Université catholique de Louvain

Ecole de physique

Simulation numérique en physique [lphy2371]

Equation d'Advection-Diffusion et Prédictibilité

Auteurs :
Arnaud Schils
Valéry Materne

Enseignant :
Pr. Michel Crucifix

Décembre 2016

Première partie

Equation d'Advection-Diffusion

1.1 Fournissez un schéma numérique explicite d'ordre $\mathcal{O}(h+k)$. Donnez-en le stencil. Nous pouvons introduire les facteurs $\lambda_d = Dk/h^2$, $\lambda_a = ak/h$ et $\lambda_b = bk$. Quelles importances ces facteurs ont-ils pour la condition de stabilité ?

Dans cette section un schéma numérique explicite est fourni pour l'équation aux dérivées partielles suivantes :

$$D\frac{\partial^2 u(x,t)}{\partial x^2} - a\frac{\partial u(x,t)}{\partial x} - bu(x,t) = \frac{\partial u(x,t)}{\partial t} \tag{1}$$

où D, a, b sont des constantes et D est positive. Dans la suite de ce texte les dépendances en x et t de la fonction u ne seront pas toujours mentionnées explicitement.

Schéma numérique explicite

Les dérivées de l'équation (1) sont remplacées par leurs expressions en différences finies. Soient $u(x_i, t_j) \equiv U_{i,j}$, h le pas d'espace et k le pas de temps, ces expressions sont :

$$\left.\frac{\partial^2 u(x,t)}{\partial x^2}\right|_{x=x_i,t=t_j} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \mathcal{O}(h^2) \tag{2}$$

$$\left.\frac{\partial u(x,t)}{\partial x}\right|_{x=x_i,t=t_j} = \frac{U_{i+1,j} - U_{i,j}}{h} + \mathcal{O}(h) \tag{3}$$

$$\left.\frac{\partial u(x,t)}{\partial t}\right|_{x=x_i,t=t_j} = \frac{U_{i,j+1} - U_{i,j}}{k} + \mathcal{O}(k)\,. \tag{4}$$

Nous avons pris une différence centrée pour la dérivée seconde par rapport à x et une différence avant pour les dérivées premières par rapport à x et t. Notons que pour la dérivée seconde par rapport à x, la formule à trois points d'ordre $\mathcal{O}(h^2)$ a été choisie au lieu de celles d'ordre $\mathcal{O}(h)$ afin d'obtenir à la fin

une matrice tridiagonale. En injectant ces différences finies dans l'équation (1) on obtient :

$$D\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \mathcal{O}(h^2) - a\frac{U_{i+1,j} - U_{i,j}}{h} + \mathcal{O}(h) - bU_{i,j} = \frac{U_{i,j+1} - U_{i,j}}{k} + \mathcal{O}(k)\ . \quad (5)$$

En multipliant l'expression par le pas de temps k, en isolant $U_{i,j+1}$ et en négligeant l'erreur en $\mathcal{O}(h^2)$ car elle est d'ordre supérieur à $\mathcal{O}(h)$ on obtient :

$$U_{i,j+1} = \frac{Dk}{h^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) - \frac{ak}{h}(U_{i+1,j} - U_{i,j}) - bkU_{i,j} + U_{i,j} + \mathcal{O}(h+k)\ . \quad (6)$$

En définissant $\lambda_d = \frac{Dk}{h^2}$, $\lambda_a = \frac{ak}{h}$ et $\lambda_b = bk$ l'expression devient :

$$U_{i,j+1} = \lambda_d(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) - \lambda_a(U_{i+1,j} - U_{i,j}) - \lambda_b U_{i,j} + U_{i,j} + \mathcal{O}(h+k)\ . \quad (7)$$

Sans spécifier l'ordre de l'erreur et en réarrangeant l'expression on a :

$$U_{i,j+1} = U_{i-1,j}\lambda_d + U_{i,j}(-2\lambda_d + \lambda_a - \lambda_b + 1) + U_{i+1,j}(\lambda_d - \lambda_a)\ . \quad (8)$$

En imposant les conditions aux bords constantes $\forall j, U_{0,j} = 0$ et $\forall j, U_{N+1,j} = 0$, le schéma numérique peut s'écrire sous forme matricielle comme ceci :

$$M = \begin{pmatrix} 1 & 0 & ... & 0 & 0 \\ \lambda_d & -2\lambda_d + \lambda_a - \lambda_b + 1 & \lambda_d - \lambda_a & 0 & ... \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & ... & \lambda_d & -2\lambda_d + \lambda_a - \lambda_b + 1 & \lambda_d - \lambda_a \\ 0 & 0 & ... & 0 & 1 \end{pmatrix} \quad (9)$$

$$\begin{pmatrix} U_{0,j+1} \\ U_{1,j+1} \\ . \\ . \\ . \\ U_{N,j+1} \\ U_{N+1,j+1} \end{pmatrix} = M \begin{pmatrix} U_{0,j} \\ U_{1,j} \\ . \\ . \\ . \\ U_{N,j} \\ U_{N+1,j} \end{pmatrix} \tag{10}$$

où M est une matrice tridiagonale. Notons que les $U_{i,0}$ sont connus grâce aux conditions initiales.

Stencil

Le stencil de ce schéma explicite est présenté à la figure 1. Chaque $U_{i,j+1}$ dépend en effet des $U_{i-1,j}$, $U_{i,j}$ et $U_{i+1,j}$.

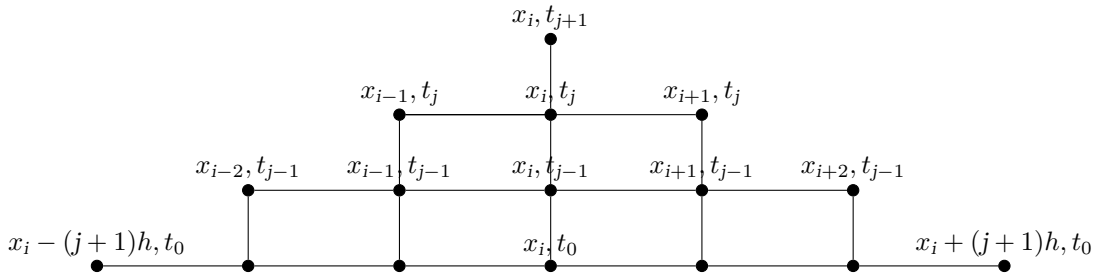

FIGURE 1 – Stencil du schéma numérique explicite.

Importance des facteurs λ_d, λ_a et λ_b pour la condition de stabilité

Ces facteurs sont reliés aux conditions de stabilité de la méthode aux différences finies. Pour le montrer, introduisons dans notre schéma numérique une solution de la forme :

$$U_{i,j} = w_j \exp(irx_i) \ , \tag{11}$$

où $i \equiv \sqrt{-1}$. Nous injectons cette solution dans l'équation (8), on obtient :

$$w_{j+1}e^{irx_i} = w_j e^{irx_{i\text{-}1}}\lambda_d + w_j e^{irx_i}(-2\lambda_d + \lambda_a - \lambda_b + 1) + w_j e^{irx_{i+1}}(\lambda_d - \lambda_a) \quad (12)$$

En tenant compte du pas d'espace $x_{i+1} - x_i = h$, après calcul :

$$\begin{aligned} w_{j+1} &= w_j e^{-irh}\lambda_d + w_j(-2\lambda_d + \lambda_a - \lambda_b + 1) + w_j e^{irh}(\lambda_d - \lambda_a) \\ &= w_j\left(\lambda_d\left(e^{-irh} - 2 + e^{irh}\right) + \lambda_a\left(1 - e^{irh}\right) - \lambda_b + 1\right) \\ &= w_j\left(1 + \lambda_d\left(2\cos(rh) - 2\right) + \lambda_a\left(1 - e^{irh}\right) - \lambda_b\right) \\ &= w_j\left(1 - 4\lambda_d \sin^2(rh/2) + \lambda_a\left(1 - e^{irh}\right) - \lambda_b\right) \ . \end{aligned} \quad (13)$$

Il s'agit d'une equation au différences, que l'on peut résoudre en posant :

$$w_j = w_0\kappa^j \ . \quad (14)$$

On obtient alors $\kappa \equiv 1 - 4\lambda_d \sin^2(rh/2) + \lambda_a\left(1 - e^{irh}\right) - \lambda_b$ qui est complexe. La condition de stabilité se ramène alors à :

$$|\kappa|^2 = \left(1 - 4\lambda_d \sin^2(rh/2) + \lambda_a\left(1 - \cos(rh)\right) - \lambda_b\right)^2 + \lambda_a^2 \sin^2(rh) \leq 1 \ . \quad (15)$$

Les valeurs des paramètres λ_d, λ_a et λ_b déterminent donc la stabilité du schéma numérique. L'expression peut se réécrire :

$$\boxed{|\kappa|^2 = \left(1 + 2\sin^2(rh/2)\left(\lambda_a - 2\lambda_d\right) - \lambda_b\right)^2 + \lambda_a^2 \sin^2(rh) \leq 1} \ . \quad (16)$$

L'expression est trop compliquée pour pouvoir trouver une condition exacte de stabilité sur les différents λ. Par contre, on peut trouver des conditions qui guarantissent stabilité pour certaines valeurs des λ sans pour autant pouvoir déterminer toutes les combinaisons de λ stables.

Une condition pessimiste peut êre obtenue en posant les valeurs des sinus à celles qui maximisent chacun des termes de l'équation (16). Le pire cas pour le terme de droite est toujours $\sin^2(rh) = 1$. Pour le terme de gauche, cela

dépend des signes et valeurs respectives des différents λ. Deux cas peuvent être distingués.

Si $|1 - \lambda_b| > |1 - \lambda_b + 2(\lambda_a - 2\lambda_d)|$, le terme de gauche est maximisé si $\sin^2(rh/2) = 0$. La condition pessimiste de stabilité est alors :

$$(1 - \lambda_b)^2 + \lambda_a^2 \leq 1 \implies \lambda_b(\lambda_b - 2) + \lambda_a^2 \leq 0 \tag{17}$$

Sinon, le terme de gauche est maximisé si $\sin^2(rh/2) = 1$. La condition pessimiste de stabilité est alors dans ce deuxième cas :

$$(1 + 2(\lambda_a - 2\lambda_d) - \lambda_b)^2 + \lambda_a^2 \leq 1 \, . \tag{18}$$

A partir de ces conditions, une fonction Matlab is_stable_expl a été implémentée. Lorsque celle-ci renvoie 1 (le booléen true) cela signifie que la méthode est stable. Si elle renvoie 0 (le booléen false), cela ne signifie rien (on ne sait pas si la méthode sera stable ou non).

Il a été vérifié pour toutes les combinaisons de paramètres possibles parmi $a \in \{-10, 1, 10\}$, $b \in \{-10, 1, 10\}, D \in \{1, 10, 50\}, h \in \{0.0001, 0.001, 0.01\}$ et $k \in \{0.000001, 0.00001, 0.0001\}$ que la méthode numérique est effectivement stable lorsque la fonction is_stable_expl renvoie true.

1.2 Fournissez la solution analytique de l'équation, ainsi que la relation de dispersion. Discutez brièvement les cas particuliers déjà vus au cours (D=0, b=0, etc.).

L'équation aux dérivées partielles à résoudre est l'équation (1). Les conditions aux bords suivantes sont imposées :

$$\begin{cases} u(x, 0) = g(x) \\ u(0, t) = u(l, t) = 0 \end{cases} \tag{19}$$

où $l > 0$. La solution u est donc recherchée dans le domaine :

$$\begin{cases} t \geq 0 \\ 0 < x < l \end{cases} \tag{20}$$

L'équation (1) peut-être résolue par la méthode de séparation de variables. La solution u est supposée être de la forme :

$$u(x,t) = v(x)w(t) \tag{21}$$

En injectant cette forme de u dans l'équation (1) on obtient :

$$Dw(t)\frac{\partial^2 v(x)}{\partial x^2} - aw(t)\frac{\partial v(x)}{\partial x} - bv(x)w(t) = v(x)\frac{\partial w(t)}{\partial t} \tag{22}$$

En divisant cette équation par $v(x)w(t)$ on obtient :

$$\frac{D}{v}\frac{\partial^2 v}{\partial x^2} - \frac{a}{v}\frac{\partial v}{\partial x} - b = \frac{1}{w}\frac{\partial w}{\partial t} \equiv C_1 \tag{23}$$

Chaque partie de l'équation est en effet égale à une constante C_1 puisque la partie gauche ne dépend que de x et la partie droite ne dépend que de t. L'équation peut maintenant être résolue en résolvant séparément la partie qui dépend du temps t et la partie qui dépend de la position x. Pour la partie dépendante du temps on a :

$$\frac{1}{w}\frac{\partial w}{\partial t} = C_1 \tag{24}$$

$$\frac{\partial w}{\partial t} = C_1 w \tag{25}$$

$$w(t) = C_2 e^{C_1 t} \tag{26}$$

où C_2 est une constante. Pour la partie dépendante de la position x on a :

$$\frac{D}{v}\frac{\partial^2 v}{\partial x^2} - \frac{a}{v}\frac{\partial v}{\partial x} = C_1 + b \tag{27}$$

$$D\frac{d^2 v}{dx^2} - a\frac{dv}{dx} - (C_1 + b)v = 0 \tag{28}$$

C'est une équation différentielle linéaire homogène du 2ème ordre. Sa solution dépend donc de son polynôme caractéristique :

$$Dr^2 - ar - (C_1 + b) = 0 \tag{29}$$
$$\rho = a^2 + 4D(C_1 + b) \tag{30}$$

Les deux racines de ce polynôme sont :

$$r_{1,2} = \frac{a \pm \sqrt{\rho}}{2D} \tag{31}$$

En fonction du signe de ρ la solution de l'équation peut avoir trois formes.

