Chapitre 17

Etude des erreurs numériques dans la résolution des équations différentielles

1 Introduction

1.1 Rappel

Définition : Soit une fonction numérique y de variable t définie sur un intervalle I de $\mathbb{R}$ et dérivable jusqu'à l'ordre p. Une équation différentielle scalaire d'ordre p est de la forme :

$$g(t, y, y', \cdots, y^{(p)}) = 0 \quad \textbf{(1)}$$

On peut définir de la même manière un système différentiel (ou équation différentielle vectorielle) en considérant g comme une fonction vectorielle (l'ordre du système est l'ordre le plus élevé de dérivation de l'ensemble des équations).
Une solution (ou intégrale) de l'équation différentielle **(1)** est une fonction y de la variable t, définie sur I et p fois dérivable sur cet intervalle vérifiant cette équation. Intégrer ou résoudre cette équation revient à trouver toutes ses solutions dépendant de p paramètres arbitraires.

Définition : Une équation différentielle scalaire s'écrit sous forme canonique :

$$y^{(p)} = g(t, y, y', \cdots, y^{(p-1)})$$

Proposition : Une équation différentielle scalaire canonique d'ordre p est équivalente à un système différentiel du premier ordre de dimension p :

(1) $Y' = f(t, Y)$ avec Y variable vectorielle $(y_1, y_2, \cdots, y_p)$
on pose $y_1 = y$, $y_2 = y'$, $\cdots$, $y_p = y^{(p-1)}$; la fonction vectorielle f est telle que $f(t, y_1, y_2, \cdots, y_p) = (y_1, y_2, \cdots, y_p)' = (y_2, y_3, \cdots, y_p, y^{(p)}) = (y_2, y_3, \cdots, y_p, g(t, y_1, y_2, \cdots, y_p))$.

Définition : Un problème de Cauchy (ou de conditions initiales) consiste à trouver une solution $y : I \to \mathbb{R}^p$ de l'équation différentielle $y' = f(t, y)$ vérifiant la condition de Cauchy : $y(t_0) = y_0$ avec $y_0 \in \mathbb{R}^p$ donné et $t_0 \in \mathring{I}$ (intérieur de l'intervalle I).

Définition : Soit Ω un ouvert de $\mathbb{R}^p$. Une fonction f définie de $I \times \Omega$ dans $\mathbb{R}^p$ est lipschitzienne en y sur $I \times \Omega$ s'il existe une constante $L > 0$ telle que

$$||f(t, y) - f(t, \widetilde{y})|| \leq L||y - \widetilde{y}||$$

pour tout $(t, y) \in I \times \Omega$ et $(t, \widetilde{y}) \in I \times \Omega$

Théorème de Cauchy-Lipschitz : Si f est continue et lipschitzienne en y sur $I \times \Omega$ alors le problème de Cauchy a une solution unique y sur I telle que $y(t_0) = y_0$.

Remarques :
La condition de Lipschitz est suffisante mais non nécessaire pour l'existence et l'unicité de la solution.

Si f est différentiable en y, si sa différentielle en y est continue et si $J_f(t,y)$ matrice carrée d'ordre p, appelée jacobienne en y, de coefficients $\dfrac{\partial f_i}{\partial y_j}(t,y) \quad i,j=1,\dots,p$, est bornée ($||J_f(t,y)|| \leq M_1$) sur $I \times \Omega$, alors f est lipschitzienne en y sur $I \times \Omega$.

1.2 Stabilité de la solution mathématique d'un problème de Cauchy

Cette notion est très importante car elle permet d'étudier l'influence des perturbations sur les données du problème : y_0 (respectivement f) peuvent être entachés d'erreurs de codage (respectivement d'erreur d'arrondi) lors de la résolution numérique sur ordinateur.
On considère l'intervalle $I_0 = [t_0, t_0 + T] \subset I$ avec $T > 0$.
Proposition : Soit le problème de Cauchy $y' = f(t,y)$ et $y(t_0) = y_0$ et le problème perturbé $z' = f(t,z) + \delta(t)$ et $z(t_0) = y_0 + \epsilon(t_0)$
Soient y et z sur I les solutions de ces deux problèmes sous les hypothèses du théorème de Cauchy-Lipschitz.
Si la fonction δ est bornée sur I par un nombre γ, l'erreur $\epsilon(t) = z(t) - y(t)$ est bornée sur I_0 et on a la majoration :

$$\max_{t \in I_0} ||\epsilon(t)|| \leq \max(||\epsilon(t_0)||, \gamma)\frac{1}{L}((1+L)e^{LT} - 1)$$

