Traitement de la raideur par les méthodes de type Runge-Kutta implicite

S. Descombes [1] T. Dumont [2] V. Louvet [2] **M. Massot** [3]

[1]Laboratoire J.A. Dieudonné, Université de Nice

[2]ICJ - Université Claude Bernard Lyon 1

[3]EM2C - Ecole Centrale Paris

ANGD Informatique Scientifique pour le Calcul - Sète 2008

Plan de la présentation

Plan de la présentation

Brusselator

Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes

$$\left.\begin{array}{rcll} A & \rightarrow & X & \omega = k_1 A, \\ B + X & \rightarrow & Y + D & \omega = k_2 B X, \\ 2X + Y & \rightarrow & 3X & \omega = k_3 X^2 Y, \\ X & \rightarrow & E & \omega = k_4 X, \end{array}\right\}$$

Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Brusselator

Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes

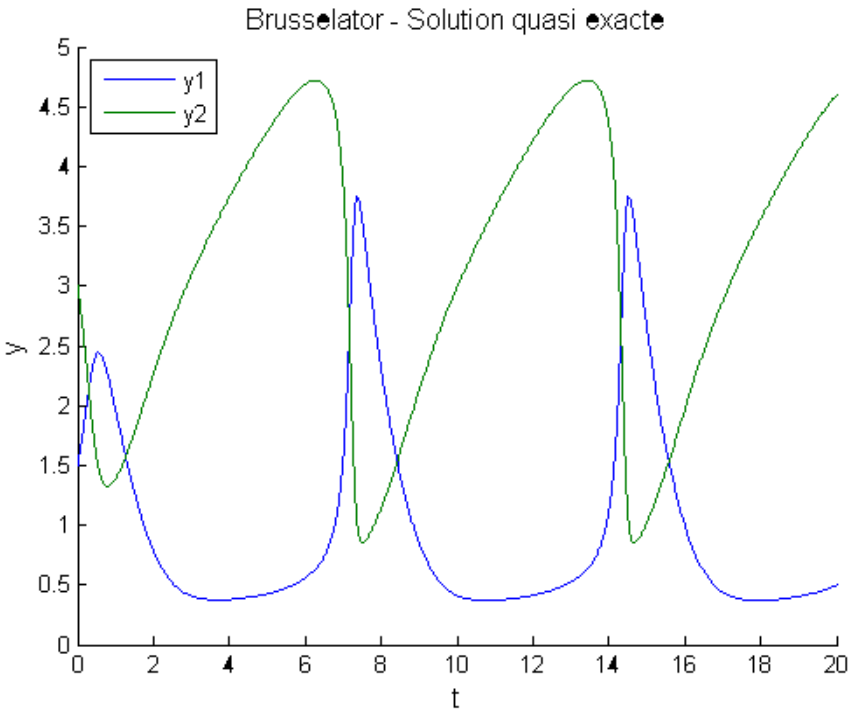
$$x' = A + x^2 y - (B+1)x,$$

$$y' = Bx - x^2 y.$$

Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Brusselator

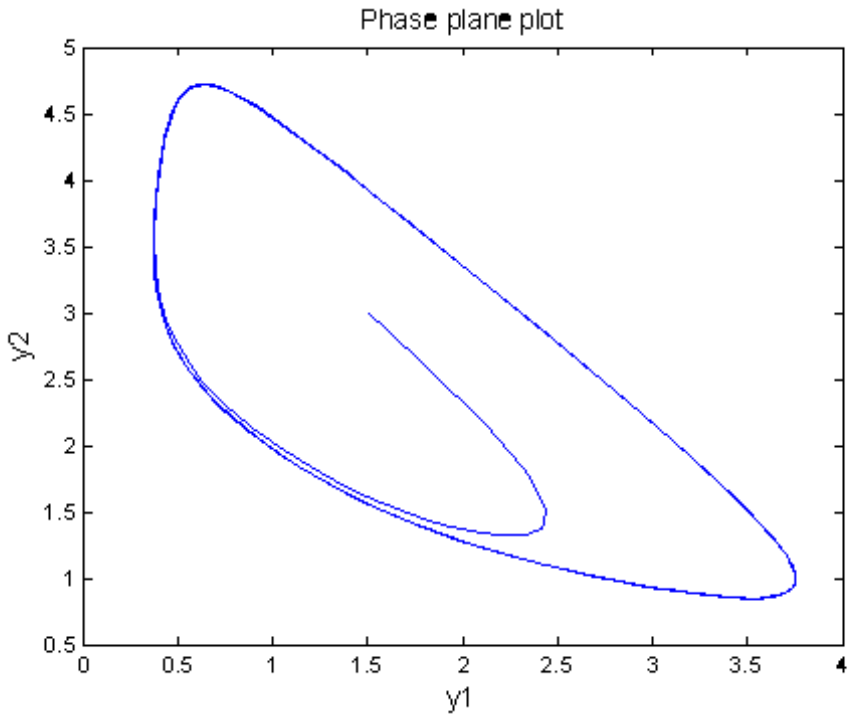
Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes



Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Brusselator

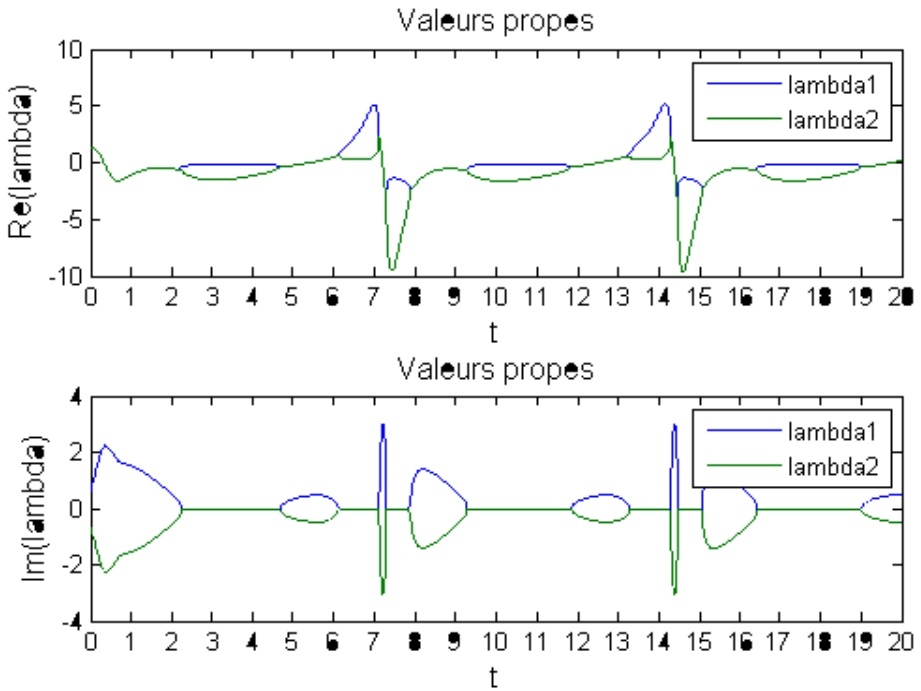
Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes



Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Brusselator

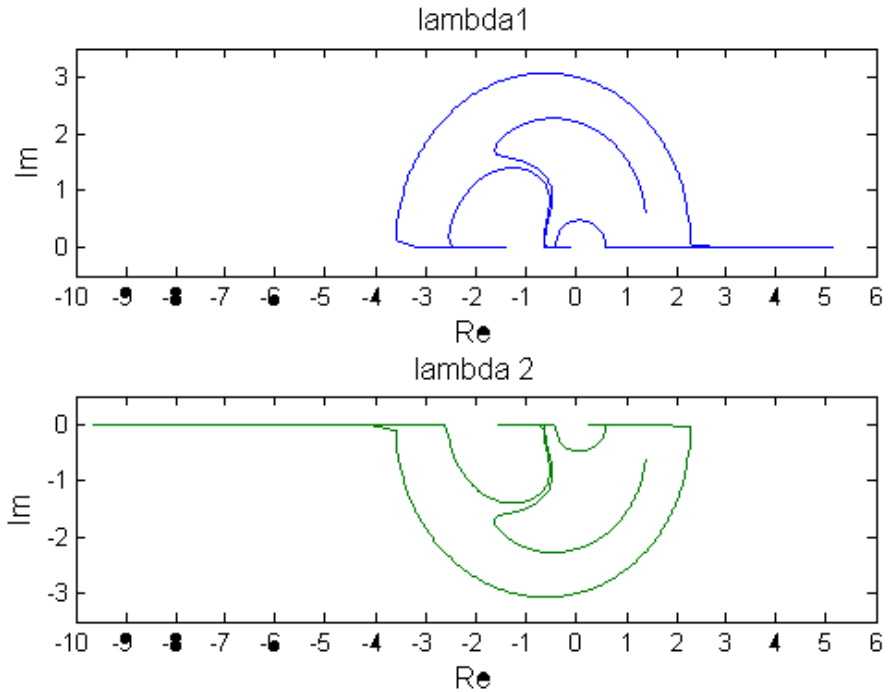
Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes



Variation importante de la valeur propre réelle des valeurs propres et, en particulier, de l'amplitude de la partie réelle négative

Brusselator

Dynamique chimique non-linéaire - un des premiers modèles de réaction chimiques oscillantes



Variation importante de la valeur propre réelle des valeurs propres et, en particulier, de l'amplitude de la partie réelle négative

Oregonator

Dynamique chimique non-linéaire - Modèle un peu plus complexe permettant d'obtenir le chaos chimique pour certaines valeurs des paramètres - plus raide que le Brusselator

$$\left.\begin{array}{rcll} A+Y & \rightarrow & X+P & \omega = k_3 AY, \\ X+Y & \rightarrow & 2P & \omega = k_2 XY, \\ A+X & \rightarrow & 2X+2Z & \omega = k_5 AX, \\ 2X & \rightarrow & A+P & \omega = k_4 X^2, \\ B+Z & \rightarrow & \frac{1}{2} fY & \omega = k_0 BZ, \end{array}\right\}$$

Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Oregonator

Dynamique chimique non-linéaire - Modèle un peu plus complexe permettant d'obtenir le chaos chimique pour certaines valeurs des paramètres - plus raide que le Brusselator

$$\begin{aligned}
\frac{dX}{dt} &= k_3 AY - k_2 XY + k_5 AX - 2k_4 X^2, \\
\frac{dY}{dt} &= -k_3 AY - k_2 XY + \frac{1}{2} f k_0 BZ, \\
\frac{dZ}{dt} &= 2k_5 AX - k_0 BZ.
\end{aligned}$$

Hypothèses : A et B sont maintenues constantes et toutes les constantes de réaction sont prises à un (adimensionnement adéquat).