Si $\rho > 0$

$$v(x) = C_3 e^{r_1 x} + C_4 e^{r_2 x} \tag{32}$$

où C_3 et C_4 sont des constantes. En utilisant la condition au bord $u(0,t) = 0 = v(0)w(t) \implies v(0) = 0$ on a :

$$C_4 = -C_3 \implies v(x) = C_3(e^{r_1 x} - e^{r_2 x}) \tag{33}$$

Et en utilisant la condition au bord $u(l,t) = 0 = v(l)w(t) \implies v(l) = 0$ on a :

$$C_3(e^{r_1 l} - e^{r_2 l}) = 0 \implies C_3 = 0 \implies v(x) = 0 \tag{34}$$

Cette solution n'est donc pas intéressante par rapport à nos conditions aux bords.

Si $\rho = 0, r_1 = r_2 \equiv r$

$$v(x) = (C_3 + C_4 x)e^{rx} \tag{35}$$

où C_3 et C_4 sont des constantes. En utilisant la condition au bord $u(0,t) = 0 = v(0)w(t) \implies v(0) = 0$ on a :

$$C_3 = 0 \implies v(x) = C_4 x e^{rx} \tag{36}$$

Et en utilisant la condition au bord $u(l,t) = 0 = v(l)w(t) \implies v(l) = 0$ on a :

$$C_4 l e^{rl} = 0 \implies C_4 = 0 \implies v(x) = 0 \tag{37}$$

Cette solution n'est donc pas intéressante par rapport à nos conditions aux bords.

Si $\rho < 0$

On définit

$$r_{1,2} \equiv \alpha \pm i\beta \tag{38}$$

avec

$$\alpha = \frac{a}{2D} , \tag{39}$$

$$\beta = \frac{\sqrt{-a^2 - 4D(C_1 + b)}}{2D} . \tag{40}$$

On a alors comme solution pour $v(x)$:

$$v(x) = (C_3 \cos(\beta x) + C_4 \sin(\beta x))e^{\alpha x} . \tag{41}$$

En utilisant la condition au bord $u(0,t) = 0 = v(0)w(t) \implies v(0) = 0$ on a :

$$C_3 = 0 \implies v(x) = C_4 \sin(\beta x)e^{\alpha x} \tag{42}$$

Et en utilisant la condition au bord $u(l,t) = 0 = v(l)w(t) \implies v(l) = 0$ on a :

$$\sin(\beta l) = 0 \implies \beta l = m\pi \implies \beta = \frac{m\pi}{l}, m \in \mathbb{N}^* . \tag{43}$$

On notera que $\beta > 0$ car $\rho < 0$ et $D > 0$.
La solution générale pour la partie de l'équation dépendante de la position est donc :

$$v(x) = \sum_{m=1}^{\infty} C_{4_m} \sin\left(\frac{m\pi x}{l}\right) e^{\alpha x} \tag{44}$$

où les C_{4_m} sont des constantes. Utilisons maintenant la condition au bord $u(x,0) = g(x)$ afin de déterminer les valeurs de ces constantes C_{4_m} :

$$u(x,0) = v(x)w(0) = v(x)C_2 = g(x)\ . \tag{45}$$

On a alors,

$$g(x) = \sum_{m=1}^{\infty} C_{4_m} C_2 \sin\left(\frac{m\pi x}{l}\right) e^{\alpha x}\ . \tag{46}$$

En définissant $C_m = C_{4_m} C_2$ on obtient :

$$g(x) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi x}{l}\right) e^{\alpha x} \tag{47}$$

On voit dès lors que les coefficients C_m sont obtenus en projetant la fonction $g(x)e^{-\alpha x}$ sur la base des fonctions $\sin\left(\frac{m\pi x}{l}\right)$:

$$C_m = \frac{2}{l}\int_0^l g(x) \sin\left(\frac{m\pi x}{l}\right) e^{-\alpha x} dx \tag{48}$$

La fonction $u(x,t)$ peut alors s'écrire :

$$u(x,t) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi x}{l}\right) e^{\alpha x} e^{C_{1_m} t} \tag{49}$$

Nous devons maintenant déterminer l'expression de la constante C_{1_m} qui dépend de β et donc de m. En partant de l'expression de β (40), on obtient :

$$C_{1_m} = \frac{-a^2 - 4D^2\beta^2}{4D} - b \tag{50}$$

et avec $\beta = \frac{m\pi}{l}$, on a :

$$C_{1_m} = -\frac{a^2}{4D} - \frac{Dm^2\pi^2}{l^2} - b\ . \tag{51}$$

En injectant l'expression de C_{1_m} dans celle de u on a alors :

$$u(x,t) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi x}{l}\right) \exp\left(\frac{ax}{2D} + \left(-\frac{a^2}{4D} - \frac{Dm^2\pi^2}{l^2} - b\right)t\right) \tag{52}$$

$$\boxed{u(x,t) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi x}{l}\right) \exp\left(\frac{a}{2D}\left(x - \frac{a}{2}t\right) - \left(\frac{Dm^2\pi^2}{l^2} + b\right)t\right)} \tag{53}$$

avec

$$\boxed{C_m = \frac{2}{l}\int_0^l g(x) \sin\left(\frac{m\pi x}{l}\right) \exp\left(-\frac{a}{2D}x\right) dx}\,. \tag{54}$$

Relation de dispersion

On utilise, comme solution à l'équation d'advection-diffusion (1), une onde plane de la forme :

$$u(x,t) = e^{i(kx-\omega t)} \tag{55}$$

avec k le nombre d'onde qui est relié à la longueur d'onde λ par la relation $\lambda = 2\pi/k$ et ω la fréquence de l'onde.

On obtient alors l'expression suivante :

$$\omega = ak - i\left(Dk^2 + b\right) \tag{56}$$

qui est la relation de dispersion de l'équation (1). Elle détermine la fréquence ω en fonction de k pour une onde plane prise comme solution.

Cas particuliers

Si $D = 1$, $a = b = 0$ on retrouve l'équation de la chaleur adimensionnelle et homogène :

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{\partial u(x,t)}{\partial t} \ . \tag{57}$$

En injectant ces valeurs de D, a et b dans l'équation (53) on a :

$$u(x,t) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi x}{l}\right) \exp\left(-\left(\frac{m^2\pi^2}{l^2}\right) t\right) \tag{58}$$

avec

$$C_m = \frac{2}{l} \int_0^l g(x) \sin\left(\frac{m\pi x}{l}\right) dx \ . \tag{59}$$

On retombe donc bien sur la solution de l'équation de la chaleur adimensionnelle homogène du cours si l'on pose $l = 1$.

Si $D = 0 = b$ on retrouve la forme générale de l'équation d'advection :

$$\frac{\partial u(x,t)}{\partial t} + a\frac{\partial u(x,t)}{\partial x} = 0 \ . \tag{60}$$

La solution présentée à l'équation (53) n'est alors plus valide car celle-ci n'est valable que si $D > 0$ et $\rho < 0$. La solution de l'équation (60) est $u(x,t) = g(x - at)$ avec comme condition initiale $u(x,0) = g(x)$ (démonstration faite en séance d'exercices).

1.3 Fournissez un schéma implicite. Est-il toujours stable ?

Le schéma implicite est obtenu en discrétisant l'Equation 1 de la façon suivante. La fonction u est remplacée par l'image du point (x_i, t_{j+1}) c'est à dire par $U_{i,j+1}$. Les dérivées sont remplacées par leurs formulations discrètes au temps t_{j+1} :

$$\left.\frac{\partial^2 u(x,t)}{\partial x^2}\right|_{x=x_i,t=t_{j+1}} = \frac{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}}{h^2} + \mathcal{O}(h^2) \quad (61)$$

$$\left.\frac{\partial u(x,t)}{\partial x}\right|_{x=x_i,t=t_{j+1}} = \frac{U_{i+1,j+1} - U_{i,j+1}}{h} + \mathcal{O}(h) \quad (62)$$

$$\left.\frac{\partial u(x,t)}{\partial t}\right|_{x=x_i,t=t_{j+1}} = \frac{U_{i,j+1} - U_{i,j}}{k} + \mathcal{O}(k)\,. \quad (63)$$

On obtient alors :

$$\frac{U_{i,j+1} - U_{i,j}}{k} + \mathcal{O}(k) = D\frac{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1}}{h^2} + \mathcal{O}(h^2) - a\frac{U_{i+1,j+1} - U_{i,j+1}}{h} + \mathcal{O}(h) - bU_{i,j+1}\,. \quad (64)$$

Le schéma est d'ordre $\mathcal{O}(h+k)$. En isolant $U_{i,j}$ et en introduisant les quantités λ_a, λ_b et λ_d depuis leurs définitions on obtient :

$$U_{i,j} = -\lambda_d U_{i\text{-}1,j+1} + U_{i,j+1}(1 + 2\lambda_d + \lambda_b - \lambda_a) + U_{i+1,j+1}(\lambda_a - \lambda_d) \quad (65)$$

La matrice M correspondante est présentée ci-dessous.

$$M = \begin{pmatrix} 1 & 0 & ... & 0 & 0 \\ -\lambda_d & 2\lambda_d - \lambda_a + \lambda_b + 1 & \lambda_a - \lambda_d & 0 & ... \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & ... & -\lambda_d & 2\lambda_d - \lambda_a + \lambda_b + 1 & \lambda_a - \lambda_d \\ 0 & 0 & ... & 0 & 1 \end{pmatrix} \quad (66)$$

$$\begin{pmatrix} U_{0,j+1} \\ U_{1,j+1} \\ . \\ . \\ . \\ U_{N,j+1} \\ U_{N+1,j+1} \end{pmatrix} = M^{-1} \begin{pmatrix} U_{0,j} \\ U_{1,j} \\ . \\ . \\ . \\ U_{N,j} \\ U_{N+1,j} \end{pmatrix} \quad (67)$$

où M est une matrice tridiagonale. Notons que les $U_{i,0}$ sont connus grâce aux conditions initiales.

Stabilité

Afin de trouver une condition de stabilité pour la méthode implicite injectons l'équation (11) dans le schéma numérique implicite (équation (65)). On obtient alors :

$$w_j e^{irx_i} = -\lambda_d w_{j+1} e^{irx_{i-1}} + w_{j+1} e^{irx_i}(1 + 2\lambda_d + \lambda_b - \lambda_a) + w_{j+1} e^{irx_{i+1}}(\lambda_a - \lambda_d) \; . \tag{68}$$

$$w_j = w_{j+1}\left(-\lambda_d e^{-irh} + 1 + 2\lambda_d + \lambda_b - \lambda_a + e^{irh}(\lambda_a - \lambda_d)\right) \tag{69}$$

Et donc,

$$w_{j+1} = \kappa w_j \Leftrightarrow w_j = \kappa w_{j-1} = \kappa^j w_0 \tag{70}$$

avec

$$\kappa \equiv \left[-\lambda_d e^{-irh} + 1 + 2\lambda_d + \lambda_b - \lambda_a + e^{irh}(\lambda_a - \lambda_d)\right]^{-1} \equiv \alpha^{-1} \; . \tag{71}$$

On a,

$$\begin{aligned} \alpha &= -\lambda_d(2\cos(rh) - 2) + 1 + \lambda_b + \lambda_a(e^{irh} - 1) \\ &= 4\lambda_d \sin^2(rh/2) + 1 + \lambda_b - 2\lambda_a \sin^2(rh/2) + i\lambda_a \sin(rh) \equiv a + ib \; . \end{aligned} \tag{72}$$

Afin que la méthode numérique soit stable on veut donc que

$$|\kappa|^2 \leq 1 \Leftrightarrow |\alpha^{-1}|^2 \leq 1 \; . \tag{73}$$

Par ailleurs,

$$\alpha^{-1} = \frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2} \; . \tag{74}$$

La condition de stabilité devient donc :

$$\frac{1}{(a^2+b^2)^2}|a-ib|^2 = \frac{a^2+b^2}{(a^2+b^2)^2} = \frac{1}{a^2+b^2} \leq 1 \implies a^2+b^2 \geq 1\,. \qquad (75)$$

On a finalement :

$$\boxed{\Big(2\sin^2(rh/2)(2\lambda_d-\lambda_a)+1+\lambda_b\Big)^2+\lambda_a^2\sin^2(rh) \geq 1}\,. \qquad (76)$$

Le schéma n'est donc pas toujours stable. En effet par exemple lorsque $\sin^2(rh/2) = 0 = \sin^2(rh)$ et $\lambda_b = -1$. Ou encore lorsque $\lambda_d = \lambda_a/2$, $\lambda_b = -1$ et $\lambda_a < 1$ (voir Figure 10).

