ce n'est qu'un cas particulier d'une inégalité plus générale :

Proposition 6. *Il existe une constante $\bar{\bar{C}}$ ne dépendant que de φ et non de ε telle que, pour tout $s > 1$, pour tout $u_0 \in H^s(\mathbb{R})$:*

$$||u_0 - u_0 \star \varphi^\varepsilon||_{H^1(\mathbb{R})} \leq \bar{\bar{C}}\varepsilon^{s-1}||u_0||_{H^s(\mathbb{R})}.$$

Démonstration. Soit s fixé plus grand que 1. Nous introduisons une fonction χ_s définie comme suit :

- $\chi_s \in \mathscr{C}^\infty(\mathbb{R})$,
- $0 \leq \chi_s \leq 1$,
- $\chi_s(0) = 1$,
- $\sup_{\lambda\in\mathbb{R}}(\frac{\chi_s(\lambda)-1}{\lambda^s}) < 1$, (ce qui est réalisé par exemple si χ_s est très « plate » autour de 0).

Nous pouvons poser $\varphi = \mathcal{F}^{-1}(\chi_s)$, où $\mathcal{F}$ représente la transformée de Fourier. Dans ce cas, $\varphi^\varepsilon = \mathcal{F}^{-1}(\chi_s(\varepsilon\cdot))$, et on a la relation $\mathcal{F}(u_0 \star \varphi^\varepsilon)(\xi) = \chi_s(\varepsilon\xi)\mathcal{F}(u_0)(\xi)$, $\forall \xi \in \mathbb{R}$.

Dans ce cas,

$$\begin{aligned}||u_0 - u_0 \star \varphi^\varepsilon||^2_{H^1(\mathbb{R})} &= \int_{\mathbb{R}} |\xi|^2(\chi_s(\varepsilon\xi) - 1)^2|\mathcal{F}(u_0)(\xi)|^2 d\xi, \\ &= \varepsilon^{2s}\int_{\mathbb{R}} |\xi|^{2(s+1)}(\frac{\chi_s(\varepsilon\xi)-1}{(\xi\varepsilon)^s})^2|\mathcal{F}(u_0)(\xi)|^2 d\xi.\end{aligned}$$

■

En utilisant la proposition 6 nous pouvons prendre u_0 moins régulière que $H^2(\mathbb{R})$: $u_0 \in H^s(\mathbb{R})$ avec $s > 1$ suffit.

Ensuite, nous pouvons encore diminuer la régularité de u_0 en utilisant les espaces de Besov.
Notation. *Nous notons $B^{\frac{s}{2}}_{p,q}(\mathbb{R})$ l'espace de Besov dont la norme est définie par $||u||_{B^{\frac{s}{2}}_{p,q}(\mathbb{R})} =$*

$\left(\sum_{j\in\mathbb{Z}}(2^{\frac{sj}{2}}||\Delta_j u||_{L^p})^q\right)^{\frac{1}{q}}$, *où* $u=\sum_{j\in\mathbb{Z}}\Delta_j u$ *est la décomposition de Littlewood-Paley de u.*

En prenant u_0 *dans l'espace* $B^{\frac{1}{2}}_{2,1}(\mathbb{R})$, *nous avons l'existence d'une constante C telle que* $||u_0||_{L^\infty(\mathbb{R})}\leq C||u_0||_{B^{\frac{1}{2}}_{2,1}(\mathbb{R})}$ *(propriété des espaces de Besov qui joue le rôle de l'inégalité de Gagliardo-Nirenberg). Nous n'avons donc plus besoin de l'inégalité du lemme 5 : nous travaillons directement dans l'espace de Besov. L'inégalité* $||u_0-u_0\star\varphi^\varepsilon||_{B^{\frac{1}{2}}_{2,1}(\mathbb{R})}\leq\tilde{C}\varepsilon^{s-\frac{1}{2}}||u_0||_{H^s(\mathbb{R})}$, *(*$\tilde{C}$ *étant une constante indépendante de* ε*), permet de conclure sur le taux de convergence de* $u_0\star\varphi^\varepsilon$ *vers* u_0.

Il nous suffit donc en réalité de prendre $u_0\in H^s(\mathbb{R})$, *pour* $s>\frac{1}{2}$.

Chapitre 3

Équation d'Airy avec contrainte

Nous transformons le problème de Cauchy $(\mathcal{P}_{\text{init non contraint}})$ *de départ en un problème de minimisation afin de pouvoir rajouter plus facilement une contrainte inégalité sur la solution. Nous exposons trois problèmes de minimisation possibles.*

Cette dernière partie s'inscrit dans une série de travaux récents sur la prise en compte de contraintes unilatérales (cf [Berthelin, 2002], [Berthelin et Bouchut, 2003] par exemple, ou encore [Després *et al.*, 2011] où une contrainte convexe est introduite). Nous prendrons dans ce rapport une contrainte générique du type contrainte inégalité, ce qui conduit au problème initial suivant :

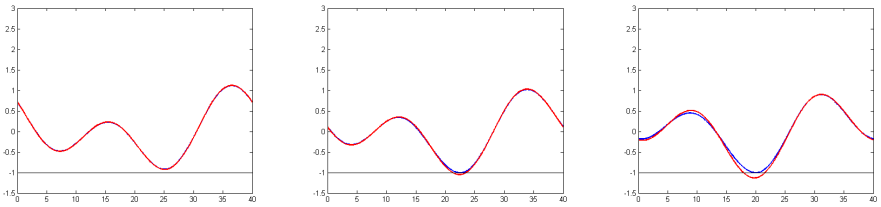
$$(\mathcal{P}_{\text{init contraint}}) \begin{cases} \partial_t u + \partial_x^3 u = 0, \\ u_{|_{t=0}} = u_0, \\ u \geq -1. \end{cases}$$

Nous recherchons là encore une solution $L-$périodique et imposons à nouveau à la donnée initiale de vérifier $\int_0^L u_0(x)dx = 0$.