Oregonator

Dynamique chimique non-linéaire - Modèle un plus complexe permettant d'obtenir le chaos pour certaines valeurs des paramètres - plus raide que le Brusselator

$$b = X/X_0, \qquad a = Y/Y_0, \qquad c = Z/Z_0, \qquad \tau = t/T_0, \tag{1}$$

$$X_0 = k_5 A/2k_4, \quad Y_0 = k_5 A/k_2, \quad Z_0 = (k_5 A)^2/k_4 k_0 B, \quad T_0 = 1/k_0 B,$$

$$\mu \frac{da}{d\tau} = -qa - ab + fc, \tag{2}$$

$$\epsilon \frac{db}{d\tau} = qa - ab + b(1-b), \tag{3}$$

$$\frac{dc}{d\tau} = b - c, \tag{4}$$

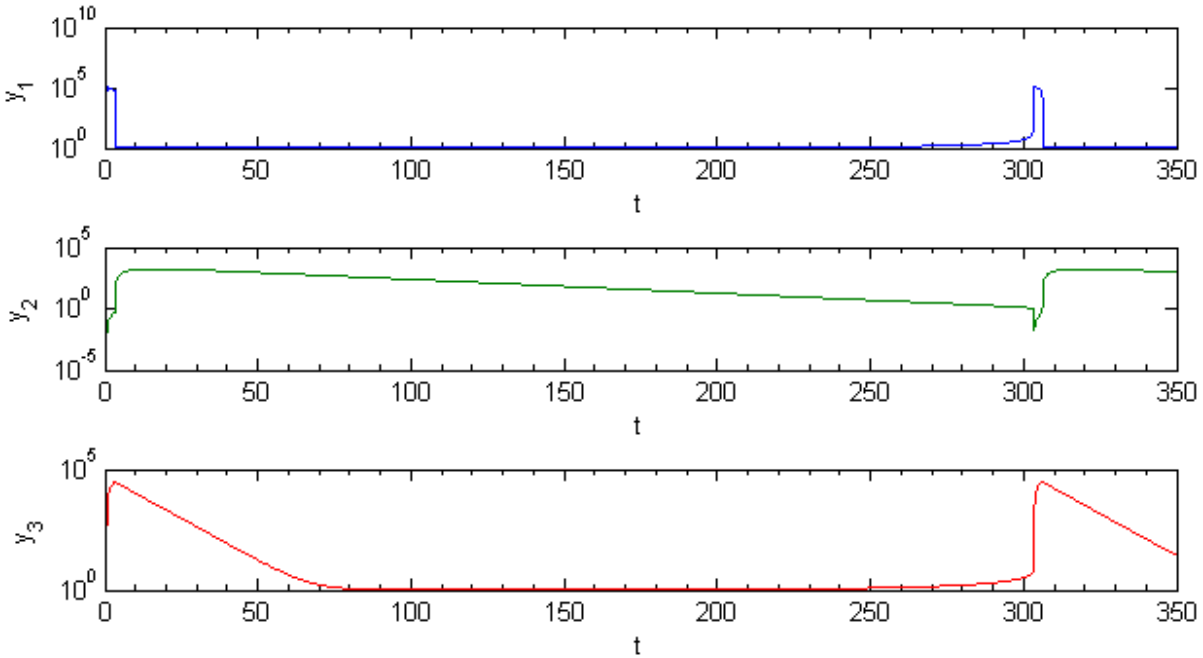
avec les paramètres

$$\epsilon = k_0 B/k_5 A, \qquad \mu = 2k_0 k_4 B/k_2 k_5 A, \qquad q = 2k_3 k_4/k_2 k_5. \tag{5}$$

En général, $\mu \ll \epsilon$ et $q \ll 1$.

Oregonator

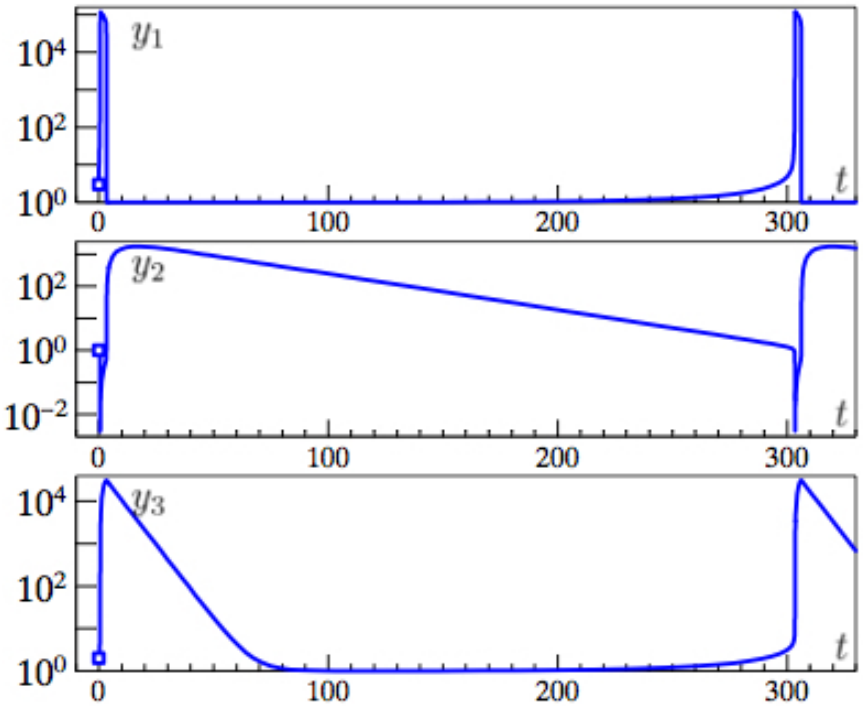
Dynamique chimique non-linéaire - Modèle un plus complexe permettant d'obtenir le chaos pour certaines valeurs des paramètres - plus raide que le Brusselator

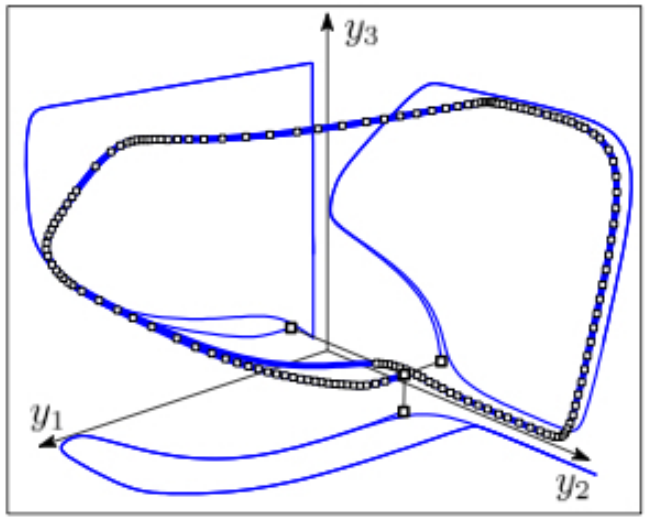


Dynamique très multi-échelles avec de très grosses variations du spectre de valeurs propres

Oregonator

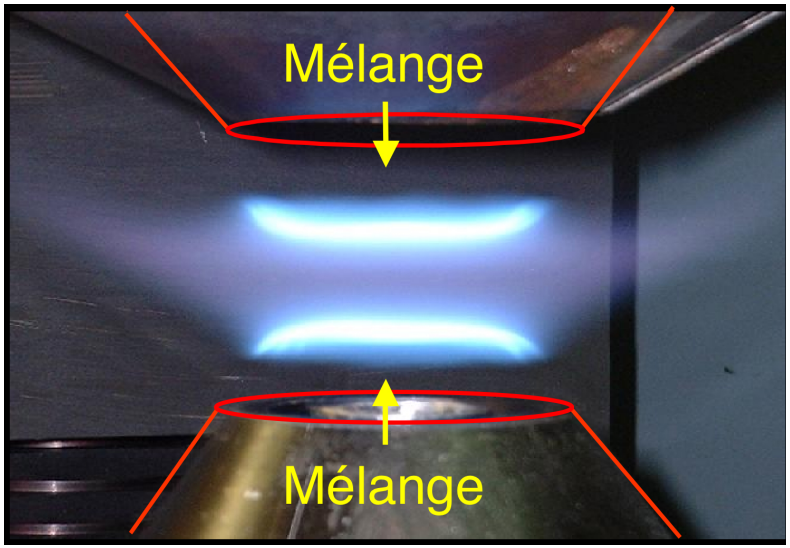
Dynamique chimique non-linéaire - Modèle un plus complexe permettant d'obtenir le chaos pour certaines valeurs des paramètres - plus raide que le Brusselator