1.4 Illustrez votre propos avec plusieurs simulations numériques, comparant schémas numériques implicites et explicites.

Pour tous les graphiques présentés dans cette section, la condition initiale utilisée est $g(x) = 1+\cos(8x\pi/l+\pi)$. On peut constater qu'elle satisfait bien aux conditions aux bords imposées $g(0) = g(l) = 0$.

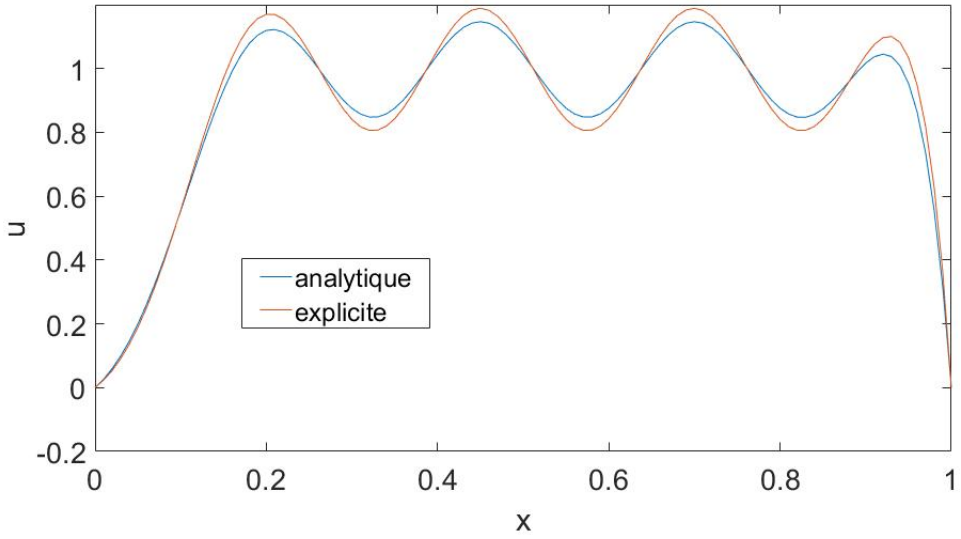

FIGURE 2 – Comparaison entre les solutions analytique et numérique explicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.003$.

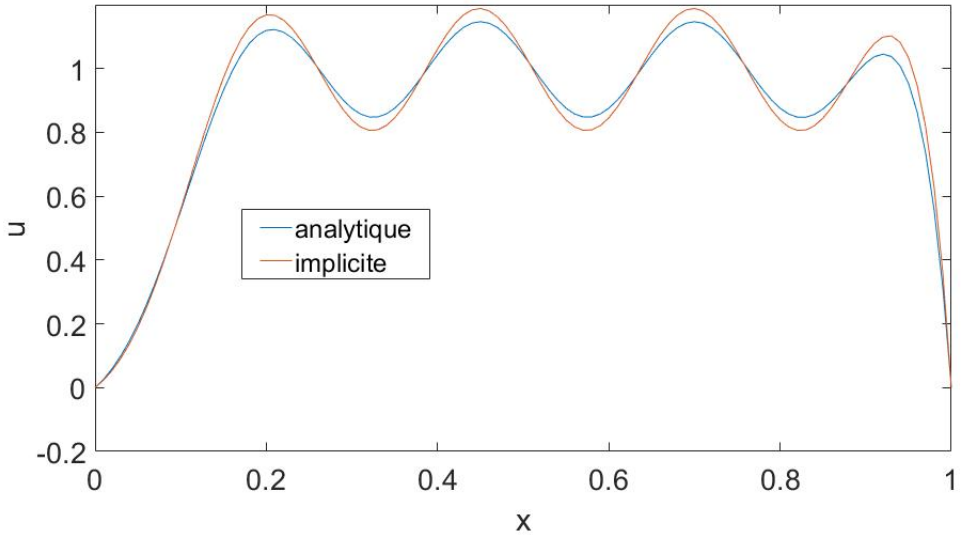

FIGURE 3 – Comparaison entre les solutions analytique et numérique implicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.003$.

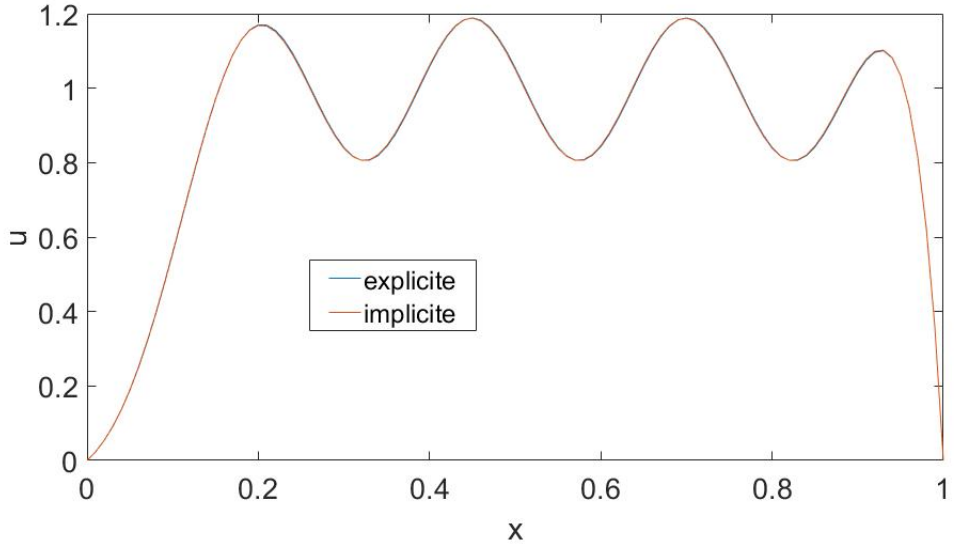

FIGURE 4 – Comparaison entre les solutions numériques implicite et explicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.003$.

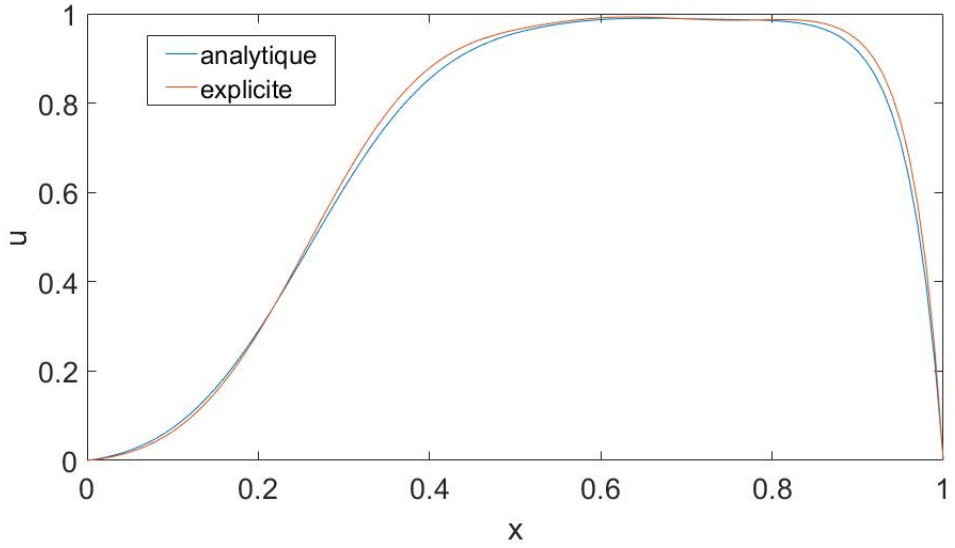

FIGURE 5 – Comparaison entre les solutions analytique et numérique explicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.01$.

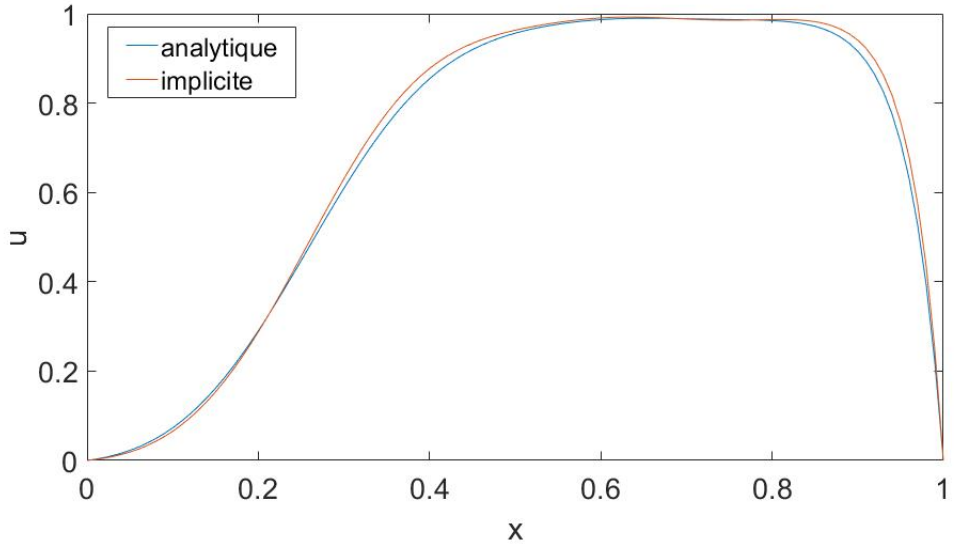

FIGURE 6 – Comparaison entre les solutions analytique et numérique implicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.01$.

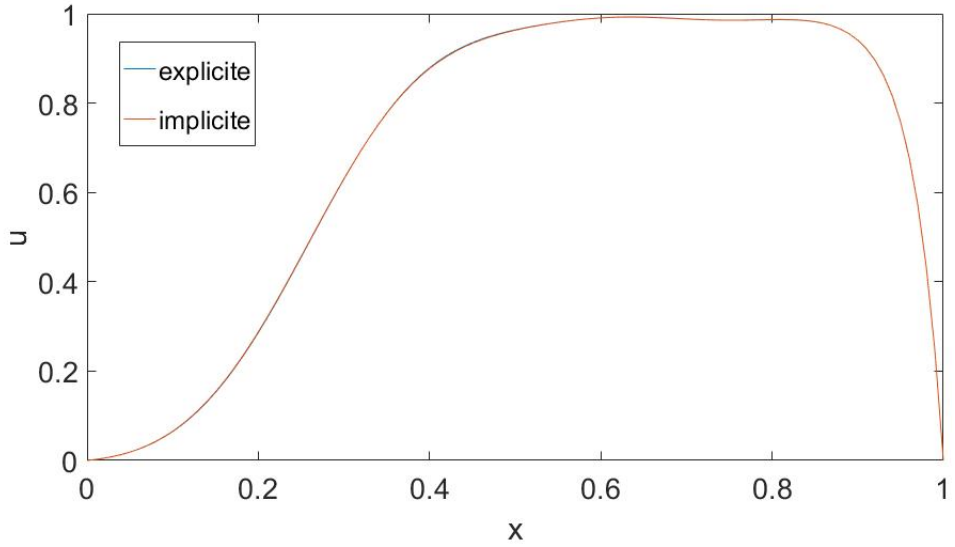

FIGURE 7 – Comparaison entre les solutions numériques implicite et explicite pour les paramètres $a = 25$, $b = 1$, $D = 1$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.01$.