Démonstration :
On a $||z' - f(t,z)|| < \gamma$ en appliquant le théorème des accroissements finis à la fonction $z(t) - \int_{t_0}^t f(s,z(s))ds$ on obtient $||z(t) - z(t_0) - \int_{t_0}^t f(s,z(s))ds|| \leq \gamma(t - t_0) \quad t \geq t_0$ d'où
$||\epsilon(t)|| \leq ||\epsilon(t_0)|| + ||\int_{t_0}^t f(s,z(s))ds - \int_{t_0}^t f(s,y(s))ds|| + \gamma(t - t_0) \qquad \mathbf{(2)}$
avec la condition de Lipschitz on a
$||\int_{t_0}^t (f(s,z(s)) - \int_{t_0}^t f(s,y(s)))ds|| \leq \int_{t_0}^t ||f(s,z(s)) - \int_{t_0}^t f(s,y(s))||ds \leq \int_{t_0}^t L||z(s) - y(s)||ds$
d'après **(2)** on obtient
$||\epsilon(t)|| \leq ||\epsilon(t_0)|| + \gamma(t - t_0) + L\int_{t_0}^t ||\epsilon(s)||ds \quad t \geq t_0$
C'est une "inéquation intégrale" en la fonction $||\epsilon(t)||$ de la forme $||\epsilon(t)|| \leq \varphi(t) + L\int_{t_0}^t ||\epsilon(s)||ds$
avec $\varphi(t) = ||\epsilon(t_0)|| + \gamma(t - t_0)$
on pose $u(t) = \int_{t_0}^t ||\epsilon(s)||ds$ on a $u'(t) = ||\epsilon(t)||$ et on pose $v(t) = u(t)e^{-L(t-t_0)}$ on a $v'(t) = u'(t)e^{-L(t-t_0)} - Lu(t)e^{-L(t-t_0)} = (u'(t) - Lu(t))e^{-L(t-t_0)} = (||\epsilon(t)|| - Lu(t))e^{-L(t-t_0)} \leq \varphi(t)e^{-L(t-t_0)}$
donc $\int_{t_0}^t v'(s)ds \leq \int_{t_0}^t \varphi(s)e^{-L(s-t_0)}ds$ puisque $v(t_0) = u(t_0) = 0$ on a $v(t) \leq \int_{t_0}^t \varphi(s)e^{-L(s-t_0)}ds$
et par définition de v on obtient $u(t)e^{-L(t-t_0)} \leq \int_{t_0}^t \varphi(s)e^{-L(s-t_0)}ds$ soit
$u(t) \leq \int_{t_0}^t \varphi(s)e^{-L(s-t_0)}ds e^{L(t-t_0)} \leq \int_{t_0}^t \varphi(s)e^{L(t-s)}ds$
d'où en utilisant l'inégalité $||\epsilon(t)|| \leq \varphi(t) + Lu(t)$ on a $||\epsilon(t)|| \leq \varphi(t) + L\int_{t_0}^t \varphi(s)e^{L(t-s)}ds$
en remplaçant $\varphi(t)$ on obtient $||\epsilon(t)|| \leq ||\epsilon(t_0)|| + \gamma(t - t_0) + L\int_{t_0}^t (||\epsilon(t_0)|| + \gamma(s - t_0))e^{L(t-s)}ds$
$||\epsilon(t)|| \leq ||\epsilon(t_0)||(1 + L\int_{t_0}^t e^{L(t-s)}ds) + \gamma(t - t_0 + L\int_{t_0}^t (s - t_0)e^{L(t-s)}ds)$
$||\epsilon(t)|| \leq ||\epsilon(t_0)||e^{L(t-t_0)} + \frac{\gamma}{L}(e^{L(t-t_0)} - 1) \quad t \geq t_0$
c'est-à-dire sur I_0

$$||\epsilon(t)|| \leq ||\epsilon(t_0)||e^{LT} + \frac{\gamma}{L}(e^{LT} - 1),\ t \in I_0$$

Définition : Un problème de Cauchy est mathématiquement bien posé s'il admet une solution et une seule et si elle dépend continuement des données.

Remarque : la proposition précédente donne des conditions pour qu'un problème de Cauchy soit mathématiquement bien posé.