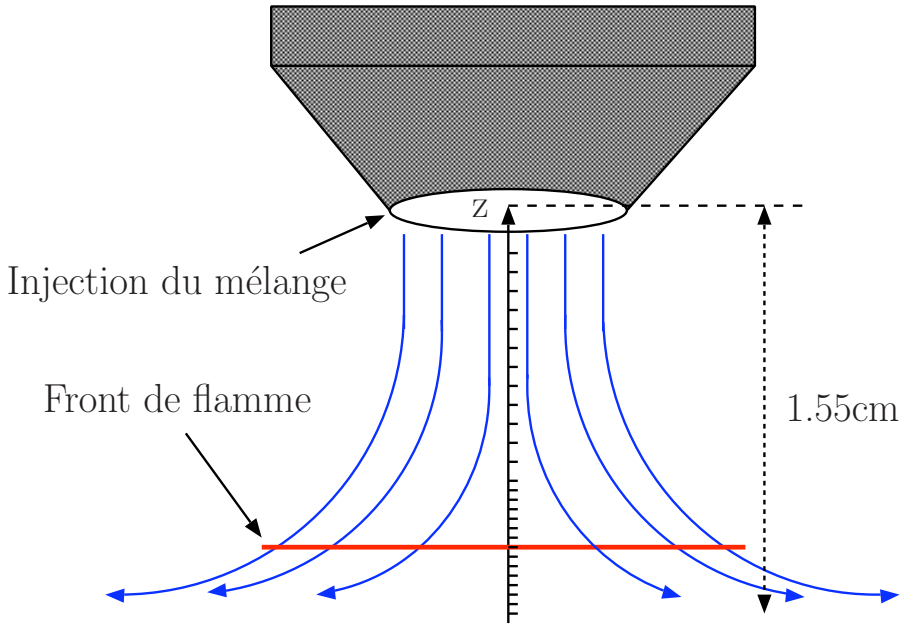




Dynamique très multi-échelles avec de très grosses variations du spectre de valeurs propres

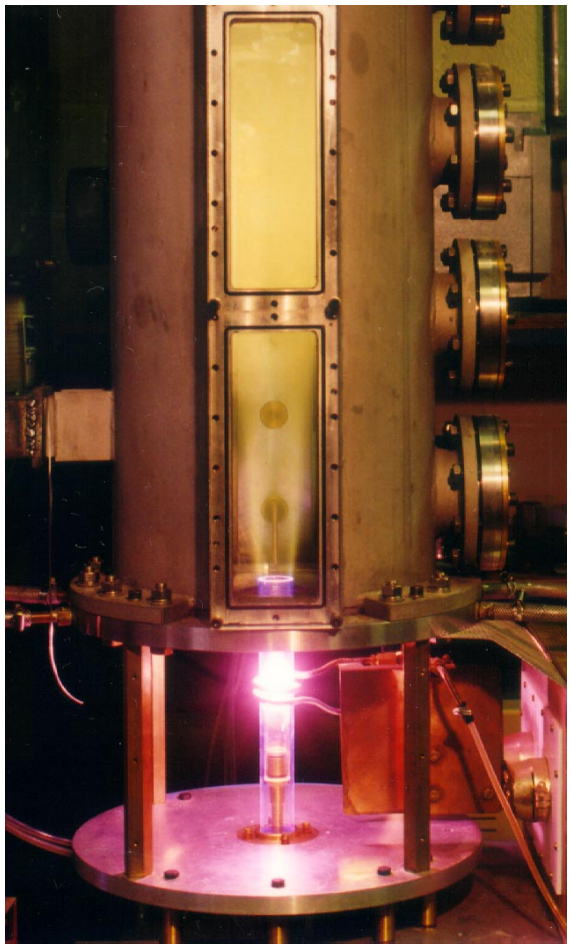
Dynamique d'une flamme de prémélange de méthane à contre-courant pulsée





Modèle avec chimie complexe et transport détaillé (45 species 250 reactions)

Un spectre encore plus large : les plasmas



Torche inductive du von Karman Institute, Bruxelles

Notion de raideur

Raideur : problème rencontré par les méthodes numériques explicites lors de l'intégration numérique du système.
Elle est liée à deux aspects :

- Le spectre de la matrice jacobienne associée au système (notamment, la dispersion des valeurs propres)
- La "distance" de la condition initiale à la "variété d'équilibre" du système

Illustration sur un exemple : Curtiss & Hirschfelder, 1952

Notion de raideur

Raideur : problème rencontré par les méthodes numériques explicites lors de l'intégration numérique du système. Exemple : Curtiss & Hirschfelder, 1952

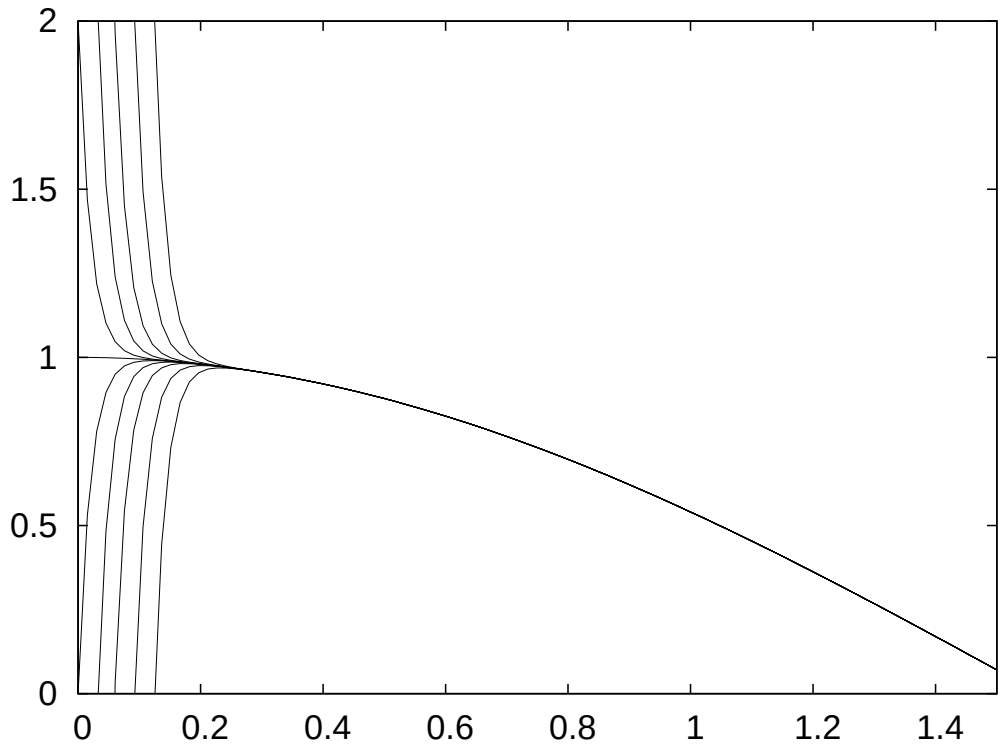
$$y' = \frac{-y + \cos t}{\varepsilon} \quad 0 < \varepsilon \ll 1$$

Solutions :

$$y(t) = Ce^{-t/\varepsilon} + \frac{\cos t}{1+\varepsilon^2} + \frac{\varepsilon \cos t}{1+\varepsilon^2} \quad C \in \mathbb{R}$$