On peut constater sur les Figures 4 et 7 que les solutions numériques implicite et explicite sont quasi identiques pour les paramètres, conditions aux bords et conditions initiales choisis.

On peut également constanter que dans certains cas, lorsque notre condition pessimiste de stabilité pour la méthode explicite est violée, le comportement est effectivement instable (voir Figure 8). En effet, pour $a = 1$, $b = -10$, $D = 10$, $h = 0.01$ et $k = 0.00001$, on a $\lambda_a = 1000$, $\lambda_b = -0.0001$ et $\lambda_d = 1$. Dès lors, $|1 - \lambda_b| = 1.0001 < |1 - \lambda_b + 2(\lambda_a - 2\lambda_d)| = 1997$ et la condition pessimiste de stabilité est donnée par l'Equation 18. Ell est en effet non respectée :

$$(1 + 2(\lambda_a - 2\lambda_d) - \lambda_b)^2 + \lambda_a^2 = 4988009 > 1 \ . \tag{77}$$

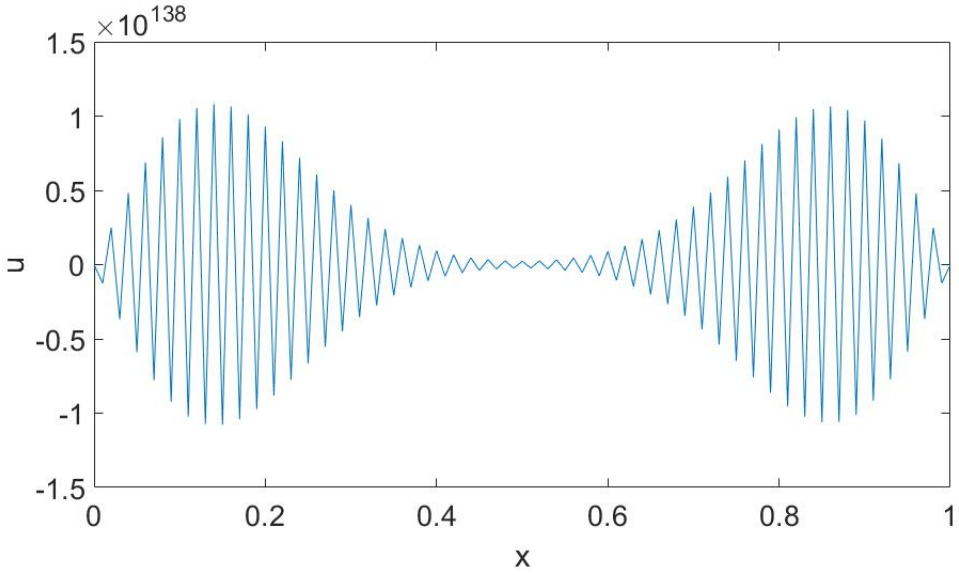

FIGURE 8 – Instabilité de la méthode numérique explicite pour $a = 1$, $b = -10$, $D = 10$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.003$.

Cependant, on constate que la méthode implicite est elle stable pour ces paramètres (voir Figure 9).

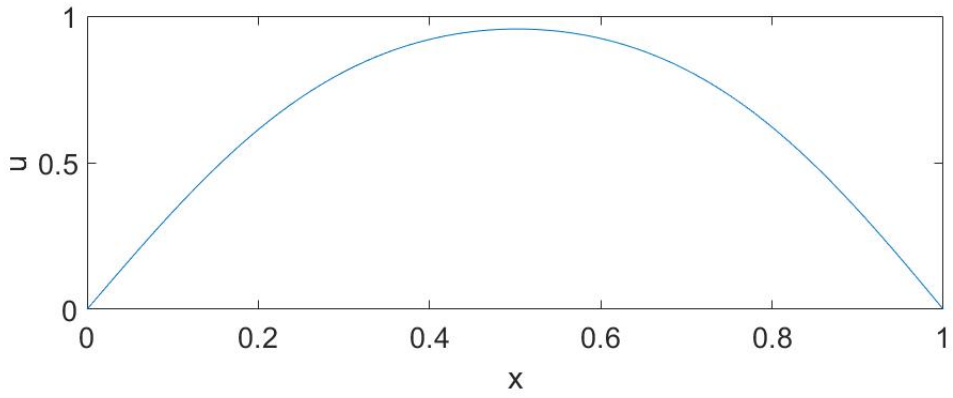

FIGURE 9 – Stabilité de la méthode numérique implicite pour $a = 1$, $b = -10$, $D = 10$, $h = 0.01$ et $k = 0.00001$, au temps $t = 0.003$.

Par contre, confirmant notre réponse à la question 1.4 (voir Section 1.3), la méthode implicite n'est en effet pas toujours stable comme le montre la Figure 10.

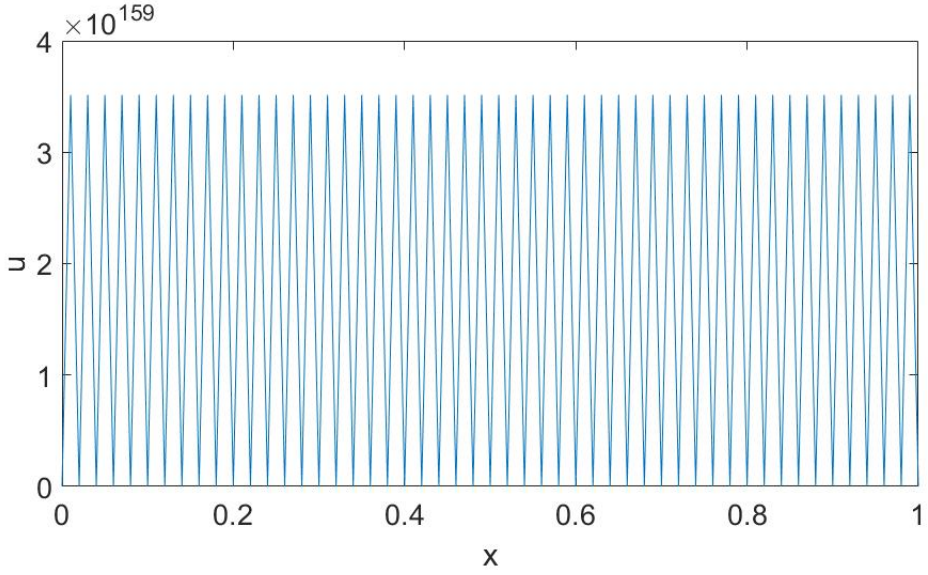

FIGURE 10 – Instabilité de la méthode numérique implicite pour $\lambda_d = \lambda_a/2$, $\lambda_b = -1$ et $\lambda_a < 1$.

1.5 Montrez, dans ces schéma numériques, quels termes sont responsables de la diffusion numérique. Quel est son ordre de grandeur par rapport à la diffusion explicitement modélisée par le terme D ?

Schéma explicite

Afin de déterminer les termes responsables d'une éventuelle diffusion numérique, calculons l'erreur de troncature $\tau_{i,j}$. Pour rappel notre schéma numérique explicite a la forme :

$$U_{i,j+1} = U_{i+1,j}(\lambda_d - \lambda_a) + U_{i,j}(-2\lambda_d + \lambda_a - \lambda_b + 1) + U_{i-1,j}\lambda_d + \mathcal{O}(k+h) \ . \quad (78)$$

Ou encore,

$$u(x_i, t_j + k) = Au(x_i + h, t_j) + Bu(x_i, t_j) + Cu(x_i - h, t_j) + k\tau_{i,j} \ . \quad (79)$$

Dès lors l'erreur de troncature $\tau_{i,j}$ peut-être exprimée comme :

$$\tau_{i,j} = \frac{1}{k}\left(u(x_i, t_j + k) - Au(x_i + h, t_j) - Bu(x_i, t_j) - Cu(x_i - h, t_j)\right) \ . \quad (80)$$

On peut développer en séries $u(x_i, t_j + k)$, $u(x_i + h, t_j)$, $u(x_i, t_j)$ et $u(x_i - h, t_j)$, en gardant uniquement les termes jusqu'à l'ordre k de sorte que on ne considère que les termes dominant de l'erreur. C'est à dire qu'on calcule $k\tau_{i,j}$ à l'ordre $\mathcal{O}(k)$.

$$u(x_i, t_j + k) = u(x_i, t_j) + k\frac{\partial u(x_i, t_j)}{\partial t} + \mathcal{O}(k^2) \quad (81)$$

Par définition de l'équation d'advection-diffusion (Equation 1), on a :

$$\frac{\partial u(x_i, t_j)}{\partial t} = D\frac{\partial^2 u(x_i, t_j)}{\partial x^2} - a\frac{\partial u(x_i, t_j)}{\partial x} - bu(x_i, t_j) \ . \quad (82)$$

Dès lors,

$$u(x_i, t_j + k) = u(x_i, t_j) + k\left(D\frac{\partial^2 u(x_i,t_j)}{\partial x^2} - a\frac{\partial u(x_i,t_j)}{\partial x} - bu(x_i,t_j)\right) + \mathcal{O}(k^2)\ . \tag{83}$$

Par ailleurs,

$$u(x_i \pm h, t_j) = u(x_i, t_j) \pm h\frac{\partial u(x_i,t_j)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x_i,t_j)}{\partial x^2} \pm \frac{h^3}{6}\frac{\partial^3 u(x_i,t_j)}{\partial x^3} + \mathcal{O}(h^4)\ . \tag{84}$$

En injectant les Equations 83 et 84 dans l'expression de l'erreur de troncature (Equation 80) on obtient :

$$\begin{aligned}\tau_{i,j} = \frac{1}{k}\Bigg(&u(x_i,t_j)(1 - bk - A - B - C) + \frac{\partial u(x_i,t_j)}{\partial x}(-ak - Ah + Ch)\\ &+\frac{\partial^2 u(x_i,t_j)}{\partial x^2}\left(Dk - \frac{Ah^2}{2} - \frac{Ch^2}{2}\right) + \frac{\partial^3 u(x_i,t_j)}{\partial x^3}\left(-\frac{Ah^3}{6} + \frac{Ch^3}{6}\right)\Bigg)\ .\end{aligned} \tag{85}$$

Dans notre schéma explicite on a :

$$A = \lambda_d - \lambda_a = \frac{Dk}{h^2} - \frac{ak}{h} \tag{86}$$

$$B = -2\lambda_d + \lambda_a - \lambda_b + 1 = -2\frac{Dk}{h^2} + \frac{ak}{h} - bk + 1 \tag{87}$$

$$C = \lambda_d = \frac{Dk}{h^2}\ . \tag{88}$$

En injectant ces valeurs de A, B et C dans l'Equation 85 on obtient :

$$\begin{aligned}\tau_{i,j} = \frac{1}{k}&\left(\frac{\lambda_a h^2}{2}\frac{\partial^2 u(x_i,t_j)}{\partial x^2} + \frac{\lambda_a h^3}{6}\frac{\partial^3 u(x_i,t_j)}{\partial x^3}\right)\\ &= \frac{ah}{2}\left(\frac{\partial^2 u(x_i,t_j)}{\partial x^2} + \frac{h}{3}\frac{\partial^3 u(x_i,t_j)}{\partial x^3}\right)\ .\end{aligned} \tag{89}$$

On voit donc qu'on a bien un effet de diffusion numérique lié au terme

$$\frac{ah}{2}\frac{\partial^2 u(x_i,t_j)}{\partial x^2} \tag{90}$$

de l'erreur de troncature. On voit en effet que ce terme contient une dérivée seconde de u par rapport à la variable spatiale x, tout comme le terme de diffusion explicitement modélisée

$$D\frac{\partial^2 u(x_i,t_j)}{\partial x^2}\ . \tag{91}$$

La constante a ainsi que le pas d'espace h vont influencer l'effet de diffusion numérique. Si on regarde l'ordre de grandeur relatif entre la diffusion numérique et la diffusion modélisée explicitement on a :

$$\frac{\lambda_a h^2/2}{\lambda_d} = \frac{ah^3}{2D}\ . \tag{92}$$

Lorsque $a < 0$, le terme de diffusion numérique (Equation 90) a pour d'effet de diminuer l'amplitude de la solution par rapport à la solution analytique (voir Figure 11).