Toute la question est de donner un sens à ce problème $(\mathcal{P}_{\text{init contraint}})$ puisque l'équation et l'inégalité sont incompatibles. Plus précisément, la solution non contrainte de l'équation d'Airy ne reste pas forcément supérieure à -1 au cours du temps ; et la « solution » contrainte ne vérifie plus l'équation d'Airy. Le travail présenté ici est très prospectif et présente linéairement quelques idées et méthodes testées, sans prétendre être exhaustif.
Nous verrons tout d'abord comment adapter l'étude faite au chapitre 1 pour prendre en compte cette contrainte, nous étudierons nottamment certaines projections à rajouter aux programmes informatiques. Cependant, devant la difficulté à savoir vers quoi la solution discrète converge, nous ramènerons l'étude à un problème de minimisation, problème pour lequel rajouter une contrainte est moins délicat. Ceci fera l'objet de la seconde section.

1 Projections et conservation de l'Hamiltonien

L'idée intuitive est de rajouter la projection suivante aux programmes informatiques permettant de résoudre l'équation d'Airy : $\mathbb{P}_1(U^n) = \max(U^n, -1)$ (U^n représente toujours le vecteur $(u_j^n)_{j\in\mathbb{Z}}$). La figure 3.1 illustre cette projection pour un pas de temps $\Delta t = \frac{1}{10}$, un pas d'espace $\Delta x = \frac{1}{25}$ et une donnée initiale $u_0(x) = \frac{1}{2}\cos(\frac{2\pi}{L}x) + \frac{2}{3}\cos(\frac{4\pi}{L}x)$.



(a) $t = 40$ (b) $t = 70$ (c) $t = 100$

FIGURE 3.1: Schéma implicite ($\theta = 1$) aux différences finies décentré à droite et projection $\mathbb{P}_1(U^n) = \max(U^n, -1)$

La solution théorique non restreinte $u(t,x) = \frac{1}{2}\cos(\frac{2\pi}{L}x + (\frac{2\pi}{L})^3 t) + \frac{2}{3}\cos(\frac{4\pi}{L}x + (\frac{4\pi}{L})^3 t)$ est représentée en rouge tandis que le schéma numérique est représenté en bleu.
Cependant, il existe une autre projection, plus spécifique à l'équation que nous étudions puisqu'elle se base sur une de ses propriétés : la conservation de l'Hamiltonien. Numériquement, la donnée à conserver est

$$\sum_{k\in\mathbb{Z}} \frac{1}{2}\left(\frac{u^n_{k+1} - u^n_k}{\Delta x}\right)^2 \Delta x, \tag{3.1}$$

pour le schéma numérique aux différences finies décentré à droite.

Proposition 7. *La projection $\mathbb{P}_2$ assure la conservation de l'Hamiltonien numérique ainsi que la contrainte $u^n_j \geq -1, \forall (j,n) \in \mathbb{Z}\times[0,N]$, où $\mathbb{P}_2$ est définie par*

$$\underline{\text{Projection } \mathbb{P}_2} : \begin{cases} [\mathbb{P}_2(U^n)]_j = u^n_M + u^n_m - u^n_{M+m-j} \\ \text{avec} \begin{cases} M = \min(l \geq j, u^n_l \geq -1), \\ m = \max(l \leq j, u^n_l \geq -1). \end{cases} \end{cases}$$

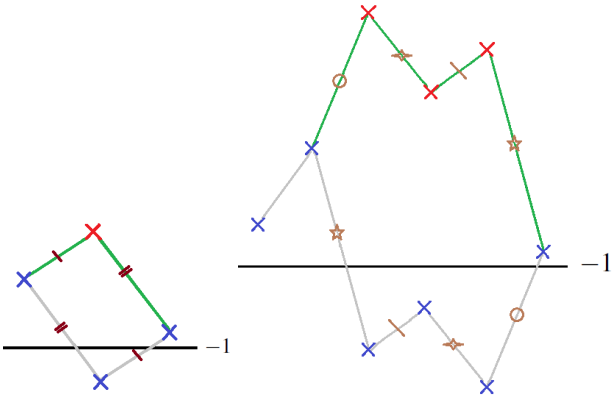
Démonstration.
- Si $u^n_j \geq -1$ alors $M = j$, $m = j$ donc $[\mathbb{P}_2(U^n)]_j = u^n_j$.
- Par définition de M et m, nous avons

$$\begin{aligned} \sum_{j=m}^{M-1} \left([\mathbb{P}_2(U^n)]_{j+1} - [\mathbb{P}_2(U^n)]_j\right)^2 &= \sum_{j=m}^{M-1} \left(u^n_M + u^n_m - u^n_{M+m-j-1} - u^n_M - u^n_m + u^n_{M+m-j}\right)^2 \\ &= \sum_{k=m}^{M-1} (u^n_k - u^n_{k+1})^2. \end{aligned}$$

  L'Hamiltonien numérique est donc bien conservé,
- de même que $[\mathbb{P}_2(U^n)]_j \geq -1 - 1 + 1 = -1$, car $u^n_M \geq -1$, $u^n_m \geq -1$ et $u^n_{M+m-j} \leq -1$.

■

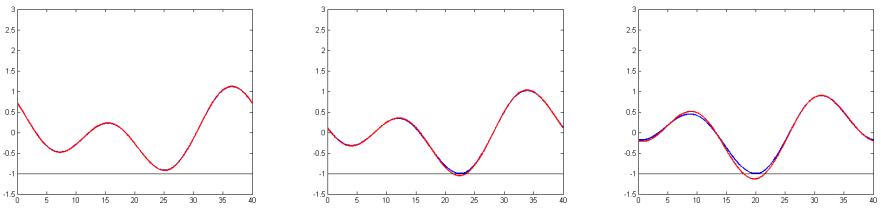
Remarque. *Graphiquement, la projection $\mathbb{P}_2$ revient à considérer le parallélogramme de la figure 3.2. Les points $(u^n_j)_{j\in\mathbb{Z}}$ initiaux sont représentés par les croix bleues. Une fois projetés, ces points deviennent les croix rouges. L'orientation et la longueur des vecteurs $\overrightarrow{u^n_{j-1}u^n_j}$ (vecteurs gris) sont préservés (vecteurs verts), ce qui conserve l'Hamiltonien numérique.*