2 Méthodes d'intégration numérique à un pas

2.1 Introduction

On considère un problème de Cauchy mathématiquement bien posé et on veut calculer une approximation de la solution $y(t)$ en certains points t_k $k = 1, ..., n$ de l'intervalle I, supposés équidistants avec un pas d'intégration $h = t_{k+1} - t_k$.

Définition : Une méthode à un pas a une équation récurrente de la forme :

(3) $\quad y_{k+1} = y_k + h\Phi(t_k, y_k, h) \quad k = 0, ..., n-1$ avec y_0 donné

C'est une méthode numérique à un pas (explicite) car la valeur de y_{k+1} se calcule à partir de y_k seulement. On suppose Φ continue.

Définition :

1. Une méthode à un pas est consistante avec le problème de Cauchy si on a :
$$\lim_{h\to 0} \max_{k=0,1,\ldots,n} ||\frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k), h)|| = 0$$
où $y(t)$ est la solution mathématique du problème.
2. Une méthode à un pas est stable s'il existe deux constantes S et M telles que :
$$\max_{k=0,1,\ldots,n} ||y_k - z_k|| \leq S||y_0 - z_0|| + M \max_{k=0,1,\ldots,n} ||\epsilon_k||$$
en considérant la méthode perturbée avec l'équation récurrente de la forme :
$$\textbf{(4)} \quad z_{k+1} = z_k + h(\Phi(t_k, z_k, h) + \epsilon_k) \text{ avec } z_0 \text{ donné}$$
3. Une méthode à un pas est convergente si :
$$\lim_{h\to 0} \max_{k=0,1,\ldots,n} ||y_k - y(t_k)|| = 0$$

Remarques :
La consistance permet de dire que Φ est une approximation de y'.
La stabilité de la méthode assure que de petites erreurs initiales ou de calcul sur Φ provoquent une erreur sur la solution numérique y_k bornée, contrôlable.

Théorème : Une méthode à un pas consistante et stable est convergente.

Démonstration :
Par consistance on a $y(t_{k+1}) = y(t_k) + h(\Phi(t_k, y(t_k), h) + \epsilon_k)$ avec $\lim_{h\to 0} \max_{k=0,1,\ldots,n} ||\epsilon_k|| = 0$ et par stabilité $\max_{k=0,1,\ldots,n} ||y_k - y(t_k)|| \leq S||y_0 - y(t_0)|| + M \max_{k=0,1,\ldots,n} ||\epsilon_k||$ or $y(t_0) = y_0$.

Théorème : Une méthode à un pas est consistante si et seulement si on a : $\Phi(t, y, 0) = f(t, y)$ pour tout $(t, y) \in I \times \Omega$, avec Ω un ouvert de $\mathbb{R}^p$

Théorème : Si la fonction Φ est lipschitzienne en y :
$$||\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)|| \leq K||y - \widetilde{y}||$$

Pour tout (t, y), $(t, \widetilde{y})$ dans $I \times \Omega$, avec K indépendant de h, alors la méthode est stable.

Démonstration : Soient y_k et z_k les deux suites calculées pour l'équation récurrente **(3)** et la méthode perturbée **(4)**. On a alors :

$$||y_{k+1} - z_{k+1}|| \leq ||y_k - z_k|| + h||\Phi(t_k, y_k, h) - \Phi(t_k, z_k, h)|| + h||\epsilon_k||$$

Φ est lipschitzienne donc :

$$||y_{k+1} - z_{k+1}|| \leq ||y_k - z_k|| + hK||y_k - z_k|| + h||\epsilon_k||$$

c'est-à-dire :

$$||y_{k+1} - z_{k+1}|| \leq (1 + hK)||y_k - z_k|| + h||\epsilon_k||$$

Par récurrence, on obtient :

$$||y_{k+1} - z_{k+1}|| \leq (1 + hK)^{k+1}||y_0 - z_0|| + \frac{(1 + hK)^{k+1} - 1}{K} \max_{k=0,\ldots,n} ||\epsilon_k||$$

Par l'inégalité $1 + hK \leq e^{hK}$ et par définition du pas h ($(k + 1)h \leq T$), on obtient :

$$||y_{k+1} - z_{k+1}|| \leq e^{KT}||y_0 - z_0|| + \frac{e^{KT} - 1}{K} \max_{k=0,\ldots,n} ||\epsilon_k||$$

Remarque : La méthode est stable si le pas h est inférieur à une valeur h_{max} et la constante K est indépendante de h lorsque celui-ci est assez petit.