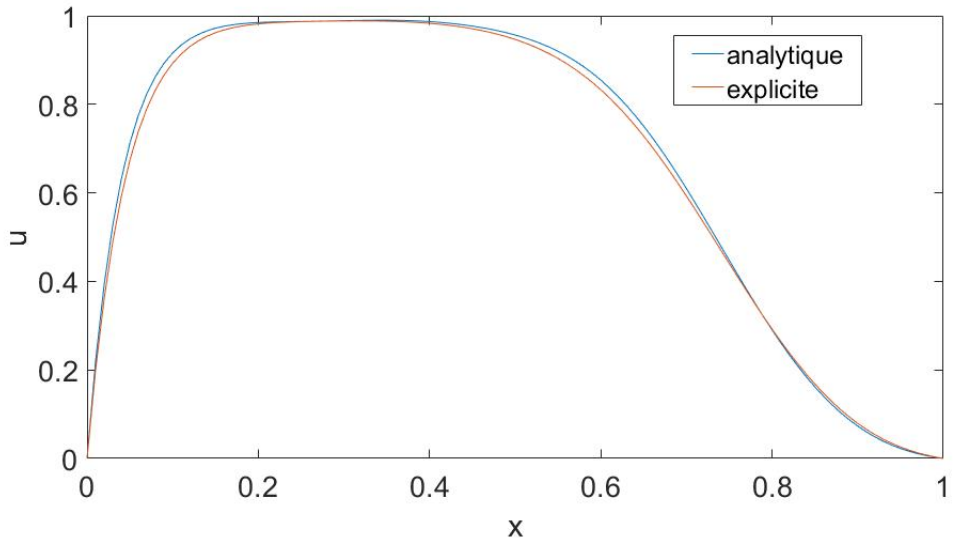

FIGURE 11 – Diffusion numérique du schéma explcite avec $a = -25$, $b = 1$, $D = 1$, $l = 1$, $h = 0.01$, $k = 0.0000001$ à t=0.01.

Lorsque $a > 0$, le terme de diffusion numérique (Equation 90) a pour d'effet d'augmenter l'amplitude de la solution par rapport à la solution analytique (voir Figure 12).

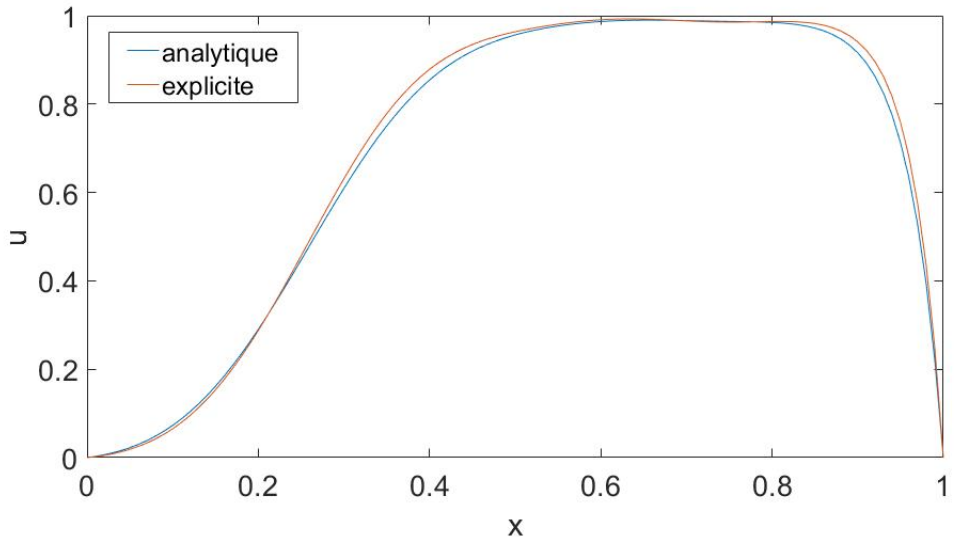

FIGURE 12 – Diffusion numérique du schéma explcite avec $a = 25$, $b = 1$, $D = 1$, $l = 1$, $h = 0.01$, $k = 0.0000001$ à t=0.01.

On voit que cet effet est moins important lorsque on diminue le pas d'espace h, comme cela est prédit par l'Equation 90 (voir Figure 13).

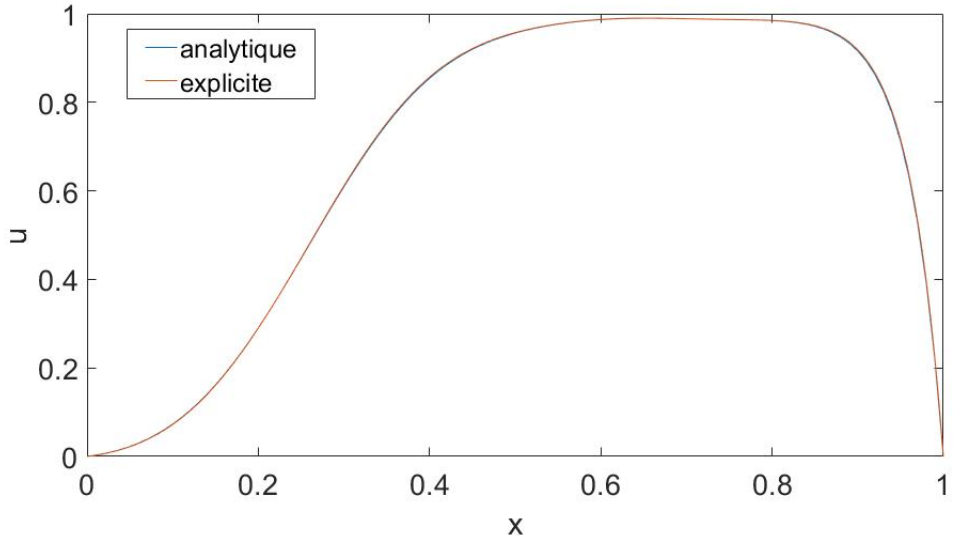

FIGURE 13 – La diffusion numérique du schéma explcite est moins marquée quand on augmente le pas d'espace. $a = 25$, $b = 1$, $D = 1$, $l = 1$, $h = 0.001$, $k = 0.0000001$ à t=0.01.

L'autre terme de l'erreur de troncature à l'ordre considéré, qui fait intervenir une dérivée troisième en x, introduira lui un effet de dispersion.

Schéma implicite

Une démarche similaire peut-être effectuée pour le schéma implicite. Celui-ci a la forme :

$$U_{i,j} = AU_{i+1,j+1} + BU_{i,j+1} + CU_{i\text{-}1,j+1} \ . \tag{93}$$

Ou encore,

$$u(x_i, t_j) = Au(x_i + h, t_j + k) + Bu(x_i, t_j + k) + Cu(x_i - h, t_j + k) + k\tau_{i,j} \ . \tag{94}$$

Dès lors en isolant $\tau_{i,j}$ on a :

$$\tau_{i,j} = \frac{1}{k} \left(u(x_i, t_j) - Au(x_i + h, t_j + k) - Bu(x_i, t_j + k) - Cu(x_i - h, t_j + k) \right) \ . \tag{95}$$

On peut remplacer $u(x_i, t_j)$ par un développement en série :

$$u(x_i, t_j) = u(x_i, t_j + k) - k\frac{\partial u(x_i, t_j + k)}{\partial t} + ... \ . \tag{96}$$

En remplaçant la dérivée par rapport au temps par l'Equation 82 on a :

$$u(x_i, t_j) = u(x_i, t_j+k) - k\left(D\frac{\partial^2 u(x_i, t_j + k)}{\partial x^2} - a\frac{\partial u(x_i, t_j + k)}{\partial x} - bu(x_i, t_j + k)\right) + ... \ . \tag{97}$$

De plus, les termes $u(x_i + h, t_j + k)$ et $u(x_i - h, t_j + k)$ sont remplacés par un développement de Taylor similaire à l'Equation 84. On obtient alors :

$$\tau_{i,j} = \frac{1}{k}\left(u(x_i, t_j + k)(1 + kb - A - B - C) + \frac{\partial u(x_i, t_j + k)}{\partial x}(ka - Ah + Ch)\right.$$
$$\left. + \frac{\partial^2 u(x_i, t_j + k)}{\partial x^2}\left(-kD - \frac{Ah^2}{2} - \frac{Ch^2}{2}\right) + \frac{\partial^3 u(x_i, t_j + k)}{\partial x^3}\left(-\frac{Ah^3}{6} + \frac{Ch^3}{6}\right)\right) . \tag{98}$$

Pour notre schéma implicite on a $A = \lambda_a - \lambda_d$, $B = 1 + 2\lambda_d + \lambda_b - \lambda_a$ et $C = -\lambda_d$. En injectant ces valeurs, les termes en u et en derivée première de u par rapport à x se simplifient. Le terme de troncature devient donc :

$$\tau_{i,j} = -\frac{h^2\lambda_a}{2k}\left(\frac{\partial^2 u(x_i, t_j + k)}{\partial x^2} + \frac{h}{3}\frac{\partial^3 u(x_i, t_j + k)}{\partial x^3}\right) . \tag{99}$$

On obtient donc un terme de diffusion numérique identique en valeur absolue, mais de signe opposé, à celui obtenu pour la méthode explicite. Son ordre de grandeur en valeur absolue par rapport au terme de diffusion explicitement modélisée est donc le même que pour la méthode explicite.

1.6 Comparez vitesse de groupe numérique avec celle du système original. Le schéma est-il dispersif ?

vitesse de groupe analytique

A partir de la définition de la vitesse de groupe :

$$v_g \equiv \frac{d\omega_r}{dk} \tag{100}$$

où ω_r est la partie réelle de la relation de dispersion (56), on obtient :

$$v_g = a \ . \tag{101}$$

vitesse de groupe numérique

On part d'une expression discrète d'une onde plane :

$$U_{i,j} = e^{i\left(\bar{k}x_i - \omega t_j\right)} \tag{102}$$

où l'on distinguera le nombre d'onde $\bar{k}$ du pas de temps k. On l'injecte dans l'équation aux différences explicite (8) de notre équation différentielle de départ. On obtient :

$$e^{i\left(\bar{k}x_i - \omega t_{j+1}\right)} = e^{i\left(\bar{k}x_{i-1} - \omega t_j\right)}\lambda_d + e^{i\left(\bar{k}x_i - \omega t_j\right)}(-2\lambda_d + \lambda_a - \lambda_b + 1) + e^{i\left(\bar{k}x_{i+1} - \omega t_j\right)}(\lambda_d - \lambda_a) \ . \tag{103}$$

En tenant compte du pas d'espace $x_{i+1} - x_i = h$ et du pas de temps $t_{j+1} - t_j = k$, après calcul on a :

$$e^{-i\omega k} = e^{-i\bar{k}h}\lambda_d + (-2\lambda_d + \lambda_a - \lambda_b + 1) + e^{i\bar{k}h}(\lambda_d - \lambda_a) \ . \tag{104}$$

Nous allons faire une approximation par le théorème de Taylor à l'ordre $\mathcal{O}(h + k)$, en tenant compte que $\bar{k}$ et les λ sont fixés, nous obtenons :

$$
\begin{aligned}
1 - i\omega k &= (1 - i\bar{k}h)\lambda_d + (-2\lambda_d + \lambda_a - \lambda_b + 1) + (1 + i\bar{k}h)(\lambda_d - \lambda_a) \\
&= \lambda_d - i\bar{k}h\lambda_d - 2\lambda_d + \lambda_a - \lambda_b + 1 + \lambda_d + i\bar{k}h\lambda_d - \lambda_a - i\bar{k}h\lambda_a \ .
\end{aligned}
$$

Après simplification et en remplaçant λ_a par sa valeur $\lambda_a = \frac{ak}{h}$, on obtient :

$$
\begin{aligned}
\omega &= \frac{\bar{k}h}{k}\lambda_a && (105) \\
&= a\bar{k} \ . && (106)
\end{aligned}
$$

Nous calculons la vitesse de groupe numérique par différence avant, on obtient :

$$
\begin{aligned}
v_{gn} &= \frac{d\omega}{d\bar{k}} && (107) \\
&= a\frac{d\bar{k}}{d\bar{k}} && (108) \\
&= a \ . && (109)
\end{aligned}
$$

Cette approximation à l'ordre $\mathcal{O}(h + k)$, nous permet de retrouver la vitesse de groupe analytique.

Schéma non dispersif au premier ordre

La vitesse de groupe numérique n'est pas fonction de $\bar{k}$, elle est égale à une constante a dans notre approximation. Notre schéma n'est donc pas dispersif au premier ordre. Ce qui est cohérent avec notre système originel qui est non dispersif $v_g = a$.

1.7 Le terme de diffusion D peut-il être responsable d'un comportement instable ? Expliquer.

Oui car le terme $\lambda_D = Dk/h^2$ apparait dans les conditions de A-stabilité des schémas numériques explicites et implicites.