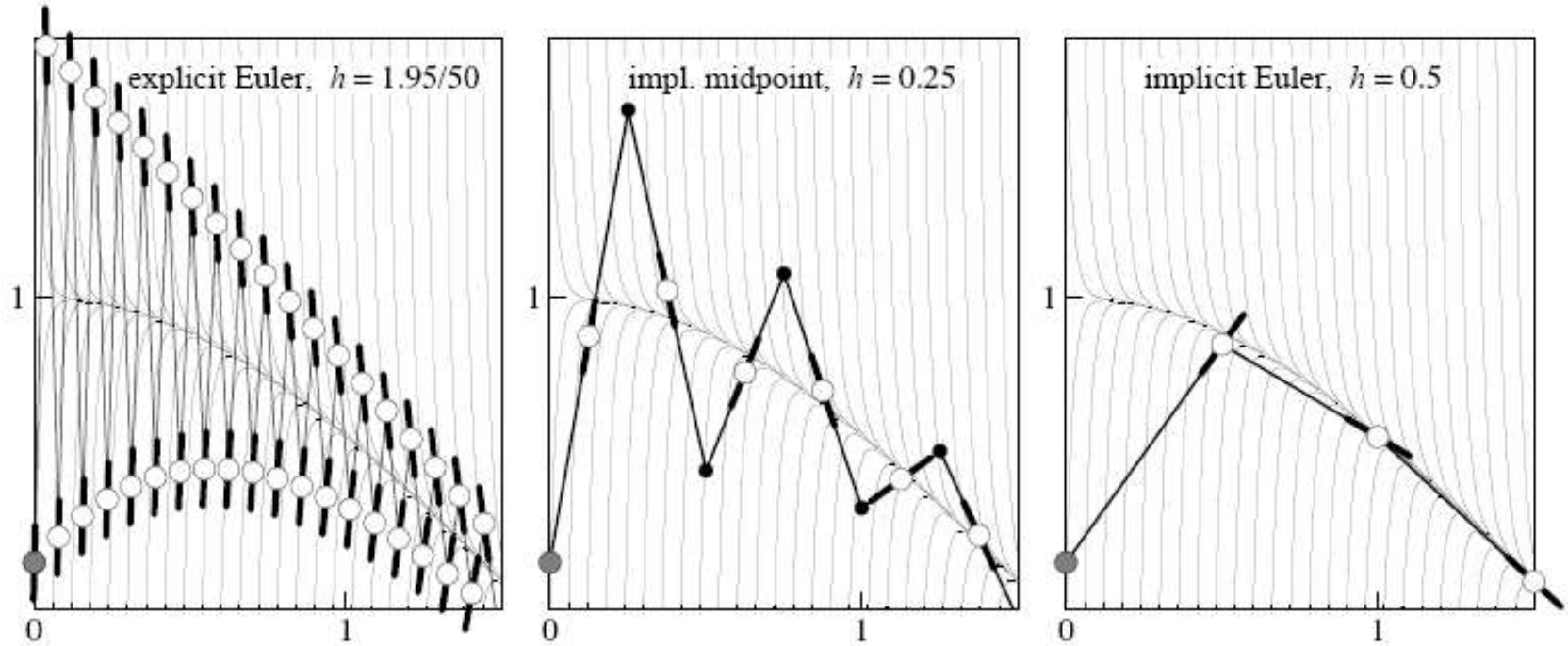
Notion de raideur

Raideur : problème rencontré par les méthodes numériques explicites lors de l'intégration numérique du système. Exemple : Curtiss & Hirschfelder, 1952
Solutions ($\varepsilon = 1/50$) :



Notion de raideur

Raideur : problème rencontré par les méthodes numériques explicites lors de l'intégration numérique du système. Différentes méthodes numériques

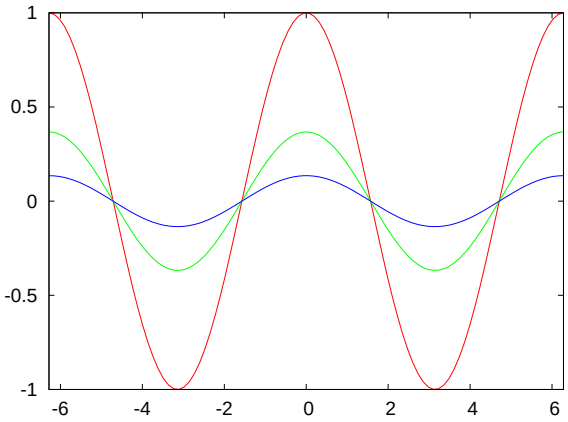


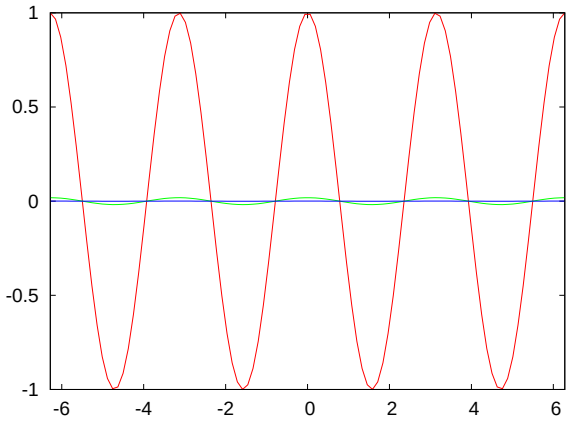
Plan de la présentation

Raideur des systèmes réaction-diffusion

La raideur des systèmes réaction-diffusion peut venir :

- du terme de réaction : différentes gammes d'échelles de temps impliquées, par exemple en combustion échelles rapides dans le terme de réaction
- du terme de diffusion car si le nombre de points de discrétisation est grand, le spectre du laplacien discrétisé est très étalé. Par exemple si dans la donnée initiale, on a une discontinuité, amortissement très rapide des hautes fréquences spatiales et une grande disparité de temps caractéristiques

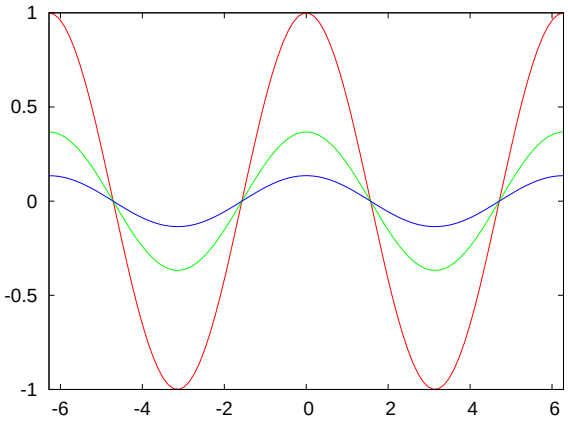


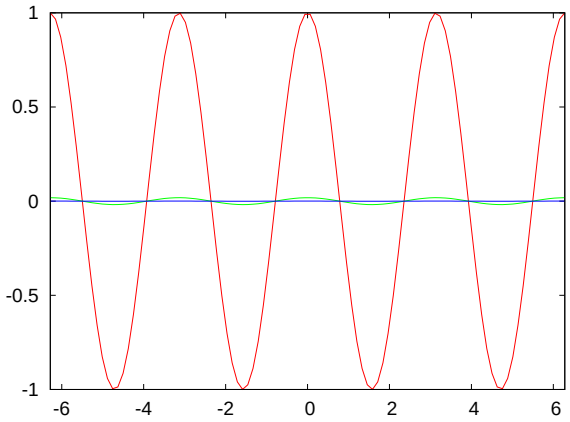


Raideur des systèmes réaction-diffusion

La raideur des systèmes réaction-diffusion peut venir :

- du terme de réaction : différentes gammes d'échelles de temps impliquées, par exemple en combustion échelles rapides dans le terme de réaction
- du terme de diffusion car si le nombre de points de discrétisation est grand, le spectre du laplacien discrétisé est très étalé. Par exemple si dans la donnée initiale, on a une discontinuité, amortissement très rapide des hautes fréquences spatiales et une grande disparité de temps caractéristiques

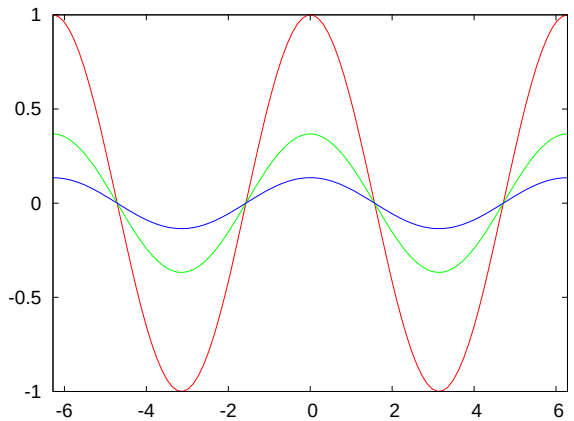


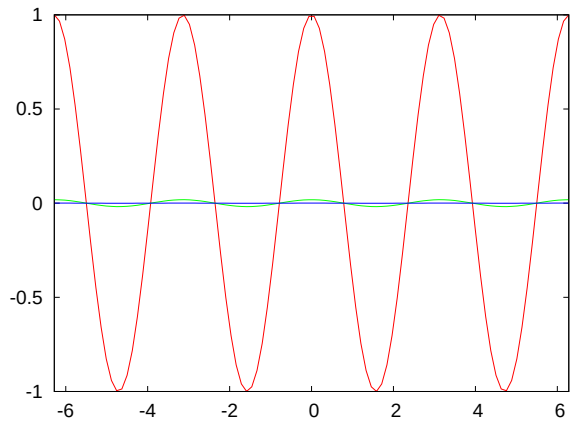


Raideur des systèmes réaction-diffusion

La raideur des systèmes réaction-diffusion peut venir :

- du terme de réaction : différentes gammes d'échelles de temps impliquées, par exemple en combustion échelles rapides dans le terme de réaction
- du terme de diffusion car si le nombre de points de discrétisation est grand, le spectre du laplacien discrétisé est très étalé. Par exemple si dans la donnée initiale, on a une discontinuité, amortissement très rapide des hautes fréquences spatiales et une grande disparité de temps caractéristiques





Donc :

Condition initiale à forts gradients $\Rightarrow$ échelles rapides
$\Rightarrow$ raideur du système

Méthode des lignes (MOL)

Première étape de la résolution numérique d'un système d'équations aux dérivées partielles : **discrétisation spatiale du système** .

- Exemple de système à discrétiser :

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\,\dfrac{\partial^2 u}{\partial x^2} + f(u) \\ \dfrac{\partial u}{\partial t}(a,t) = 0 \\ \dfrac{\partial u}{\partial t}(b,t) = 0 \\ u(x,0) = u_0(x) \end{cases} \qquad x \in [a,b], \quad t \in [0,T]$$

(Condition aux limites de type Neumann)

- **étape 1 :** subdivision spatiale régulière de $[a, b]$, $(x_i)_{i\in[0,n]}$, de pas $h = \frac{b-a}{n}$.

$$U(t) \in \mathbb{R}^{n+1}, \forall i \in [0, n],\ U_i(t) = u(t, x_i)$$

- **étape 2 :**Evaluation des dérivées spatiales par une méthode de différence finies

$$\forall i \in [1, n-1], \frac{\partial^2 u}{\partial x^2}(x_i) \approx \frac{1}{h^2}(U_{i+1} - 2U_i + U_{i-1})$$

- **étape 3 :** Evaluation des conditions aux limites.