Définition : Une méthode à un pas est d'ordre r si on a :

$$\max_{k=0,\ldots,n} |\frac{1}{h}(y(t_{k+1}) - y(t_k)) - \Phi(t_k, y(t_k), h)| \leq Ch^r$$

où C est une constante indépendante de h et $y(t)$ est solution mathématique du problème de Cauchy (p=1).

Remarque : Toute méthode d'ordre r est consistante.

Théorème : Si Φ est lipschitzienne en y et la méthode à un pas d'ordre r alors l'erreur de discrétisation (ou de méthode) au pas k :
$e_k = y_k - y(t_k)$ vérifie la majoration $\max\limits_{k=1,\ldots,n} |e_k| \leq Rh^r$

Démonstration :
On intègre l'équation différentielle $y' = f(t, y)$ entre t_k et t_{k+1} :

$$y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} f(t, y(t))dt$$

Par définition de la méthode numérique on obtient

$$e_{k+1} = e_k + h\left(\Phi(t_k, y_k, h) - \frac{1}{h}\int_{t_k}^{t_{k+1}} f(t, y(t))dt\right)$$

Soit $e_{k+1} = e_k + h(\Phi(t_k, y_k, h) - \Phi(t_k, y(t_k), h)) \, + \, h \left(\Phi(t_k, y(t_k), h) - \dfrac{1}{h} \displaystyle\int_{t_k}^{t_{k+1}} f(t, y(t))dt \right)$
La méthode est d'ordre r, d'où :
$\Phi(t_k, y(t_k), h) - \dfrac{1}{h} \displaystyle\int_{t_k}^{t_{k+1}} f(t, y(t))dt \, = \, \Phi(t_k, y(t_k), h) - \dfrac{y(t_{k+1}) - y(t_k)}{h} \leq Ch^r$
Φ est lipschitzienne en y donc :

$$\Phi(t_k, y_k, h) - \Phi(t_k, y(t_k), h) \leq K|y_k - y(t_k)|$$

d'où $|e_{k+1}| \leq (1 + hK)|e_k| \, + \, Ch^{r+1}$
Soit par récurrence :

$$|e_k| \leq \frac{C}{K}(e^{(t_k - t_0)K} - 1)h^r$$

Remarque :
L'ordre d'une méthode permet de donner une précision sur la solution numérique. Mais ce calcul d'erreur ne prend pas en compte les erreurs d'arrondi. Les valeurs informatiques $\widetilde{y_k}$ calculées en machine par l'équation récurrente sont :

$$\widetilde{y_{k+1}} = \widetilde{y_k} + h\Phi(t_k, \widetilde{y_k}, h) + h\rho_k + \sigma_k$$

où ρ_k est l'erreur d'arrondi sur $\Phi(t_k, \widetilde{y_k}, h)$ et
σ_k est l'erreur d'arrondi sur le calcul de $\widetilde{y_{k+1}}$, avec $\widetilde{y_0} = y_0 \, + \, \epsilon_0$
On suppose ρ_n (respectivement σ_n) borné par ρ (resp. σ)
ρ et σ dépendent de l'arithmétique machine.
Si la méthode est stable on a :

$$\max_{k=0,1,\ldots,n} |\widetilde{y_k} - y_k| \leq S|\widetilde{y_0} - y_0| + M(\max_{k=0,1,\ldots,n} |\rho_k| + \frac{1}{h} \max_{k=0,1,\ldots,n} |\sigma_k|)$$

Soit :

$$\max_{k=0,1,\ldots,n} |\widetilde{y_k} - y_k| \leq S|\epsilon_0| + M(\rho + \frac{\sigma}{h})$$