Pour le schéma explicite, en regardant l'Equation 16, on voit que plus on augmente la valeur de D (et donc de λ_D) plus on tend avec certitude vers une situation où l'inégalité (et donc la condition de stabilité) n'est pas satisfaite. En effet, alors que pour une certaine combinaison des paramètres , dont $D = 5$, la méthode explicite présente un comportement stable (voir Figure 14), elle ne l'est pas si on augmente la valeur de D à 6 (voir Figure 15).

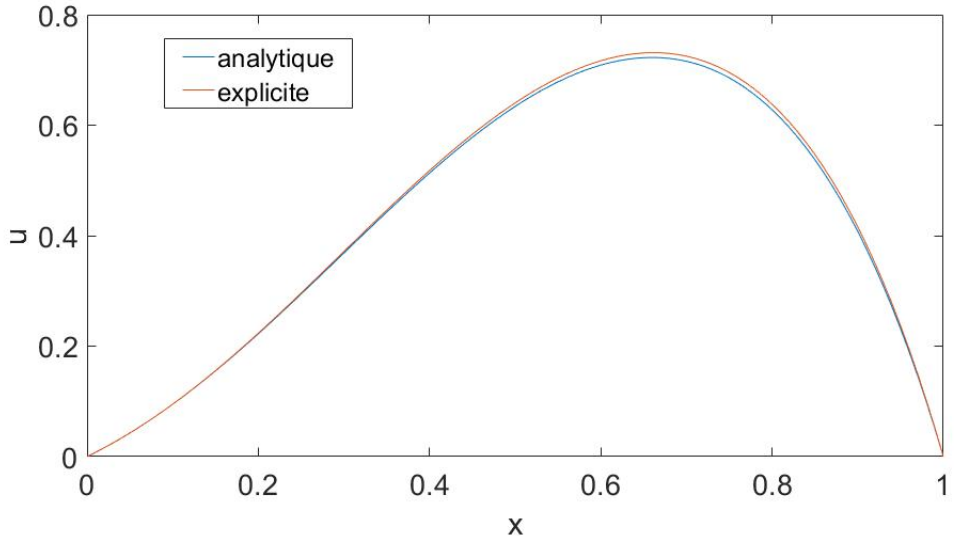

FIGURE 14 – Méthode explicite stable pour $a = 25$, $b = 5$, $D = 5$, $l = 1$, $h = 0.01$, $k = 0.00001$ à t=0.01.

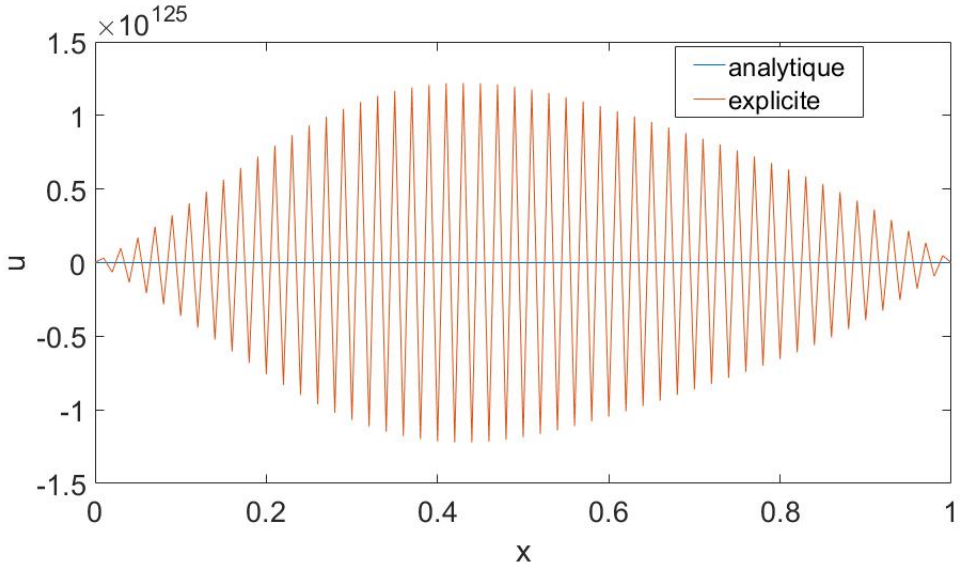

FIGURE 15 – Méthode explicite instable pour $a = 25$, $b = 5$, $D = 6$, $l = 1$, $h = 0.01$, $k = 0.00001$ à t=0.01.

Pour le schéma implicite, varier D et donc le terme λ_d peut également engendrer non-satisfaction de la condition de stabilité. Cependant il y a, contraitement au cas du schéma explicite, des combinaisons de paramètres pour lesquelles le schéma sera stable pour toutes les valeurs de D possibles.

1.8 Fournissez un schéma de Lax-Wendroff à trois points $(\boldsymbol{i}-1, \boldsymbol{i}, \boldsymbol{i}+1)$ dans l'espace. Quel est l'ordre de ce schéma ? Est-il monotone ? Discutez les avantages et les inconvénients au schéma que vous avez donné au point 1 ci-dessus.

Ce schéma de Lax-Wendroff peut-être obtenu en exigeant que l'erreur de troncature $k\tau_{i,j}$ de l'Equation 85 tende vers 0 lorsque $h \to 0$. L'Equation 85 de l'erreur de troncature du schéma explicite peut se réécrire comme :

$$k\tau_{i,j} = u(x_i,t_j)(1-\lambda_b - A - B - C) + \frac{\partial u(x_i,t_j)}{\partial x} h(-\lambda_a - A + C) + \frac{\partial^2 u(x_i,t_j)}{\partial x^2}\frac{h^2}{2}(2\lambda_d - A - C) + \frac{\partial^3 u(x_i,t_j)}{\partial x^3}\frac{h^3}{6}(C-A) + \mathcal{O}(k^2+h^4)\,. \tag{110}$$

Pour que $k\tau_{i,j} \to 0$ quand $h \to 0$ on doit avoir :

$$1 - \lambda_b - A - B - C = 0 \tag{111}$$
$$-\lambda_a + C - A = 0 \tag{112}$$
$$2\lambda_d - A - C = 0\,. \tag{113}$$

En résolvant ce système linéaire on obtient :

$$A = -\frac{\lambda_a}{2} + \lambda_d \tag{114}$$
$$B = 1 - \lambda_b - 2\lambda_d \tag{115}$$
$$C = \frac{\lambda_a}{2} + \lambda_d\,. \tag{116}$$

Le schéma numérique est donc :

$$U_{i,j+1} = \left(-\frac{\lambda_a}{2} + \lambda_d\right) U_{i+1,j} + (1 - \lambda_b - 2\lambda_d) U_{i,j} + \left(\frac{\lambda_a}{2} + \lambda_d\right) U_{i\text{-}1,j}\,. \tag{117}$$

L'ordre du schéma est obtenu en analysant l'élement restant du terme de troncature, qui ne s'annule pas par nos choix de A, B et C :

$$k\tau_{i,j} = \frac{\partial^3 u(x_i,t_j)}{\partial x^3}\frac{h^3}{6}(C-A) + \mathcal{O}(k^2+h^4) \tag{118}$$

$$\mathcal{O}(k\tau_{i,j}) = \mathcal{O}\left(\frac{h^3}{6}\left(\frac{\lambda_a}{2} + \lambda_d + \frac{\lambda_a}{2} - \lambda_d\right)\right) = \mathcal{O}\left(h^3\lambda_a\right) = \mathcal{O}\left(h^2 k\right) + \mathcal{O}(k^2 + h^4) \tag{119}$$

$$\mathcal{O}(\tau_{i,j}) = \mathcal{O}\left(h^2 + k\right) \ . \tag{120}$$

Un schéma aux différences finies du type

$$U_{i,j+1} = \sum_p A_p U_{p,j} \tag{121}$$

est monotone si :

$$A_p \geq 0, \forall p \ . \tag{122}$$

En comparant ce théorême avec le schéma numérique de Lax-Wendroff (présenté à l'Equation 117), on constate que ce schéma est monotone si les trois conditions $A \geq 0, B \geq 0$ et $C \geq 0$ sont remplies. C'est à dire si :

$$\lambda_d \geq \frac{\lambda_a}{2} \tag{123}$$

$$\lambda_d \leq \frac{1 - \lambda_b}{2} \tag{124}$$

$$\lambda_d \geq -\frac{\lambda_a}{2} \ . \tag{125}$$

Le premier avantage du schéma de Lax-Wendroff par rapport aux schémas implicite et explicite introduits dans les questions précédentes est que son erreur est moins importante ($\mathcal{O}\left(h^2 + k^2\right)$ au lieu de $\mathcal{O}\left(h + k\right)$). De plus, par choix des coefficients A, B et C on a éliminé le terme qui était responsable de la diffusion numérique (le terme en dérivée seconde de u par rapport à x dans l'erreur de troncature $\tau_{i,j}$). Le schéma de Lax-Wendroff permet donc d'éviter le problème lié à cette diffusion numérique.

On observe en effet, dans les cas observés, que le schéma Lax-Wendroff correspond mieux à la solution analytique que les schémas implicite et explicite précédents (voir Figures 16 , 2 et 3).

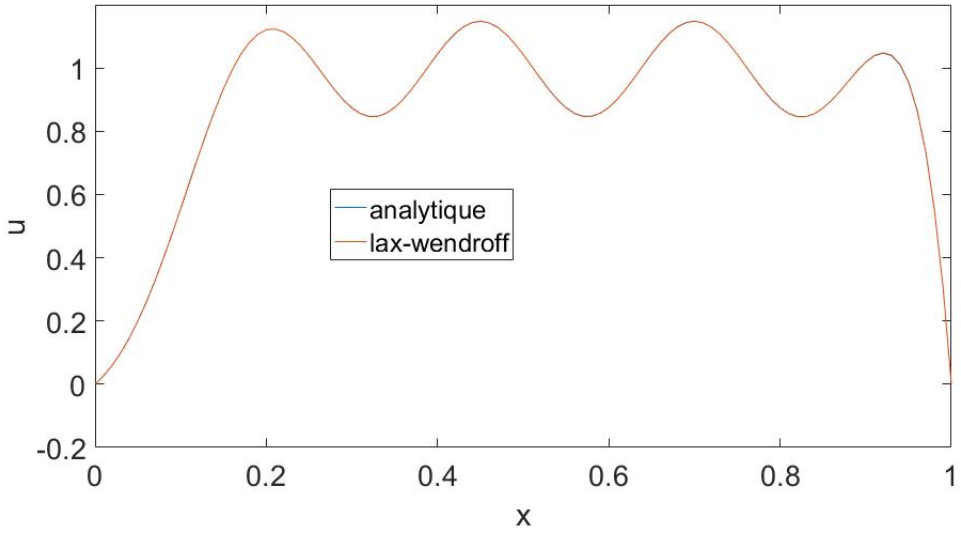

FIGURE 16 – Schéma Lax-Wendroff pour $a = 25$, $b = 1$, $D = 1$, $l = 1$, $h = 0.01$, $k = 0.00001$ à t=0.003. Les courbes numérique et analytique correspondent parfaitement.

Deuxième partie

Prédictibilité

2.1 Essayer avec différentes conditions initiales. Qu'est-ce qui est particulier à la condition initiale choisie dans l'exemple précédent ?

Une fonction exercice1(i,x0) a été implémentée en Matlab (qui utilise des nombres flottants à double précision par défaut, i.e. codés sur 64bits). Celle-ci renvoie le vecteur des $x0, x1, ..., x_i$. La valeur x_i est comprise dans l'intervalle $[0, 1]$. Les différents éléments du vecteur sont calculés à l'aide de la relation de récurrence :

$$x_{i+1} = \begin{cases} 2x_i, \text{ si } x_i \leq 0.5 \\ 2 - 2x_i, \text{ si } x_i > 0.5 \ . \end{cases} \tag{126}$$

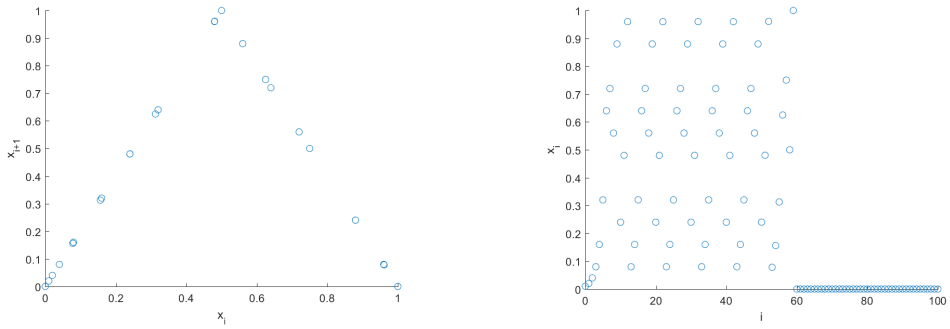

FIGURE 17 – Condition initiale $x_0 = 0.01$.