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}(x_0) \approx \frac{1}{h^2}(U_1 - U_0) \\ \frac{\partial^2 u}{\partial x^2}(x_n) \approx \frac{1}{h^2}(-U_n + U_{n-1}) \end{cases}$$

- **étape 1 :** subdivision spatiale régulière de $[a, b]$, $(x_i)_{i\in[0,n]}$, de pas $h = \frac{b-a}{n}$.

$$U(t) \in \mathbb{R}^{n+1}, \forall i \in [0, n],\ U_i(t) = u(t, x_i)$$

- **étape 2 :**Evaluation des dérivées spatiales par une méthode de différence finies

$$\forall i \in [1, n-1], \frac{\partial^2 u}{\partial x^2}(x_i) \approx \frac{1}{h^2}(U_{i+1} - 2U_i + U_{i-1})$$

- **étape 3 :** Evaluation des conditions aux limites.

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}(x_0) \approx \frac{1}{h^2}(U_1 - U_0) \\ \frac{\partial^2 u}{\partial x^2}(x_n) \approx \frac{1}{h^2}(-U_n + U_{n-1}) \end{cases}$$

- **étape 1 :** subdivision spatiale régulière de $[a, b]$, $(x_i)_{i\in[0,n]}$, de pas $h = \frac{b-a}{n}$.

$$U(t) \in \mathbb{R}^{n+1}, \forall i \in [0, n],\ U_i(t) = u(t, x_i)$$

- **étape 2 :**Evaluation des dérivées spatiales par une méthode de différence finies

$$\forall i \in [1, n-1], \frac{\partial^2 u}{\partial x^2}(x_i) \approx \frac{1}{h^2}(U_{i+1} - 2U_i + U_{i-1})$$

- **étape 3 :** Evaluation des conditions aux limites.

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}(x_0) \approx \frac{1}{h^2}(U_1 - U_0) \\ \frac{\partial^2 u}{\partial x^2}(x_n) \approx \frac{1}{h^2}(-U_n + U_{n-1}) \end{cases}$$

La méthode des lignes

- U est alors solution de l'EDO :

$$\frac{dU}{dt} = D(\frac{1}{h^2})A_{nh}U + F(U)$$

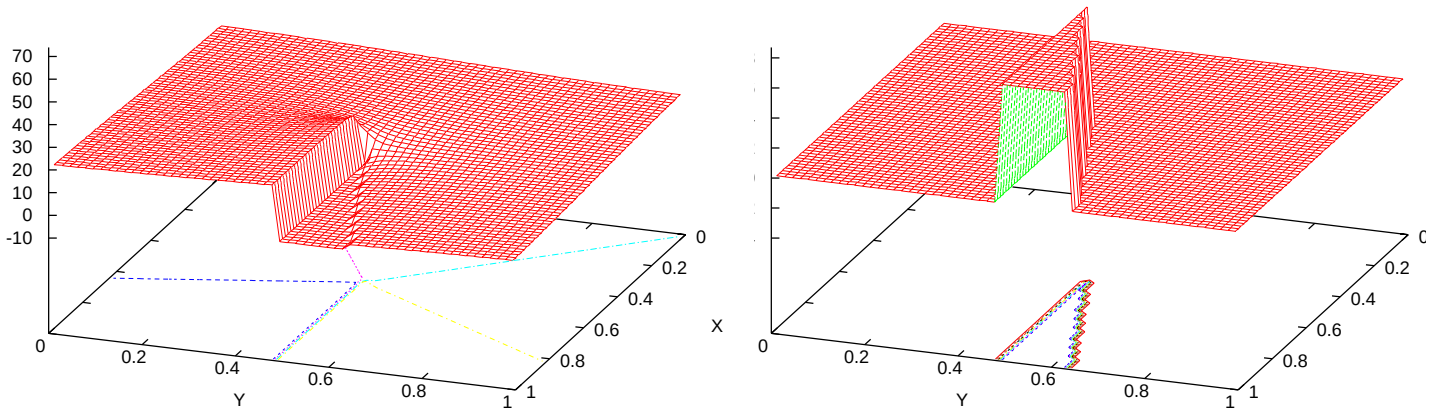
avec

$$A_{nh} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

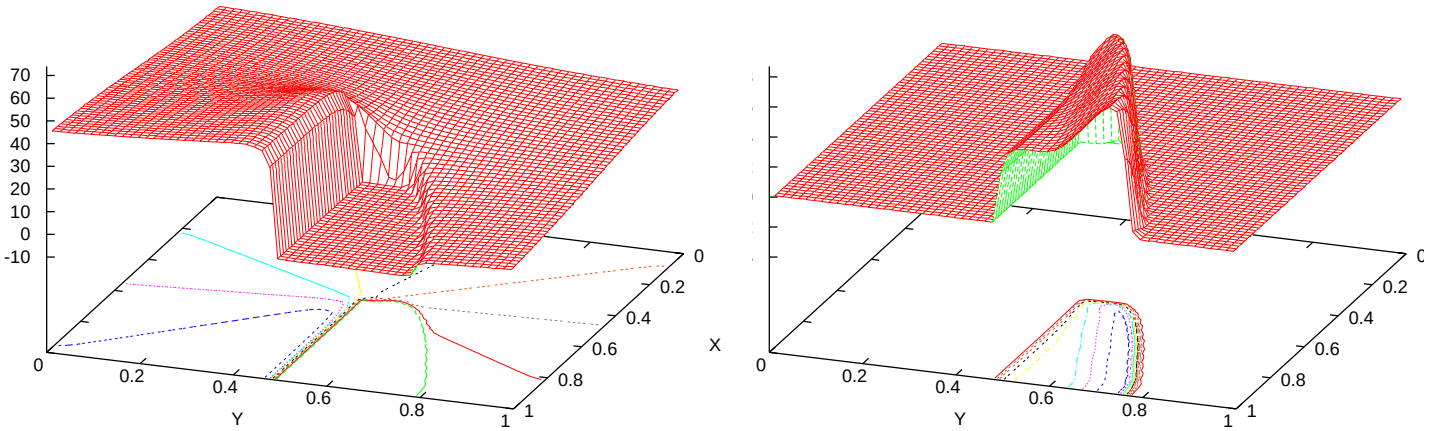
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

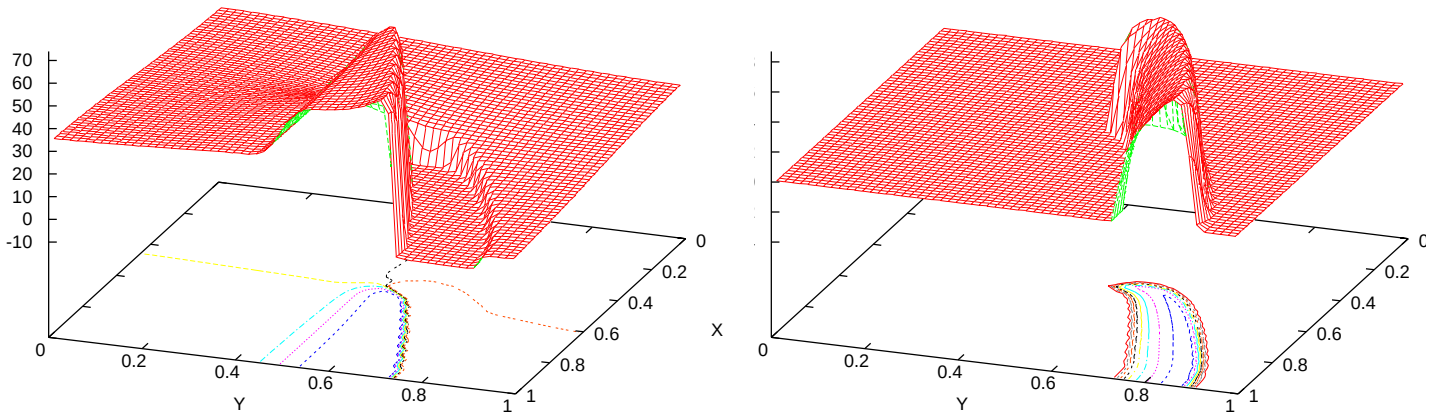
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

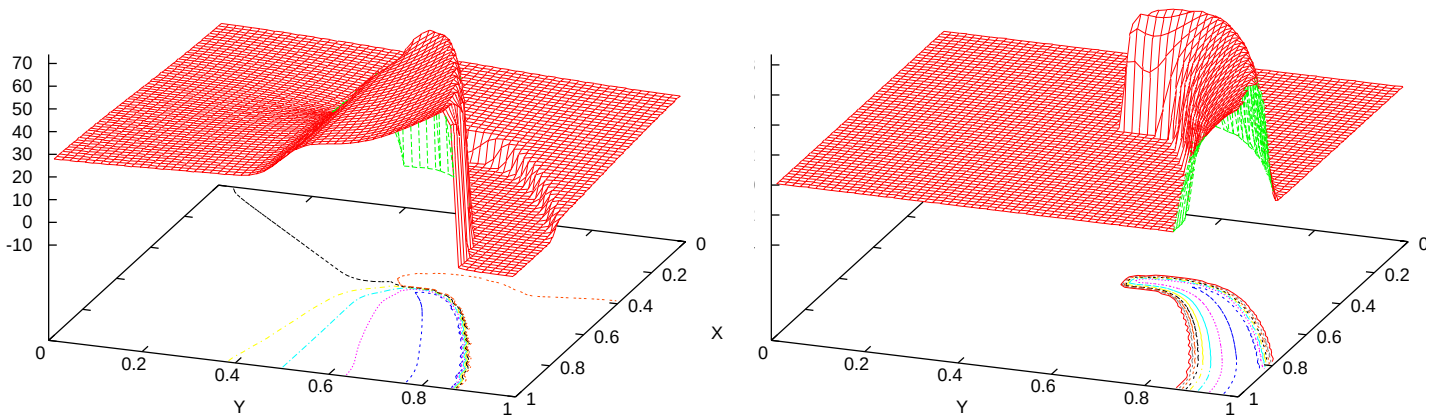
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