(a) Un seul j tel que $u_j^n \leq -1$ (b) Plusieurs j tels que $u_j^n \leq -1$

FIGURE 3.2: Projection $\mathbb{P}_2$

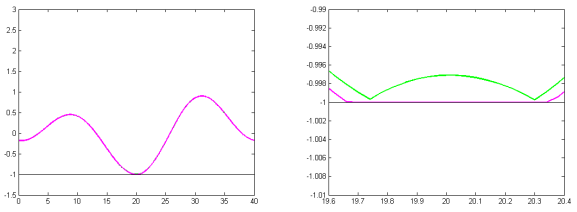
Avec cette projection, nous obtenons la figure 3.3 lorsque nous prenons les mêmes pas d'espace et de temps que pour la projection $\mathbb{P}_1$ ainsi que la même donnée initiale.



(a) $t = 40$ (b) $t = 70$ (c) $t = 100$

FIGURE 3.3: Schéma implicite ($\theta = 1$) aux différences finies décentré à droite et projection $\mathbb{P}_2(U^n)$

Le schéma numérique est toujours représenté en bleu, alors que la courbe rouge représente la solution exacte non contrainte.
L'erreur relative entre les deux solutions obtenues avec les projections $\mathbb{P}_1$ et $\mathbb{P}_2$ vaut 22%. Sur la figure suivante, la projection $\mathbb{P}_1$ est représentée en mauve et la projection $\mathbb{P}_2$ en vert.



(a) $t = 100$ (b) zoom

FIGURE 3.4: Projection $\mathbb{P}_1$ (mauve) et $\mathbb{P}_2$ (vert)

2 Problèmes de minimisation

Ne sachant pas quelle équation doit être vérifiée par la solution contrainte, nous ne pouvons déterminer quelle projection est la plus judicieuse parmi celles proposées. Nous recherchons donc une formulation intrinsèque à l'équation d'Airy afin que le traitement de la contrainte soit inclu dans la formulation : nous essayons ainsi de nous ramener à un problème de minimisation.

2.1 Minimisation et enveloppe convexe

Nous cherchons u solution de $(\mathcal{P}_{\text{init non contraint}})$ comme la solution d'un problème de minimisation, il ne faudra alors plus que rajouter la contrainte $u \geq -1$. Nous posons $u = \partial_x v$; le système satisfait par v est donc le suivant :

$$\begin{cases} \partial_t \partial_x v + \partial_x^4 v = 0, \\ \partial_x v_{|_{t=0}} = u_0. \end{cases}$$

Remarque. *La condition $\int_0^L u_0(x)dx = 0$ est justifiée ici : l'imposer permet à la fonction v d'être elle-encore $L-$périodique.*

Nous allons introduire une méthode que nous réutiliserons par la suite : pour calculer v^{n+1} à partir de v^n, nous posons $v^{n+1} = \lim\limits_{k\to\infty} v^{n,k}$ avec $(v^{n,k})_{k\in\mathbb{Z}}$ solution de

$$\begin{cases} v^{n,0} = v^n, \\ (v^{n,k+1})^{(4)} + \dfrac{(v^{n,k})' - (v^{n,0})'}{\Delta t} = 0, \end{cases} \tag{3.2}$$

Proposition 8. *Si $v^{n,k+1}$ est solution du problème de minimisation suivant :*

$$(\mathscr{P}_1) \qquad v^{n,k+1} = \operatorname*{argmin}_{f\in H^2_{\text{per}}} \int_0^L \frac{(f'')^2}{2}(x) + \left(\frac{(v^{n,k})' - (v^{n,0})'}{\Delta t}\right)(x) f(x) dx,$$

alors $v^{n,k+1}$ est solution de (3.2).

Démonstration. Nous allons montrer que (3.2) est l'équation d'Euler-Lagrange du problème $(\mathscr{P}_1)$.
Soit φ une variation admissible, (*i.e.* $\varphi \in H^2_{\text{per}}$), alors

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\int_0^L \frac{((v^{n,k+1})'' + \varepsilon\varphi'')^2}{2}(x) + \left(\frac{(v^{n,k})' - (v^{n,0})'}{\Delta t}\right)(x)(v^{n,k+1} + \varepsilon\varphi)(x)dx\right]_{|_{\varepsilon=0}} = 0.$$

Nous obtenons :

$$\begin{cases} \forall \varphi \in H^2_{\text{per}}, \\ \displaystyle\int_0^L (v^{n,k+1})''(x)\varphi''(x) + \left(\frac{(v^{n,k})' - (v^{n,0})'}{\Delta t}\right)(x)\varphi(x)dx = 0. \end{cases}$$

Nous notons $\mathscr{C}^\infty_{\text{c, per}}([0,L])$ l'ensemble des fonctions de classe $\mathscr{C}^\infty$, $L-$périodiques, nulles en tous les kL avec $k \in \mathbb{Z}$. Puisque $\mathscr{C}^\infty_{\text{c, per}}([0,L]) \subset H^2_{\text{per}}$, nous avons alors

$$(v^{n,k+1})^{(4)} + \frac{(v^{n,k})' - (v^{n,0})'}{\Delta t} = 0, \text{ dans } \mathscr{D}'.$$

Mais puisque $\frac{(v^{n,k})'-(v^{n,0})'}{\Delta t} \in L^2$, nous en concluons, par la régularité elliptique, que $v^{n,k+1} \in H^4_{\text{per}}$ et que

$$(v^{n,k+1})^{(4)} + \frac{(v^{n,k})' - (v^{n,0})'}{\Delta t} = 0, \text{ presque partout.}$$