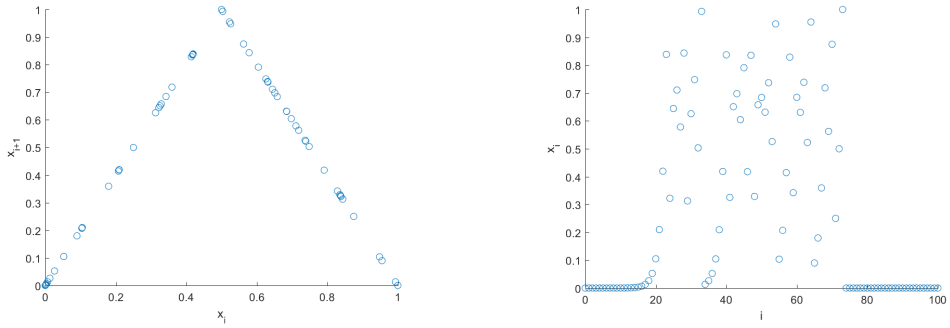

FIGURE 18 – Condition initiale $x_0 = 0.0000001$.

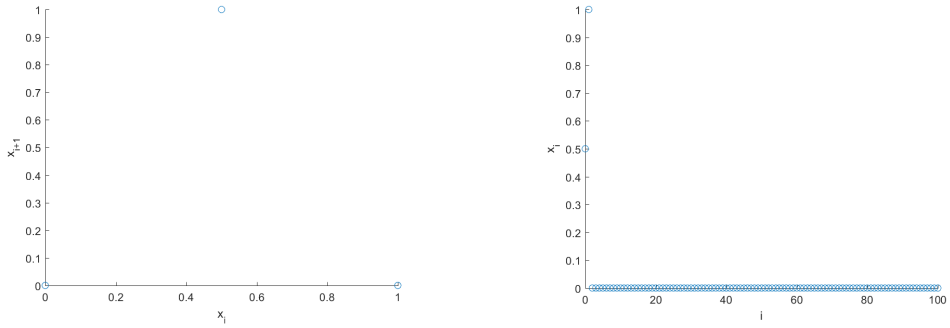

FIGURE 19 – Condition initiale $x_0 = 0.5$.

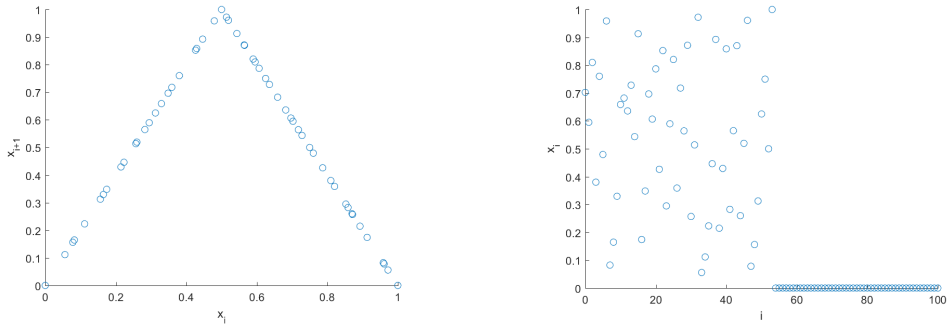

FIGURE 20 – Condition initiale $x_0 = \pi/(2\sqrt{5})$.

On constate qu'une fois l'élément $x_i = 0$ atteint, toutes les valeurs suivantes $x_j, j > i$, sont également égales à zéro. Ceci est facilement compréhensible en regardant la définition de la relation de récurrence : lorsque $x_i = 0$ on est coincé dans le premier cas, $x_i \leq 0.5$.

Par ailleurs, on voit que, pour $x_0 = \pi/(2\sqrt{5})$, les x_i tendent plus vite vers zéro en utilisant des flottants codés sur 64 bits (comme c'est le cas dans ce rapport) que en utilisant des flottants codés sur 128bits (comme c'est le cas dans les slides présentant l'énoncé de ce projet). Ceci est du au fait que avec moins de bits de précision, on arrivera plus rapidement aux limites de précision de chiffres derrière la virgule du nombre flottant. Un résultat de faible amplitude sera dès lors plus rapidement considéré comme étant zéro par l'ordinateur.

La condition initiale $x_0 = \pi/(2\sqrt{5})$ a de particulier que c'est un nombre irrationnel. C'est à dire que son développement décimal ne se termine jamais et ne se répète jamais. Il ne peut donc dans aucun cas être représenté de

manière parfaite dans la mémoire d'un ordinateur qui a elle une capacité finie. Puisque x_0 est irrationnel cela implique que les x_i fournis par la formule de récurrence ont plus de chance d'être différents entre eux. En effet, si il y a peu de décimales, le schéma cyclique de la formule va impliquer que on aura le même x_i pour différents i (voir Figure 21) : $x_0 = 0.1, x_1 = 0.2, x_2 = 0.4, x_3 = 0.8, x_4 = 0.4, x_5 = 0.8, ...$ Au contraire, si il y a beaucoup de décimales (comme c'est le cas pour les nombres irrationnels), on aura de nombreux x_i différents.

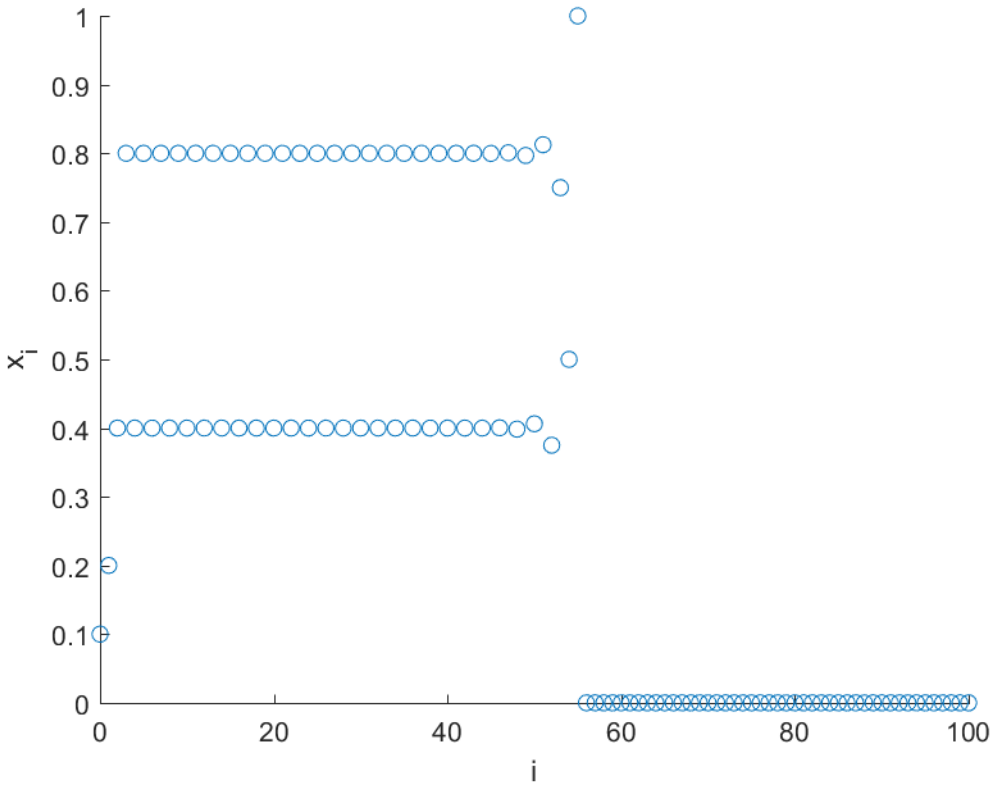

FIGURE 21 – Condition initiale $x_0 = 0.1$.

2.2 Calculer l'itération à la main, en représentation binaire. Que se passe-t-il si on utilise du real(16) ou real(32) ?

La représentation des nombres réels dans la mémoire physique de l'ordinateur dépend en fait du CPU et de son ALU (unité arithmétique et logique). Celle-ci est en effet conçue pour manipuler (additionner, multiplier, diviser, soustraire) des nombres flottants (des réels sur un nombre fini de bits).

Sur la plupart des ordinateurs, les nombres flottants sont représentés en binaire suivant la norme IEEE 754. Dans leur version en simple précision, ces nombres sont codés sur 32bits et sont dans la plupart des langages typés (C, Java, C++,...) désignés par le type float. La représentation en double précision est elle encodée sur 64bits et est habituellement désignée par le type double. Matlab utilise par défaut des nombres flottants en double précision.

Dans le cadre de notre analyse, on considère la représentation simplifiée binaire où le premier bit correspond à la plus grande puissance de 2 du nombre représenté et le dernier à sa plus petite. On représente donc un nombre réel par :

$$x_i = \sum_{i=N}^{N+n-1} a_i 2^{-i} \tag{127}$$

où $-N$ est la plus grande puissance de 2, n est le nombre de bits affectés à la représentation du nombre et les coefficients a_i peuvent prendre la valeur 0 ou 1. Si on applique à cette définition la formule de récurrence (126), on doit traiter deux cas.

Si $x_i \leq 0.5$ **:**

$$x_{i+1} = 2 \sum_{i=N}^{N+n-1} a_i 2^{-i} = \sum_{i=N}^{N+n-1} a_i 2^{-(i-1)} = \sum_{i=N}^{N+n-2} a_{i+1} 2^{-i} + 0 \cdot 2^{N+n-1} \,. \tag{128}$$

On voit donc qu'une multiplication par 2, en représentation binaire consiste en un décalage «shift» vers la gauche de tous les chiffres du nombre de départ. Comme le nombre est codé sur un nombre fini de bits, on rajoute un zéro à la droite de la représentation binaire du nombre.

Si $x_i > 0.5$ **:**

$$x_{i+1} = 2\left(1 - \sum_{i=N}^{N+n-1} a_i 2^{-i}\right) . \tag{129}$$

Où l'opération entre parenthèse consiste à inverser tous les bits après la virgule sauf le dernier 1. Ensuite l'opération de multiplication par 2 est appliquée.

L'opération de multiplication par 2 à chaque itération de notre système dynamique, nous fait perdre un bit de précision. On comprends que la précision de notre modèle dépendra du nombre de bits utilisés pour décrire nos nombres réels.Pour un float16, le nombre de bits de la mantisse est de 10 alors que pour un float32, il est de 23. Il faudra donc 10 (23) itérations en float16 (float32) pour que x_i tende vers zéro. Plus le nombre de bits réservés pour la mantisse est petit plus notre système atteint la valeur zéro rapidement.

2.3 Reproduire l'attracteur de Lorenz et la figure précédente pour $\rho = 5, 10, 28, 45$. Quels sont les cas chaotiques ?

On doit résoudre le système d'équations :

$$\begin{cases} \dfrac{dx}{dt} = \sigma(y - x) \\ \dfrac{dy}{dt} = x(\rho - z) - y \\ \dfrac{dz}{dt} = xy - \beta z \ . \end{cases} \tag{130}$$

Pour se faire on va simplement discrétiser ces équations en remplaçant les dérivées par rapport au temps par une différence finie avant. On obtient alors :

$$\begin{cases} x_{i+1} = \sigma k y_i + x_i(1 - \sigma k) \\ y_{i+1} = x_i k(\rho - z_i) + y_i(1 - k) \\ z_{i+1} = k x_i y_i + z_i(1 - \beta k) \end{cases} \tag{131}$$

où k est le pas de temps. On trouve facilement les différents x_i, y_i, z_i avec une boucle sur le pas de temps et en se donnant une condition initiale x_0, y_0

et z_0. Les graphiques de gauche ci-dessous représentent le mouvement d'une particule suivant ces équations. Les graphiques de droite sont les z_i maximums locaux en fonctions des z_j maximums locaux qui les précèdent. Par exemple, si z_2, z_5 et z_8 sont maximums locaux, on aura les points (z_2, z_5) et (z_5, z_8) dans le graphique.