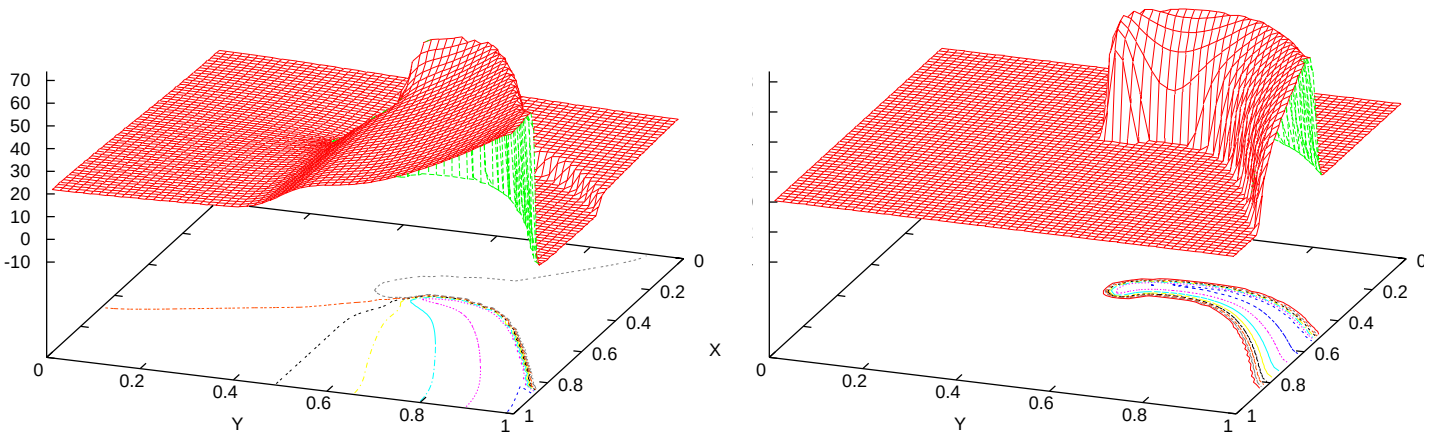
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

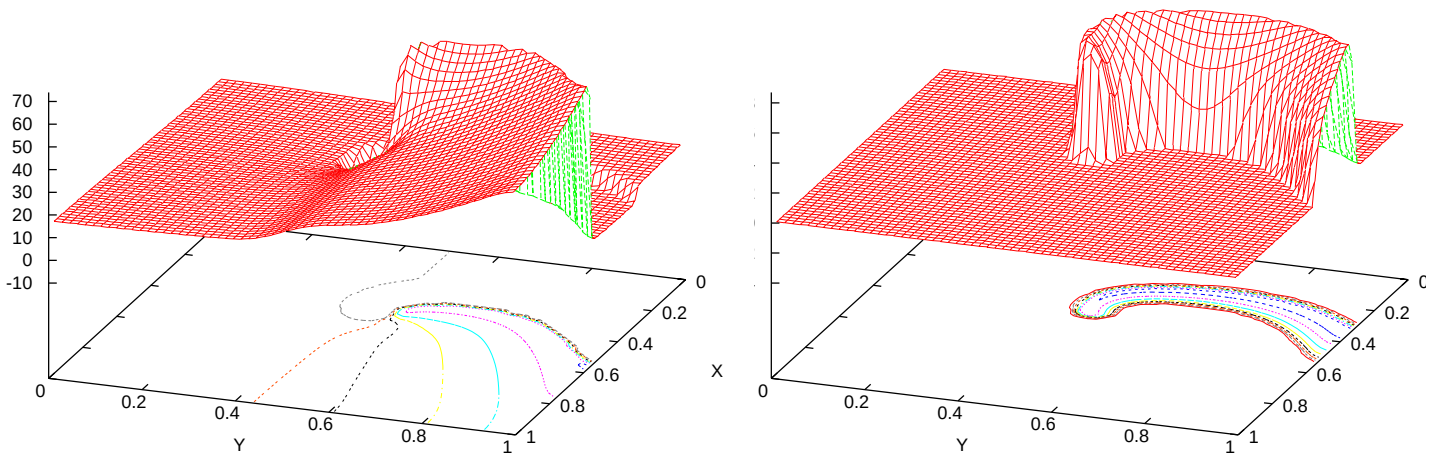
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

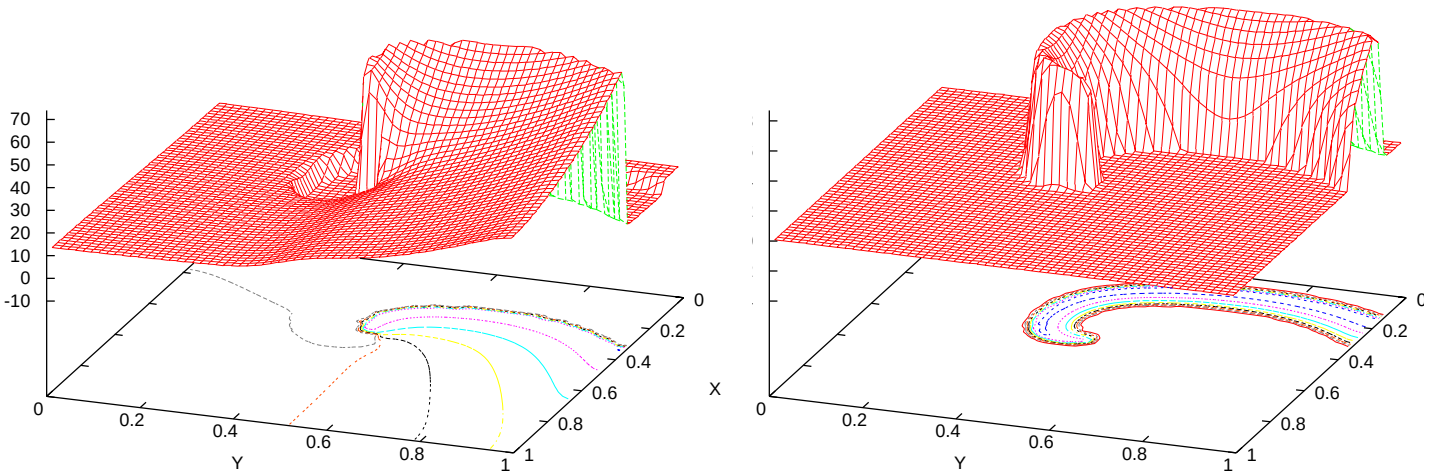
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

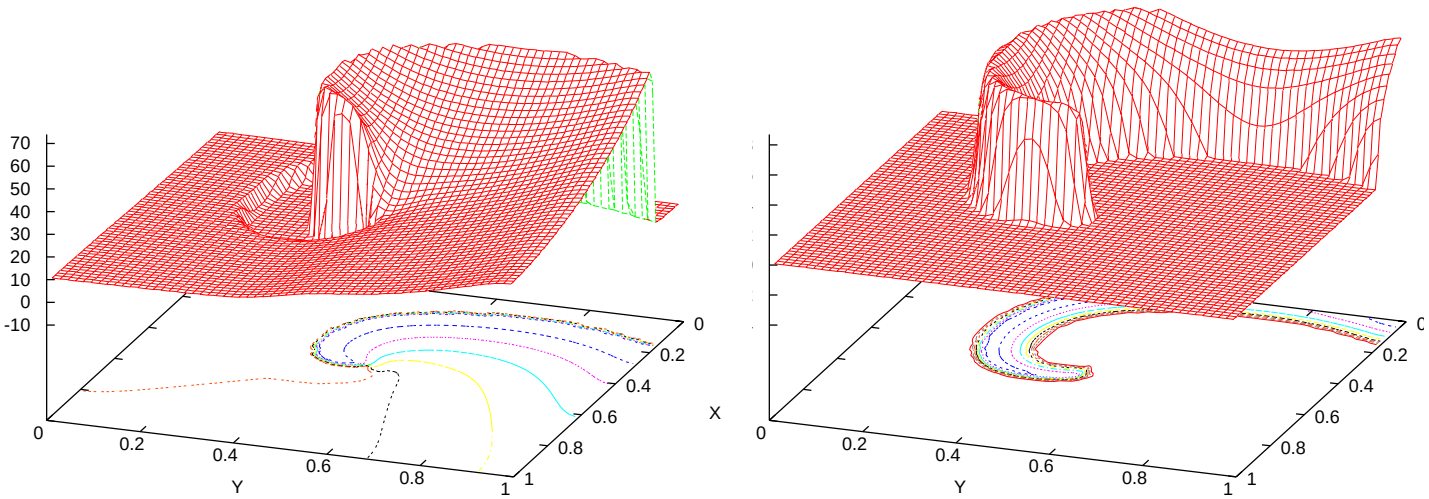
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