■

Nous étudions donc le problème de minimisation sous contrainte suivant :

$$(\mathscr{P}_{1 \text{ contraint}}) \quad v^{n,k+1} = \underset{f \in H^2_{\text{per}} \cap \{f' \geq -1\}}{\operatorname{argmin}} \int_0^L \frac{(f'')^2}{2}(x) + \left(\frac{(v^{n,k})' - (v^{n,0})'}{\Delta t} \right)(x) f(x) dx.$$

Implémentation. Pour traiter informatiquement la contrainte $f' \geq -1$, une idée est de se ramener à calculer une enveloppe convexe. Il s'agit plus particulièrement de suivre la méthode suivante, dite du gradient projeté. À partir de $v^{n,k}$, nous avançons d'un pas dans la direction du gradient (nous notons $v^{n,k+\frac{1}{2}}$ la solution) puis projetons $v^{n,k+\frac{1}{2}}$ sur $H^2_{\text{per}} \cap \{f' \geq -1\}$ grâce à la projection $\mathbb{P}_3$:

$$\underline{\text{Projection } \mathbb{P}_3} : v^{n,k+1} = \mathbb{P}_3(v^{n,k+\frac{1}{2}}) = (\text{conv}(v^{n,k+\frac{1}{2}} + \text{id}))'. \tag{3.3}$$

Nous avons noté $\text{id} : x \to x$, et $\text{conv}(f)$ l'enveloppe convexe de f. Nous répétons en boucle cette méthode pour atteindre la convergence de la suite $(v^{n,k})_{k \in \mathbb{Z}}$.

2.2 Minimisation sur un sous-espace affine

Une autre façon de traiter la contrainte est de semi-discrétiser directement le système $(\mathscr{P}_{\text{init contraint}})$ en temps et d'utiliser la même méthode que précédemment pour avancer de u^n à u^{n+1} à savoir : poser $u^{n,0} = u^n$ puis si $(u^{n,k})_{k \in \mathbb{Z}}$ converge, définir u^{n+1} comme $\lim\limits_{k \to \infty} u^{n,k}$ où $(u^{n,k})_{k \in \mathbb{Z}}$ vérifie

$$\begin{cases} (u^{n,k+1})^{(3)} + \dfrac{u^{n,k} - u^{n,0}}{\Delta t} = 0, & \text{(3.4a)} \\ u^{n,k} \geq -1. & \text{(3.4b)} \end{cases}$$

Nous traitons là encore dans un premier temps l'équation sans la contrainte.

Proposition 9. *Soit $u^{n,k+1}$ solution du problème de minimisation sur l'espace affine $H^1_{\text{per}} \cap \{\int_0^L f(x)dx = \int_0^L u^n(x)dx\}$:*

$$(\mathscr{P}_2) \quad \begin{aligned} u^{n,k+1} &= \frac{1}{L}\int_0^L u^n(x)dx + \bar{u}^{n,k+1}, \\ \bar{u}^{n,k+1} &= \underset{f \in H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}}{\operatorname{argmin}} \int_0^L \frac{(f')^2}{2}(x) - U^{n,k}(x) f(x) dx, \end{aligned}$$

où $U^{n,k}$ est une primitive de $\frac{u^{n,k}-u^{n,0}}{\Delta t}$, alors $u^{n,k+1}$ est solution de l'équation (3.4a).

Démonstration. Recherchons la formulation variationnelle liée à ce problème.
Soit a définie sur $H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\} \times H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}$ par $a(u,v) = \int_0^L u'(x)v'(x)dx$. Alors, a est une forme bilinéaire, symétrique, continue et coercive (avec l'inégalité de Poincaré-Wirtinger qui revient à l'inégalité de Poincaré ici car la moyenne de nos fonctions est nulle). De plus, $H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}$ est un espace de Hilbert, nous pouvons alors appliquer le théorème

de Lax-Milgram.
Il existe donc $\bar{u}^{n,k+1} \in H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}$ tel que

$$\int_0^L (\bar{u}^{n,k+1})'(x)\varphi'(x) - U^{n,k}(x)\varphi(x)dx = 0, \forall \varphi \in H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}.$$

De plus, a étant symétrique, $\bar{u}^{n,k+1}$ est aussi solution du problème $(\mathscr{P}_2)$.(*Nous pouvons aussi considérer l'équation* (3.4a)*, comme l'équation d'Euler-Lagrange du problème* $(\mathscr{P}_2)$*, obtenue par le calcul des variations suivant : soit* $\varphi \in H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\}$*, (*$\varphi$ *est une variation admissible) alors*

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\int_0^L \frac{((u^{n,k+1})' + \varepsilon\varphi')^2}{2}(x) - U^{n,k}(x)(u^{n,k+1} + \varepsilon\varphi)(x)dx\right]_{|\varepsilon=0} = 0.)$$

Soit ensuite $\psi \in H^2_{\text{per}}$ telle que $\psi' = \varphi$, (ceci est possible car $0 = \int_0^L \varphi(x)dx = \int_0^L \psi'(x)dx = \psi(L) - \psi(0)$), alors la formulation variationnelle devient

$$\int_0^L (\bar{u}^{n,k+1})'(x)\psi''(x) - U^{n,k}(x)\psi'(x)dx = 0,$$

soit encore, en prenant $\psi \in \mathscr{C}^\infty_{\text{c, per}}([0,L])$,

$$(\bar{u}^{n,k+1})''' + \underbrace{(U^{n,k})'}_{\frac{u^{n,k}-u^{n,0}}{\Delta t}}(x) = 0, \text{ dans } \mathscr{D}', \text{ ainsi que dans } L^2_{\text{per}},$$

car $\frac{u^{n,k}-u^{n,0}}{\Delta t} \in L^2_{\text{per}}$. Donc l'équation précédente a lieu aussi presque partout.

■

Lorsque nous rajoutons la contrainte, il suffit alors d'étudier le problème

$$(\mathscr{P}_{2 \text{ contraint}}) \quad \begin{aligned} u^{n,k+1} &= \frac{1}{L}\int_0^L u^n(x)dx + \bar{u}^{n,k+1}, \\ \bar{u}^{n,k+1} &= \underset{\substack{f \in H^1_{\text{per}} \cap \{\int_0^L f(x)dx = 0\} \\ \cap \{f \geq -\frac{1}{L}\int_0^L u^n(x)dx - 1\}}}{\operatorname{argmin}} \int_0^L \frac{(f')^2}{2}(x) - U^{n,k}(x)f(x)dx, \end{aligned}$$

où $U^{n,k}$ est une primitive de $\frac{u^{n,k}-u^{n,0}}{\Delta t}$.