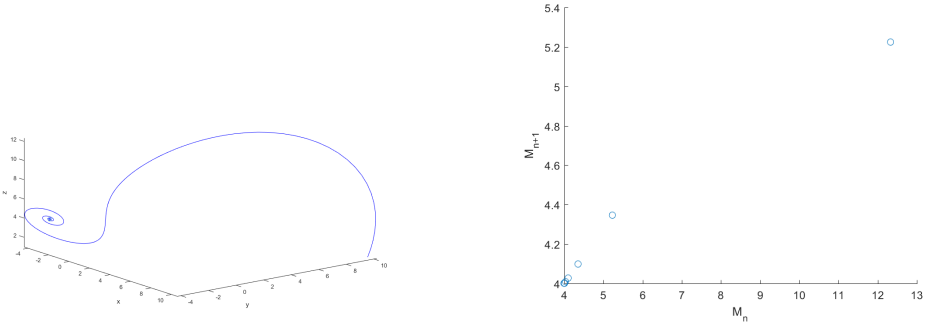

FIGURE 22 – Condition initiale $x_0 = 10, y_0 = 10, z_0 = 1$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 5$, $k = 0.01$ et $t = 50$.

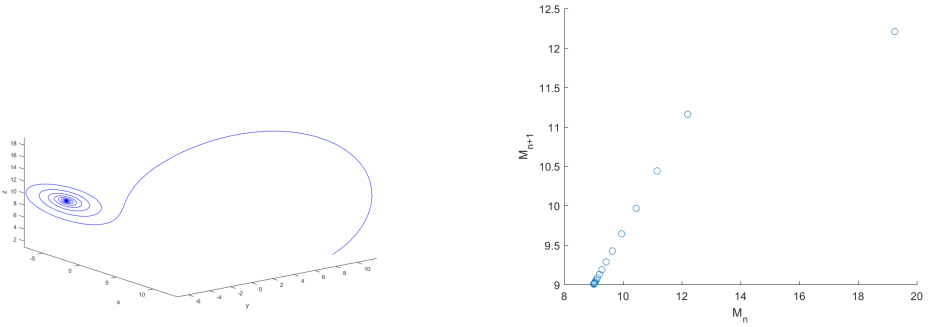

FIGURE 23 – Condition initiale $x_0 = 10, y_0 = 10, z_0 = 1$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 10$, $k = 0.01$ et $t = 50$.

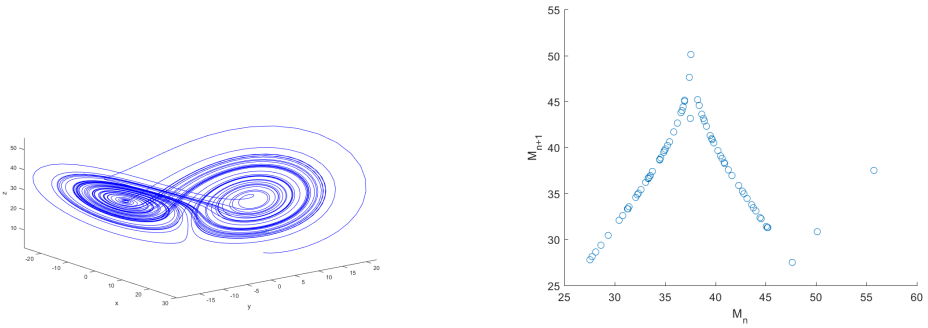

FIGURE 24 – Condition initiale $x_0 = 10, y_0 = 10, z_0 = 1$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 28,\ k = 0.01$ et $t = 50$.

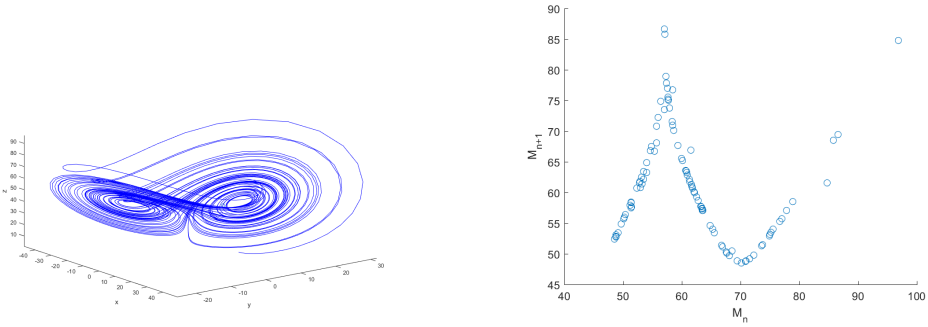

FIGURE 25 – Condition initiale $x_0 = 10, y_0 = 10, z_0 = 1$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 45,\ k = 0.01$ et $t = 50$.

Un système est chaotique si une légère modification de ses conditions initiales engendre une solution très différente aux équations. Et donc dans ce cas des trajectoires fortement différentes pour la particule. Pour déterminer si ces systèmes sont chaotiques analysons les solutions pour une condition initiale légèrement différente ($x_0 = 12, y_0 = 12, z_0 = 3$).

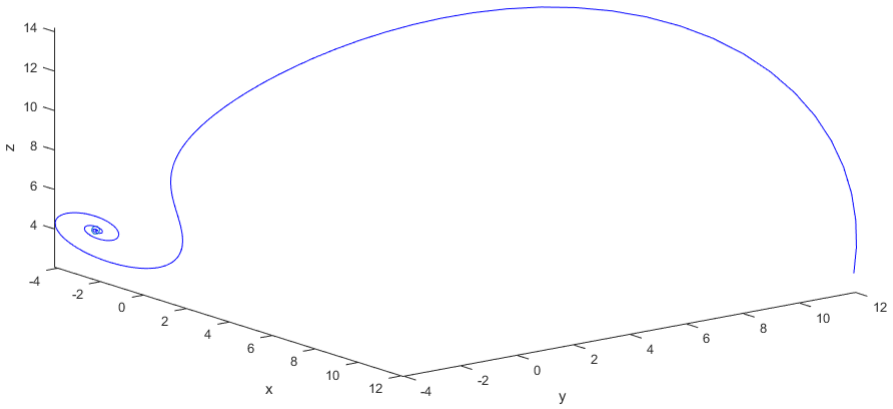

FIGURE 26 – Condition initiale $x_0 = 12, y_0 = 12, z_0 = 3$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 5$, $k = 0.01$ et $t = 50$.

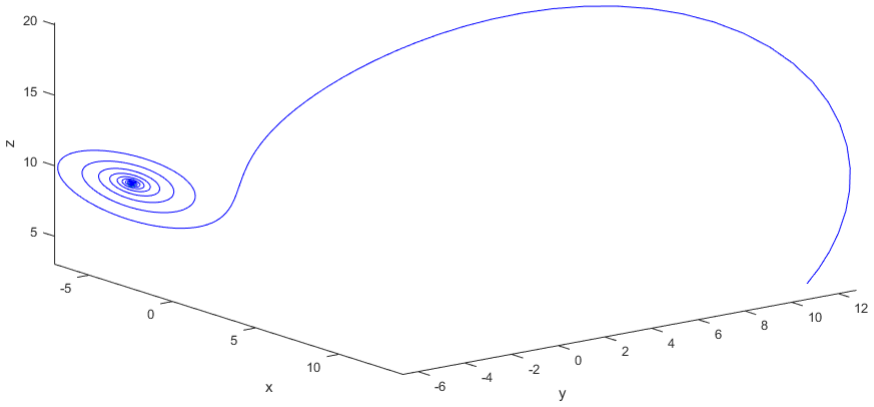

FIGURE 27 – Condition initiale $x_0 = 12, y_0 = 12, z_0 = 3$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 10$, $k = 0.01$ et $t = 50$.

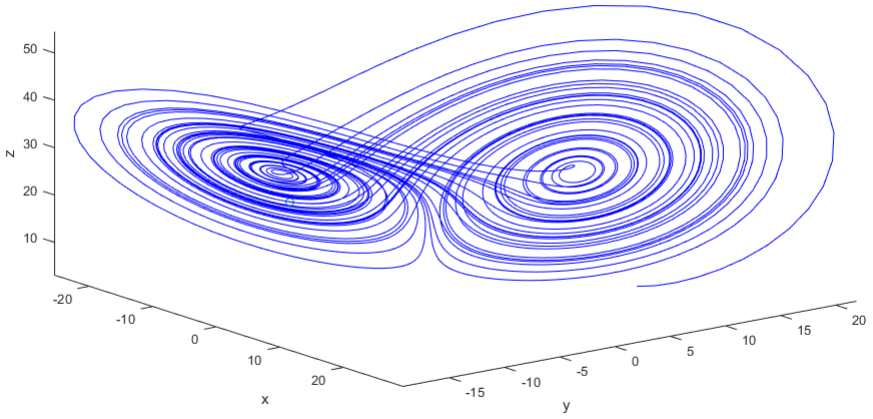

FIGURE 28 – Condition initiale $x_0 = 12, y_0 = 12, z_0 = 3$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 28$, $k = 0.01$ et $t = 50$.

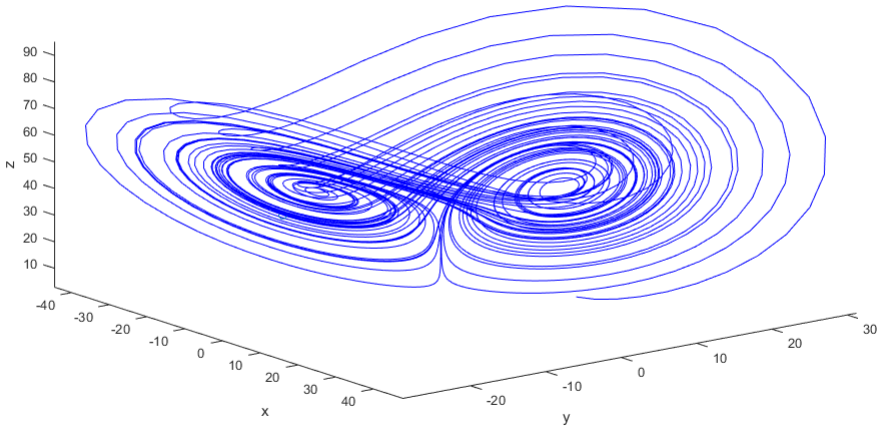

FIGURE 29 – Condition initiale $x_0 = 12, y_0 = 12, z_0 = 3$. Paramètres $\sigma = 10, \beta = 8/3, \rho = 45$, $k = 0.01$ et $t = 50$.

On observe que les cas $\rho = 5$ et $\rho = 10$ sont très proches pour les deux conditions initiales. Les cas $\rho = 28$ et $\rho = 45$ présentent eux un caractère chaotique.

2.4 Quel est le plus grand coefficient de Lyaponov dans le cas de la décroissance radioactive $\dot{x} = -kx$?

Soit l'équation de décroissance radioactive :

$$\dot{x} = -kx \tag{132}$$

$$\frac{dx}{dt} = F(x,t) = -kx \ . \tag{133}$$

L'opérateur linéaire tangent au premier ordre d'une perturbation δx est :

$$\frac{d\delta x}{dt} = J(x,t)\delta x \ . \tag{134}$$

Or $J(x,t)$ est le Jacobien :

$$J(x,t) = \sum_i \frac{\partial F}{\partial x_i} \ . \tag{135}$$

On a donc pour l'équation de la décroissance radioactive :

$$J(x,t) = -k \ . \tag{136}$$

A partir de l'équation (134), on a :

$$\ln(\delta x) = -kt + \text{Constante} \ . \tag{137}$$

On utilise la condition initiale δx_0, ce qui donne :

$$\ln(\delta x) = -kt + \ln(\delta x_0) \ . \tag{138}$$

On a donc :

$$\lambda_{max} = \lim_{t\to\infty} \frac{1}{t} \ln(\|\delta\vec{x}\|) = \lim_{t\to\infty} \frac{1}{t} \ln(\delta x) = \lim_{t\to\infty} \left(\frac{-kt + \ln(\delta x_0)}{t} \right) = -k \ . \tag{139}$$

En effet, puisqu'on est dans un cas à une dimension, $\delta\vec{x} = \delta x$ et $\delta x = \|\delta x\|$ car l'argument du logarithme est positif.

En outre, le système est chaotique si $\lambda_{max} > 0$. Pour la décroissance radioactive, cela signifie que $k < 0$. Dans ce cas, l'équation (138) montre alors qu'une perturbation appliquée au système va croître de manière exponentiel avec le temps.

2.5 Estimer λ_{max} pour $\rho = 5, 10, 28, 45$. Quels sont les cas chaotiques ?

En utilisant fonction implémentée dans le fichier exercice5.m on obtient $\lambda_{max} = -0.9205$ pour $\rho = 5$, $\lambda_{max} = -0.5766$ pour $\rho = 10$, $\lambda_{max} = 0.9194$ pour $\rho = 28$ et $\lambda_{max} = 1.2364$ pour $\rho = 45$.

Un système est chaotique si $\lambda_{max} > 0$. Dès lors on voit que les cas chaotiques sont $\rho = 28$ et $\rho = 45$.