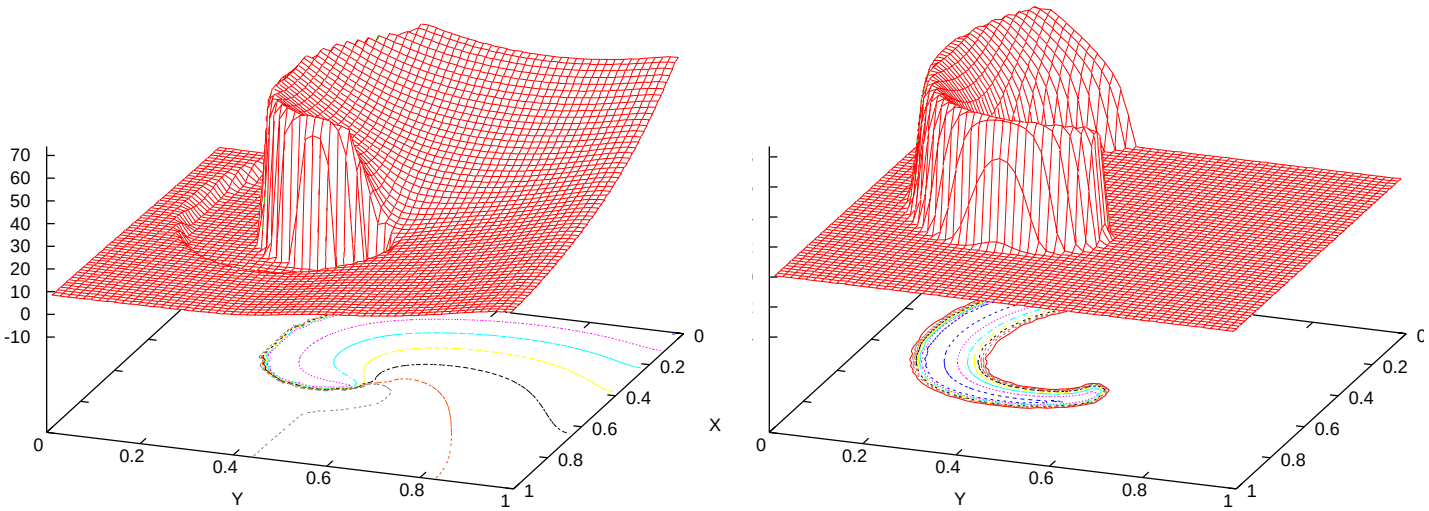
- Système de Belousov-Zhabotinsky à trois variables



Réaction diffusion pure

$$\partial_t U + \sum_{i \in C} \partial_i \left(\Phi_i(U, \partial_{\boldsymbol{x}} U) \right) = \Omega(U)$$

- Système de Belousov-Zhabotinsky à trois variables



Plan de la présentation

Problème de Dahlquist

On considère le problème test de Dahlquist :

$$y' = \lambda y$$

Sa solution exacte est

$$y(t) = e^{\lambda t} C$$

et elle reste bornée pour $t \geq 0$ si $\Re\lambda \leq 0$
La solution numérique d'une méthode de Runge Kutta ou d'une méthode multipas, appliquée avec des pas constants ne dépend que du produit $h\lambda$.
Il est alors intéressant d'étudier pour quelle valeur de $h\lambda$ la solution numérique reste bornée.

Domaine de stabilité - A-stabilité

Definition (A-stabilité)

Considérons une méthode dont la solution numérique $y_n{}_{n\geq 0}$ pour l'équation de test est une fonction de $z = h\lambda$. Alors l'ensemble

$$S := \{z \in \mathbb{C}; \{y_n\}_{n\geq 0} \text{ est bornée}\} \tag{6}$$

s'appelle domaine de stabilité de la méthode. On dit que la méthode est A-stable si

$$S \supset \mathbb{C}^- \text{ où } \mathbb{C}^- = \{z \in \mathbb{C}; \Re z \leq 0\} \tag{7}$$

Classes de méthodes

- Dans le cas des méthodes à un pas explicites, le domaine de stabilité S est borné et la condition de stabilité $h\lambda \in S$ impose une restriction sévère à h. Ces méthodes ne sont donc pas appropriées pour les problèmes raides.
- Les méthodes d'Adams explicites et implicites (à l'exception de la méthode d'Euler implicite et de la règle du trapèze) ont toutes un domaine de stabilité borné et petit. Ces méthodes ne sont donc pas utilisables pour des problèmes raides.
- Les méthodes BDF, par contre, ont un domaine de stabilité plus important, ce qui explique qu'elles sont beaucoup utilisées pour résoudre des problèmes raides. La méthode BDF avec $k = 2$ est même A-stable.
- La *barrière de Dahlquist* dit que l'ordre d'une méthode multipas A-stable ne peut être plus grand que 2. Ce résultat a contribué à la détermination d'autre méthodes d'intégration permettant de combiner A-stabilité et ordre élevé.

↪ Besoin de créer des méthodes à un pas très stable et d'ordre élevé :

Runge Kutta implicite

Plan de la présentation

IRK

Definition (Méthode de Runge-Kutta à s étages)

Soient b_i, a_{ij} $(i, j = 1 \ldots, s)$ des réels. La méthode de Runge-Kutta à s étages s'écrit :

$$\begin{aligned} k_i &= f(x_0 + c_i h, y_0 + h \sum_{j=1}^{s} a_{ij} k_j) \quad i = 1, \ldots, s \\ y_1 &= y_0 + h \sum_{i=1}^{s} b_i k_i \end{aligned}$$

Quand $a_{ij} = 0$ pour $i \geq j$, la méthode est explicite. Si $a_{ij} = 0$ pour $i > j$, et qu'au moins un des $a_{ii} \neq 0$, la méthode est dite diagonalement implicite (diagonal implicit RK, DIRK). Si de plus tous les termes diagonaux sont égaux $a_{ii} = \gamma$ pour $i = 1, \ldots, s$, on parle de "singly diagonal implicit method" (SDIRK). Dans tous les autres cas, la méthode est implicite.

Stabilité et ordre des IRK

$$
\begin{aligned}
B(p): &\quad \sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q} && q = 1, \dots, p \\
C(\eta): &\quad \sum_{j=1}^{s} a_{ij} c_j^{q-1} = \frac{c_i^q}{q} && i = 1, \dots, s,\ q = 1, \dots, \eta \\
D(\zeta): &\quad \sum_{i=1}^{s} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1 - c_j^q) && j = 1, \dots, s,\ q = 1, \dots, \zeta
\end{aligned}
$$

La condition $B(p)$ signifie que la formule de quadrature (b_i, c_i) est d'ordre p.

Théorème (Butcher)

Si les coefficients b_i, c_i, et a_{ij} de la méthode de RK vérifie les hypothèses $B(p)$, $C(\eta)$, et $D(\zeta)$ avec $p \leq \eta + \zeta + 1$ et $p \leq 2\eta + 2$ alors la méthode est d'ordre p.

Plan de la présentation

Origine de la méthode

- Ces méthodes ne sont pas forcément A-stables. Ehle reprend les idées de Butcher et construit des méthodes de type I, II et III avec d'excellentes conditions de stabilité.
- La méthode d'Ehle de type II, appelée méthode de Radau IIA, s'obtient par application de la condition $C(s)$. On peut montrer qu'elle constitue ainsi une méthode de collocation basé sur les zéros de la fonction $\frac{d^{s-1}}{dx^{s-1}}(x^{s-1}(x-1)^s)$. Elle est d'ordre $2s-1$ et est A-stable
- cette méthode pour les valeurs $s=3$ et $p=5$ est aussi L-stable et est implémentée dans le code radau5.

Coefficients de RadauIIa

- cette méthode pour les valeurs $s = 3$ et $p = 5$ est aussi L-stable et est implémentée dans le code Radau5.

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

Plan de la présentation

Réécriture de la méthode

$$\begin{array}{rcl} g_i & = & y_0 + h\sum_{j=1}^{s} a_{ij} f(x_0 + c_j h, g_j) \quad i = 1, \ldots, s \\ y_1 & = & y_0 + h\sum_{j=1}^{s} b_j f(x_0 + c_j h, g_j) \end{array}$$

- On choisit de travailler avec des quantités plus petites afin de réduire les problèmes d'erreurs d'arrondis :

$$z_i = g_i - y_0$$

ce qui permet de réécrire la première équation :

$$z_i = h\sum_{j=1}^{s} a_{ij} f(x_0 + c_j h, y_0 + z_j) \quad i = 1, \ldots, s$$

- Ainsi, lorsque les variables $z_1, \ldots, z_s$ sont connues, la deuxième équation devient explicite pour y_1. Une application directe de ce raisonnement nécessite s évaluations de fonctions additionnelles.

Réécriture de la méthode

On peut éviter ces calculs dans le cas où la matrice $A = (a_{ij})$ n'est pas singulière

$$\begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix} = A \begin{pmatrix} hf(x_0 + c_1 h, y_0 + z_1) \\ \vdots \\ hf(x_0 + c_s h, y_0 + z_s) \end{pmatrix}$$

Et ainsi

$$y_1 = y_0 + \sum_{i=1}^{s} d_i z_i, \qquad (d_1, \ldots, d_s) = (b_1, \ldots, b_s) A^{-1}$$

Dans le cas qui nous intéresse, c'est-à -dire $s = 3$, $d = (0, 0, 1)$, puisque $b_i = a_{si}\ \forall i$.