Remarque. *Attention, il faut que* $-\frac{1}{L}\int_0^L u^n(x)dx - 1$ *soit négatif pour vérifier en même temps ces deux contraintes* $\{\int_0^L f(x)dx = 0\} \cap \{f \geq -\frac{1}{L}\int_0^L u^n(x)dx - 1\}$*, ce qui est vrai puisque* $\int_0^L u^n(x)dx = \int_0^L u^0(x)dx = 0$.

Implémentation. La prise en compte de la contrainte est plus simple avec cette formulation puisque l'inégalité porte directement sur l'inconnue $\bar{u}^{n,k+1}$ et non plus sur une des ses dérivées. Deux constructions sont possibles afin d'implémenter les conditions $\int_0^L f(x)dx = 0$ et $f \geq -\frac{1}{L}\int_0^L u^n(x)dx - 1$.
La première projection est la plus intuitive mais converge très lentement (et parfois, ne converge pas du tout).

$$\underline{\text{Projection } \mathbb{P}_4} : \begin{cases} \mathbb{P}_4(f) := \max(f - \frac{1}{L}\int_0^L f(x)dx, -\frac{1}{L}\int_0^L u^n(x)dx - 1), \\ \bar{u}^{n,k+1} = \underbrace{\mathbb{P}_4 \circ \mathbb{P}_4 \circ \mathbb{P}_4 \circ \mathbb{P}_4 \circ \cdots \circ \mathbb{P}_4}_{l \text{ fois}}(\bar{u}^{n,k+\frac{1}{2}}), \end{cases}$$

où l est déterminé pour que $|\int_0^L \bar{u}^{n,k+1}(x)dx| \leq \varepsilon$ et $\bar{u}^{n,k+1} \geq -\frac{1}{L}\int_0^L u^n(x)dx - 1$, avec ε la précision voulue, l peut être très grand car la convergence (quand elle a lieu) est souvent lente. $\bar{u}^{n,k+\frac{1}{2}}$ est la solution obtenue en effectuant un pas dans la direction du gradient. Nous recommençons ce calcul de $u^{n,k+1}$ en boucle jusqu'à convergence.

La deuxième projection a l'avantage de converger en un nombre fini de fois, et pour n'importe quelle donnée initiale.
Notation. Nous notons pour plus de lisibilité $\mathcal{C} = -\frac{1}{L}\int_0^L u^n(x)dx - 1$ et définissons $f_{\sup}$ et $f_{\inf}$ pour toute fonction f, de la manière suivante : $f = \underbrace{f\mathbb{1}_{\{f>\mathcal{C}\}}}_{=f_{\sup}} + \underbrace{f\mathbb{1}_{\{f\leq\mathcal{C}\}}}_{=f_{\inf}}$.

Nous définissons alors :

$$\underline{\text{Projection } \mathbb{P}_5} : \begin{cases} \mathbb{P}_5(f) := \begin{cases} \text{si } f \neq f_{\inf} \text{ alors :} \\ (f - \min\left(\inf|f_{\sup} - \mathcal{C}|, \dfrac{\int_0^L (f_{\sup} + \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}})(x)dx}{|\mathbb{1}^{-1}_{\{f>\mathcal{C}\}}(\{1\})|}\right))\mathbb{1}_{\{f>\mathcal{C}\}} + \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}}, \\ \text{si } f = f_{\inf} \text{ alors :} \\ \mathbb{P}_5(f) := 0, \end{cases} \\ \bar{u}^{n,k+1} = \underbrace{\mathbb{P}_5 \circ \mathbb{P}_5 \circ \mathbb{P}_5 \circ \mathbb{P}_5 \circ \cdots \circ \mathbb{P}_5}_{q \text{ fois}}(\bar{u}^{n,k+\frac{1}{2}}), \end{cases}$$

où q est ici toujours fini.

Démonstration. La projection $\mathbb{P}_5$ permet de projeter sur l'ensemble $\{\int_0^L f(x)dx = 0\} \cap \{f \geq -\frac{1}{L}\int_0^L u^n(x)dx - 1\}$.

- En effet, si $\mathbb{P}_5(f) = (f - \frac{\int_0^L (f_{\sup}+\mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}})(x)dx}{|\mathbb{1}^{-1}_{\{f>\mathcal{C}\}}(\{1\})|})\mathbb{1}_{\{f>\mathcal{C}\}} + \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}}$, alors la condition $\mathbb{P}_5(f) \geq \mathcal{C}$ est automatiquement vérifiée ainsi que $\int_0^L \mathbb{P}_5(f)(x)dx = 0$ car

  $$\int_0^L \mathbb{P}_5(f)(x)dx = \int_0^L \underbrace{f(x)\mathbb{1}_{\{f>\mathcal{C}\}}}_{=f_{\sup}} dx - \int_0^L f_{\sup}(x)dx - \int_0^L \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}}(x)dx + \int_0^L \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}}(x)dx = 0.$$

  En une étape la projection est donc réalisée.
- Si $\mathbb{P}_5(f) = (f - \inf|f_{\sup} - \mathcal{C}|)\mathbb{1}_{\{f>\mathcal{C}\}} + \mathcal{C}\mathbb{1}_{\{f\leq\mathcal{C}\}}$. Numériquement, $f_{\sup}$ représente un vecteur, l'infimum est donc atteint en un certain indice j. Alors $\mathbb{P}_5(f) \geq \mathcal{C}$ et

  $$\begin{cases} \forall k \text{ tel que } f_k = [f_{\sup}]_k, \text{ alors } [\mathbb{P}_5(f)]_k = [f_{\sup}]_k - [f_{\sup}]_j + \mathcal{C} \in [\mathcal{C}, [f_{\sup}]_k], \\ \forall k \text{ tel que } f_k = [f_{\inf}]_k, \text{ alors } [\mathbb{P}_5(f)]_k = \mathcal{C}, \\ [\mathbb{P}_5(f)]_j = \mathcal{C}. \end{cases}$$

  Le vecteur $\mathbb{P}_5(f)_{\sup}$ a donc diminué d'un point. À l'étape suivante, ou bien nous sommes ramenés au premier •, l'algorithme a donc convergé. Ou bien nous diminuons encore d'un point le vecteur $\mathbb{P}_5(f)_{\sup}$, en au plus $J_{\sup}$ itérations (où $J_{\sup}$ est la taille de $f_{\sup}$) nous avons donc obtenu : $[\mathbb{P}_5(f)]_k = \mathcal{C}$, $\forall k$ nous sommes donc ramené à la définition de $\mathbb{P}_5$ dans ce cas.

■

Informatiquement, nous approchons $\int_0^L f(x)dx$ par $\sum_{j=1}^{J} f_j \Delta x$; la projection $\mathbb{P}_5$ est donc effectuée en au plus $J+1$ étapes car à chaque passage dans la boucle, les valeurs égales à $\mathcal{C}$ ne sont plus prises en compte, et si aucune valeur est différente de $\mathcal{C}$, alors l'algorithme s'arrête en une seule itération : il suffit de soustraire à $\bar{u}^{n,k+\frac{1}{2}}$ sa moyenne.
Afin de visualiser cette projection, nous avons schématisé les différentes étapes qui la constituent, lorsque $L=1$, ce qui revient à intégrer sur le tore $\mathbb{T}$. Considérons tout d'abord des données discrètes (représentées en vert sur la figure 3.5) correspondant à $\bar{u}^{n,k+\frac{1}{2}}$ calculées avec un pas dans la direction du gradient. La moyenne de $\bar{u}_{\text{sup}}^{n,k+\frac{1}{2}} + \mathcal{C}$ est symbolisée par le trait vert et la contrainte par le trait plein noir. Attention, ce n'est pas tout à fait la moyenne de $\bar{u}^{n,k+\frac{1}{2}}$ que l'on calcule en vert mais une moyenne où nous avons déjà projeté $\bar{u}_{\text{inf}}^{n,k+\frac{1}{2}}$ sur $\mathcal{C}$ car seuls les points de $\bar{u}_{\text{sup}}^{n,k+\frac{1}{2}}$ peuvent encore baisser, nous l'appellerons M.

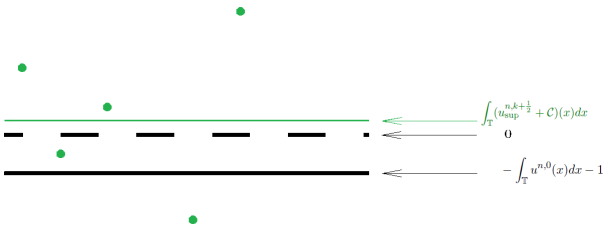


FIGURE 3.5: Donnée $\bar{u}^{n,k+\frac{1}{2}}$ avant projection $\mathbb{P}_5$

Il s'agit ensuite de comparer la valeur à soustraire (ou ajouter) pour imposer M à zero (distance bleue sur la figure 3.6) et les valeurs possibles que l'on peut soustraire à $\bar{u}_{\text{sup}}^{n,k+\frac{1}{2}}$ avant d'atteindre la contrainte $\mathcal{C}$ (valeurs marrons sur la figure 3.6).

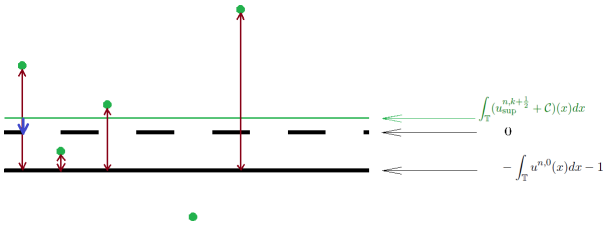


FIGURE 3.6: Annuler M, diminuer $\bar{u}_{\text{sup}}^{n,k+\frac{1}{2}}$ sans dépasser $\mathcal{C}$

Il faut ensuite projeter $\bar{u}_{\text{inf}}^{n,k+\frac{1}{2}}$ sur la contrainte, les valeurs $\bar{u}_{\text{sup}}^{n,k+\frac{1}{2}}$ restent donc inchangées (ce qui se traduit sur la figure 3.7 par la superposition des deux couleurs rouge et bleu).

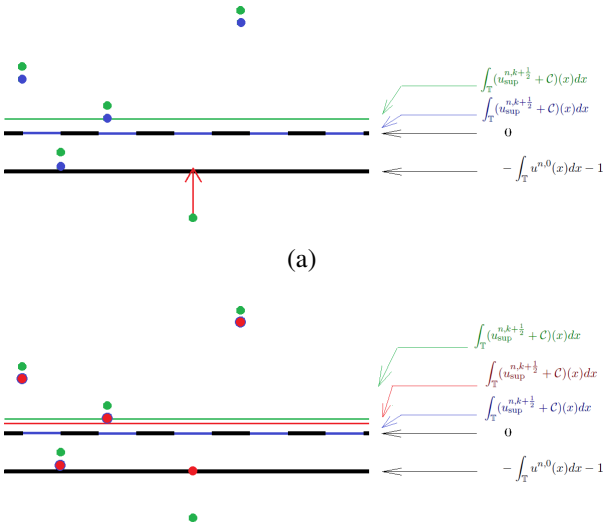


FIGURE 3.7: Imposer $\bar{u}^{n,k+\frac{1}{2}} \geq C$

Puis nous recommençons la projection, mais cette fois en ne déplaçant que les valeurs $\bar{u}_{\sup}^{n,k+\frac{1}{2}}$ (*i.e.* les valeurs rouges-bleues de la figure 3.8). Il faut à nouveau comparer les distances entre chaque point et la contrainte (distances marrons sur la figure 3.8) à la valeur à soustraire pour annuler la moyenne (flèche grise sur la figure 3.8).