Algorithme de Newton modifié

Méthode itérative conduit à chaque itération à l'inversion de :

$$\begin{pmatrix} I - ha_{11}\dfrac{\partial f}{\partial y}(x_0 + c_1 h, y_0 + z_1) & \cdots & -ha_{1s}\dfrac{\partial f}{\partial y}(x_0 + c_s h, y_0 + z_s) \\ \vdots & & \vdots \\ -ha_{s1}\dfrac{\partial f}{\partial y}(x_0 + c_1 h, y_0 + z_1) & \cdots & I - ha_{ss}\dfrac{\partial f}{\partial y}(x_0 + c_s h, y_0 + z_s) \end{pmatrix}$$

Approximation de la Jacobienne $J \approx \frac{\partial f}{\partial y}(x_0, y_0)$

$$(I - hA \otimes J)\Delta Z^k = -Z^k + h(A \otimes I)F(Z^k)$$

$$Z^{k+1} = Z^k + \Delta Z^k$$

$Z^k = (z_1^k, \ldots, z_s^k)^T$ est la $k$ème approximation, $\Delta Z^k = (\Delta z_1^k, \ldots, \Delta z_s^k)^T$ les incréments, $F(Z^k) = (f(x_0 + c_1 h, y_0 + z_1^k), \ldots, f(x_0 + c_s h, y_0 + z_s^k))^T$

↪ Chaque itération nécessite s évaluations de f, et la résolution d'un système linéaire de taille ns. Par contre, la matrice $(I - hA \otimes J)$ est la même pour toutes les itérations. Sa décomposition LU n'est donc réalisée qu'une seule fois.

Approximation initiale

Comme la solution exacte vérifie $z_i = \mathcal{O}(h)$, le choix le plus simple est :

$$z_i^0 = 0 \; i = 1, \ldots, s$$

On peut faire des choix plus efficaces. Si q est le polynôme d'interpolation de degré s définit par : $q(0) = 0$, $q(c_i) = z_i$, $i = 1 \ldots, s$, alors la condition initiale :

$$\begin{aligned} z_i^0 &= q(1 + wc_i) + y_0 - y_1 \; i = 1, \ldots, s \\ w &= \frac{h_{new}}{h_{old}} \end{aligned}$$

donne une convergence numérique plus rapide

Estimation d'erreur et critère d'arrêt

Comme la convergence est linéaire, on a, en espérant que $\Theta < 1$:

$$\|\Delta Z^{k+1}\| \leq \Theta \|\Delta Z^k\|$$

Si on applique l'inégalité triangulaire au développement (Z^* est la solution exacte) $Z^{k+1} - Z^* = (Z^{k+1} - Z^{k+2}) + (Z^{k+2} - Z^{k+3}) + \ldots$, on obtient l'estimation d'erreur :

$$\|Z^{k+1} - Z^*\| \leq \frac{\Theta}{1-\Theta} \|\Delta Z^k\|$$

Le taux de convergence peut être estimé à partir de quantités calculées :

$$\Theta_k = \frac{\|\Delta Z^k\|}{\|\Delta Z^{k-1}\|}$$

L'erreur numérique ne doit pas être supérieure à l'erreur de discrétisation locale proche de *Tol*. Les itérations sont donc arrêtées quand :

$$\eta_k \|\Delta Z^k\| \leq \kappa . Tol \text{ avec } \eta_k = \frac{\Theta_k}{1-\Theta_k}$$

Cela ne peut s'appliquer qu'après 2 itérations de la méthode.

Estimation d'erreur et critère d'arrêt

Il reste encore à choisir le paramètre κ. Des expériences numériques montrent une meilleure efficacité du code pour des valeurs de κ autour de 10^{-1} ou 10^{-2}.
De même, il semble que le code soit plus efficace quand le nombre maximum d'itérations k_{max} est élevé ($k_{max} = 7$ ou 10). Durant ces itérations, le calcul est interrompu et redémarré avec un pas plus petit (par exemple avec $h := h/2$) dans le cas où on se trouve dans une des situations suivantes :

- $\Theta_k \geq 1$ pour un k,
- Pour quelques k,

$$\frac{\Theta_k^{k_{max}-k}}{1-\Theta_k} \|\Delta Z^k\| > \kappa . Tol$$

Si une seule itération de Newton est suffisante pour satisfaire le critère d'arrêt ou si le dernier $\Theta_k \leq 10^{-3}$, le jacobien n'est pas recalculé au pas suivant.

Plan de la présentation

Résolution efficace du système linéaire

La résolution se fait en exploitant la structure particulière de la matrice $I - hA \otimes J$. L'idée est de multiplier par $(hA)^{-1} \otimes I$ et de transformer A^{-1} en une matrice simple (diagonal par bloc, triangulaire ...) :

$$T^{-1} A^{-1} T = \Lambda$$

Si on considère le changement de variables $W^k = (T^{-1} \otimes I) Z^k$, on obtient :

$$(h^{-1}\Lambda \otimes I - I \otimes J)\Delta W^k = -h^{-1}(\Lambda \otimes I) W^k + (T^{-1} \otimes I) F((T \otimes I) W^k)$$

$$W^{k+1} = W^k + \Delta W^k$$

Résolution efficace du système linéaire

Si on suppose que A^{-1} a une valeur propre réel $\hat{\gamma}$ et une paire de valeurs propres complexes conjuguées $\hat{\alpha} \pm i\hat{\beta}$ (ce qui est le cas pour Radau IIA), la matrice se réécrit :

$$\begin{pmatrix} \gamma I - J & 0 & 0 \\ 0 & \alpha I - J & -\beta I \\ 0 & \beta I & \alpha I - J \end{pmatrix}$$

avec $\gamma = h^{-1}\hat{\gamma}$, $\alpha = h^{-1}\hat{\alpha}$, $\beta = h^{-1}\hat{\beta}$.
Ainsi, le système peut se décomposer en deux systèmes linéaires de taille n et $2n$.
D'autres idée sont possibles pour exploiter la structure particulière de la matrice $2n \times 2n$.

Plan de la présentation

Calcul de l'erreur pour le contrôle

Comme la méthode est d'ordre optimale, on ne peut pas se baser sur une autre plus précise pour estimer l'erreur. Donc, on considère une méthode d'ordre moins élevé de la forme :

$$\hat{y_1} = y_0 + h(\hat{b_0} f(x_0, y_0) + \sum_{i=1}^{3} \hat{b_i} f(x_0 + c_i h, g_i))$$

où les g_1, g_2, g_3 sont les valeurs obtenues pour la méthode de Radau IIA, et $\hat{b_0} \neq 0$.
On peut écrire la différence entre les solutions des deux méthodes ($\hat{b_0} = \gamma_0 = \hat{\gamma}^{-1}$) :

$$\hat{y_1} - y_1 = \gamma_0 h f(x_0, y_0) + \sum_{i=1}^{3} (\hat{b_i} - b_i) h f(x_0 + c_i h, g_i) = \gamma_0 h f(x_0, y_0) + e_1 z_1 + e_2 z_2 + e_3 z_3,$$

dont on se sert pour l'évaluation de l'erreur :

$$err = (I - h\gamma_0 J)^{-1} (\hat{y_1} - y_1)$$

Dans le cas du premier pas, et après chaque pas rejeté pour lequels $\|err\| > 1$, l'erreur est calculé par :

$$\widetilde{err} = (I - h\gamma_0 J)^{-1} (\gamma_0 h f(x_0, y_0 + err) + e_1 z_1 + e_2 z_2 + e_3 z_3)$$

Contrôle standard

Les expressions se comportent en $\mathcal{O}(h^4)$, la prédiction du pas s'écrit donc :

$$h_{new} = fac.h_{old}.\|err\|^{-1/4}$$

avec $\|err\| = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{err_i}{sc_i})^2}$ et $sc_i = Atol_i + \max(|y_{0i}|, |y_{1i}|).Rtol_i$. Le coefficient fac dépend du nombre d'itérations de Newton $Newt$
$fac = 0.9 \times (2k_{max} + 1)/(2k_{max} + Newt)$

Pour limiter le nombre de décomposition LU de la matrice, si le jacobien n'est pas recalculé et si $c_1 h_{old} < h_{new} < c_2 h_{old}$ avec $c_1 = 1.0$ et $c_2 = 1.2$, alors on conserve h_{old} pour le pas suivant.

Contrôle prédictif

Dans le cas de systèmes très raides, la décroissance du pas de temps peut devoir être extrêmement rapide alors que la précédente est limitée par le coefficient *fac*.
Si on note err_{n+1} l'erreur donnée et si h_n correspond au pas du calcul de l'étape n, la prédiction du pas est issue de la formule asymptotique :

$$\|err_{n+1}\| = C_n h_n^4$$

L'évaluation de h_{new} est basé sur l'hypothèse que $C_{n+1} \approx C_n$. Un meilleur modèle peut être obtenu en considérant que $\log C_n$ est une fonction linéaire de n, et donc que $\log C_{n+1} - \log C_n$ est constant ou encore :

$$C_{n+1}/C_n \approx C_n/C_{n-1}$$

On obtient ainsi une nouvelle estimation du pas :

$$h_{new} = fac.h_n\left(\frac{1}{\|err_{n+1}\|}\right)^{1/4}.\frac{h_n}{h_{n+1}}\left(\frac{\|err_n\|}{\|err_{n+1}\|}\right)^{1/4}$$

Dans le code, le nouveau pas correspond au minimum des résultats obtenus précédemment