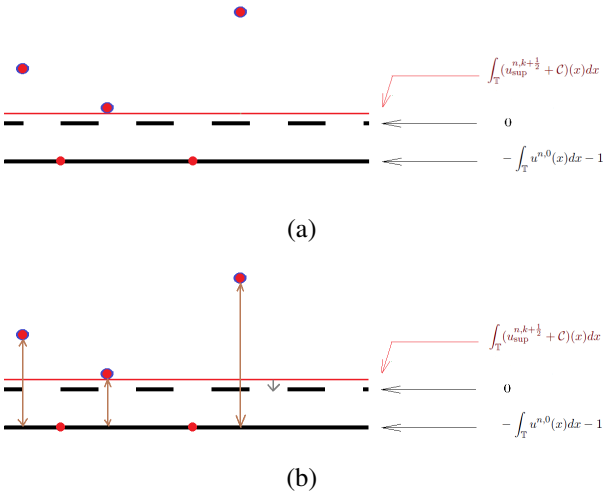


FIGURE 3.8: Une nouvelle projection recommence

2.3 Minimisation de la partie paire

Nous allons chercher une formulation plus intrinsèque encore de l'équation d'Airy. Nous voulons en effet que la solution non contrainte vérifie directement un problème de minimisation sans passer au préalable par une semi-discrétisation, car rien n'assure la convergence de la suite $(u^{n,k})_{k\in\mathbb{Z}}$. Pour cela, nous faisons un rapprochement entre l'équation d'Airy que nous étudions

et un système hamiltonien générique :

$$\begin{cases} \dot{q} = p, \\ \dot{p} = -V(q), \end{cases}$$

avec V convexe (ici, l'hamiltonien vaut $H(q,p) = \frac{p^2}{2} + \mathcal{V}(q)$ avec $\mathcal{V}$ une primitive de V, et le lagrangien $\mathcal{L}(q,\dot{q}) = \frac{(\dot{q})^2}{2} - \mathcal{V}(q)$). L'équation des ondes associée à ce système hamiltonien est $\ddot{q} + V(q) = 0$. C'est l'équation d'Euler-Lagrange associé à la minimisation de l'action $\mathcal{A}$ suivante :

$$q = \underset{\substack{q(0)=q_0 \\ q(T)=q_T}}{\operatorname{argmin}} \underbrace{\int_0^T \mathcal{L}(q(t),\dot{q}(t))dt}_{=\mathcal{A}}.$$

Pour adapter cette méthode à l'équation qui nous intéresse, nous séparons tout d'abord les parties paires et impaires de la solution d'Airy. Soit $u = v + w$ avec v la partie paire de u pour la variable spatiale (*i.e.* $v(t,x) = v(t,-x),\ \forall (t,x) \in [0,T] \times [0,L]$), et w la partie impaire pour la variable spatiale (*i.e.* $w(t,x) = -w(t,-x),\ \forall (t,x) \in [0,T] \times [0,L]$). Alors le système $(\mathcal{P}_{\text{init non contraint}})$ est équivalent à (nous notons v_0 la partie paire de u_0 et w_0 la partie impaire de u_0) :

$$(\mathcal{P}_{v,w}) \begin{cases} \partial_t v + \partial_x^3 w = 0, \\ \partial_t w + \partial_x^3 v = 0, \\ v_{|_{t=0}} = v_0, \\ w_{|_{t=0}} = w_0. \end{cases}$$

Nous imposons de plus la condition $\int_0^L v_0(x)dx = 0$.

Remarque. *L'équivalence entre ces deux système se montre comme suit :*
$\Rightarrow$ *Si u est solution de* $(\mathcal{P}_{\text{init non contraint}})$*, alors en séparant les parties paire et impaire de u nous retrouvons bien les deux premières équations de* $(\mathcal{P}_{v,w})$*. La dernière relation vient du fait que* $\int_0^L w_0(x)dx = 0$ *car* w_0 *est impaire.*
$\Leftarrow$ *si* (v,w) *vérifie le système* $(\mathcal{P}_{v,w})$ *alors il suffit de poser* $u = v + w$ *pour que u soit solution de* $(\mathcal{P}_{\text{init non contraint}})$.

Donc pour résoudre $(\mathcal{P}_{\text{init non contraint}})$, il suffit de résoudre $(\mathcal{P}_{v,w})$, de poser $u = v + w$ et par unicité de la solution de $(\mathcal{P}_{\text{init non contraint}})$, $v + w$ est bien la solution recherchée. Pour trouver v et w, il suffit de rechercher l'équation des ondes satisfaite par la partie paire de u, de calculer son lagrangien pour connaître l'action à minimiser.

L'équation des ondes de $(\mathcal{P}_{v,w})$ est :

$$(\mathcal{P}_{\text{partie paire}}) \begin{cases} \partial_t^2 v - \partial_x^6 v = 0, & \text{(3.5a)} \\ v_{|_{t=0}} = v_0, & \text{(3.5b)} \\ \partial_t v_{|_{t=0}} + \partial_x^3 w_0 = 0, & \text{(3.5c)} \\ \text{on pose } w = -\partial_x^{-3}\partial_t v. & \text{(3.5d)} \end{cases}$$

Nous imposons toujours $\int_0^L v(x)dx = 0$. L'équation en v (3.5a) étant d'ordre 2, il est nécessaire d'ajouter une condition pour assurer l'unicité de la solution, c'est le but de la condition initiale (3.5c). Nous nous sommes donc ramenés à étudier un problème de Cauchy : le système

$(\mathcal{P}_{\text{partie paire}})$. Le lagrangien de $(\mathcal{P}_{\text{partie paire}})$ vaut $\mathcal{L}(v) = \frac{1}{2}|\partial_t v|^2 - \frac{1}{2}|\partial_x^3 v|^2$, l'action à minimiser est donc la suivante

$$(\mathscr{P}_3) \qquad v = \underset{\substack{v(0)=v_0 \\ v(T)=v_T \\ \int_0^L v(x)dx=0 \\ v \text{ paire}}}{\operatorname{argmin}} \int_0^T \int_0^L \frac{1}{2}|\partial_t v|^2 - \frac{1}{2}|\partial_x^3 v|^2 dxdt. \tag{3.6}$$

Lorsque nous rajoutons la contrainte $u \geq -1$ (nous considérons $(\mathcal{P}_{\text{init contraint}})$ au lieu de $(\mathcal{P}_{\text{init non contraint}})$), cela revient à considérer $(\mathcal{P}_{\text{partie paire contraint}})$:

$$(\mathcal{P}_{\text{partie paire contraint}}) \begin{cases} \partial_t^2 v - \partial_x^6 v = 0, & \text{(3.7a)} \\ v_{|t=0} = v_0, & \text{(3.7b)} \\ \partial_t v_{|t=0} + \partial_x^3 w_0 = 0, & \text{(3.7c)} \\ \text{on pose } w = -\partial_x^{-3}\partial_t v, & \text{(3.7d)} \\ v - \partial_x^{-3}\partial_t v \geq -1. & \text{(3.7e)} \end{cases}$$

Le problème de minimisation associé est le suivant :

$$(\mathscr{P}_{3 \text{ contraint}}) \qquad v = \underset{\substack{v(0)=v_0 \\ v(T)=v_T \\ \int_0^L v(x)dx=0 \\ v \text{ paire} \\ v-\partial_x^{-3}\partial_t v \geq -1}}{\operatorname{argmin}} \int_0^T \int_0^L \frac{1}{2}|\partial_t v|^2 - \frac{1}{2}|\partial_x^3 v|^2 dxdt. \tag{3.8}$$

L'inconvénient de cette formulation est que l'action n'est pas convexe. Or la recherche d'un minimiseur passe par la recherche des points critiques, il y a même équivalence lorsque la fonctionnelle à minimiser est convexe. Si ce n'est pas le cas, comme ici, il faut encore prouver que le point critique trouvé numériquement est bien un minimiseur recherché.

Conclusion

Pour une équation dispersive à dérivée en espace d'ordre quelconque $\partial_t u + \partial_x^{2p+1} u = 0,\ p \in \mathbb{N}$, deux paramètres principaux influent sur la convergence des schémas numériques aux différences finies décentrées à gauche, centrées ou décentrées à droite.
Tout d'abord, la parité de p joue un grand rôle dans la stabilité de ces schémas : l'approximation de la dérivée spatiale par des différences finies décentrées à droite sera stable sous condition CFL (du type $\theta \geq \frac{1}{2} - \frac{(\Delta x)^{2p+1}}{2^{2p+1}\Delta t}$ pour un θ-schéma) si p est impair, instable pour p pair. La conclusion est inversée si nous utilisons les différences finies décentrées à gauche.
Ensuite, la régularité de la donnée initiale u_0 agit, elle aussi, sur l'ordre de convergence. Il est à noter qu'il s'agit ici de régularité H^k : l'ordre dépend fortement de l'appartenance ou non de u_0 à $H^{4p+3}(\mathbb{R})$.

La prise en compte de la contrainte inégalité a fait l'objet de nombreuses projections possibles. Les projections $\mathbb{P}_1$ et $\mathbb{P}_2$ ont l'avantage d'être un peu plus intuitives et de vérifier une spécificité de l'équation d'Airy : la conservation de l'Hamiltonien (pour la projection $\mathbb{P}_2$). Les autres projections ont l'avantage, quant à elles, d'être écrites sous forme d'un problème de minimisation (problème dans lequel la prise en compte de la contrainte est plus évident). La dernière formulation $(\mathscr{P}_{3\text{ contraint}})$ est vraisemblablement celle qui se rapproche plus d'une formulation intrinsèque à l'équation d'Airy car elle ne fait pas intervenir de suite intermédiaire $(u^{n,k})_{k\in\mathbb{Z}}$ et permet d'éviter les cas où cette suite intermédiaire ne convergerait pas ou très lentement.

Bibliographie

[Benzoni-Gavage *et al.*, 2006] BENZONI-GAVAGE, S., DANCHIN, R. et DESCOMBES, S. (2006). Well-posedness of one-dimensional korteweg models. *Electron. J. Differential Equations*, pages No. 59, 35 pp. (electronic).

[Berthelin, 2002] BERTHELIN, F. (2002). Existence and weak stability for a pressureless model with unilateral constraint. *M3AS, Mathematical Models and Methods in Applied Science*, Vol.12, No.2, pp.249-272.

[Berthelin et Bouchut, 2003] BERTHELIN, F. et BOUCHUT, F. (2003). Weak solutions for a hyperbolic system with unilateral constraint and mass loss. *Ann. Inst. H. Poincaré Anal. Non Linéraire*, 20 (6), pp. 975-997.

[Brezis, 2005] BREZIS, H. (2005). *Analyse fonctionnelle*. Dunod, collection : Sciences Sup.

[Demailly, 2006] DEMAILLY, J.-P. (2006). *Analyse numérique et équations différentielles*. EDP Sciences, 3e édition.

[Després *et al.*, 2011] DESPRÉS, B., LAGOUTIÈRE, F. et SEGUIN, N. (2011). Weak solutions to friedirchs sytems with convex constraints. Nonlinearity 24, pp.3055-3081.

[Kanazawa *et al.*, 2012] KANAZAWA, H., MATSUO, T. et YAGUCHI, T. (2012). A conservative compact finite difference scheme for the kdv equation. *JSIAM Letters*, Vol.4, pp.5-8.

[Zuily, 2002] ZUILY, C. (2002). *Éléments de distributions et d'équations aux dérivées partielles*. Dunod, collection : Sciences Sup.

[Zuily et Queffélec, 2007] ZUILY, C. et QUEFFÉLEC, H. (2007). *Analyse pour l'agrégation*. Dunod, 3e édition.