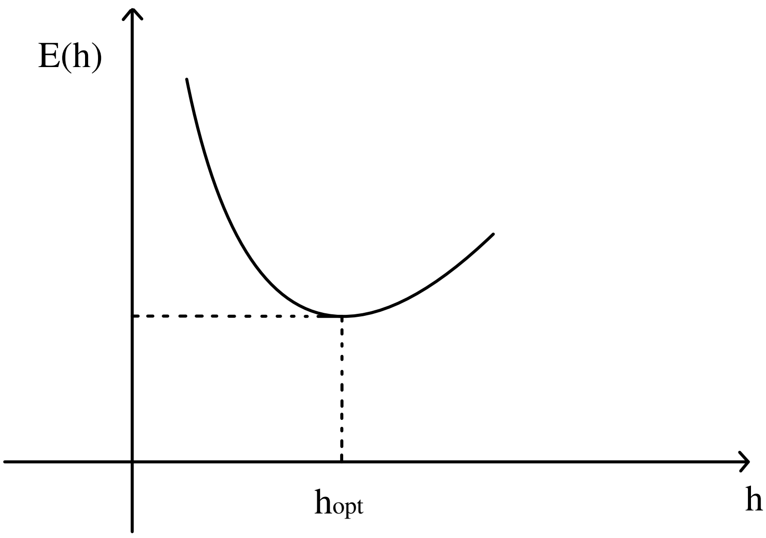
A ces erreurs d'arrondi s'ajoute l'erreur globale de méthode :

$$\max_{k=1,\ldots,n} |y_k - y(t_k)| \leq Rh^r$$

L'erreur totale est donc :

$$\max_{k=0,1,\ldots,n} |\widetilde{y_k} - y(t_k)| \leq S|\epsilon_0| + M\rho + \frac{M\sigma}{h} + Rh^r$$

Notons $E(h) = S|\epsilon_0| + M\rho + \dfrac{M\sigma}{h} + Rh^r$
L'étude de $E(h)$ donne le graphique suivant :



L'erreur passe par un minimum pour $h_{opt} = (\frac{M\sigma}{rR})^{\frac{1}{r+1}}$

Si on prend un pas h plus petit que h_{opt}, l'erreur augmente : le nombre de pas n augmente, et avec lui les erreurs d'arrondi qui l'emportent sur l'erreur globale de discrétisation.
Exemple : $y' = y$, $y_0 = 1$ sur $[0, 1]$ avec une arithmétique standard flottante

Définition : Un problème de Cauchy est numériquement bien posé si la continuité de la solution mathématique par rapport aux données est numériquement satisfaisante pour une arithmétique flottante donnée.

Remarque : La proposition précédente sur la stabilité de la solution mathématique d'un problème de Cauchy montre que la continuité de cette solution par rapport aux données est numériquement bonne lorsque la quantité e^{LT} est petite par rapport aux paramètres (précision,...) de l'arithmétique flottante.
Exemple : $y' = 3y - 1$, $y_0 = 1/3$ sur $[0, 10]$ avec une arithmétique standard flottante.

Remarque : La définition précédente ne prend pas en compte la méthode numérique utilisée.

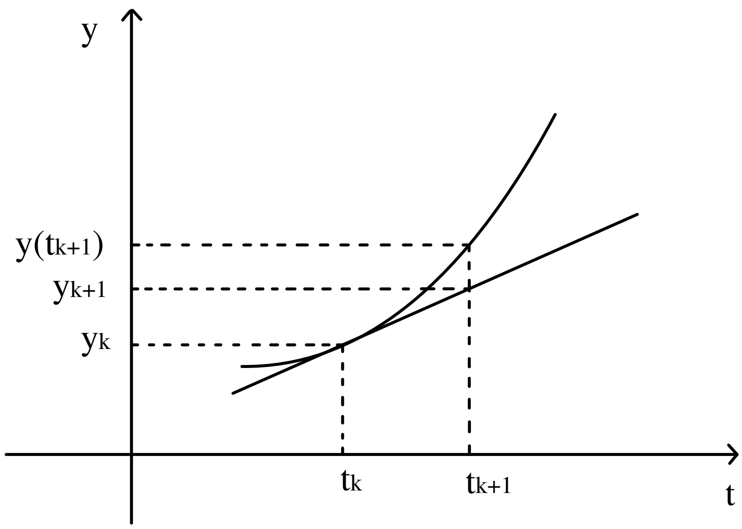
Définition : Un problème de Cauchy est numériquement bien conditionné si les méthodes numériques usuelles donnent pour un coût raisonnable une solution numérique satisfaisante.

Remarque : Le théorème précédent sur la stabilité de la solution numérique d'un problème de Cauchy montre que pour les méthodes à un pas, ce problème est numériquement bien conditionné lorsque la quantité e^{KT} est petite par rapport aux paramètres de l'arithmétique flottante.
Exemple : $y' = -150y + 30$, $y_0 = 1/5$ sur $[0, 1]$ avec une arithmétique standard flottante (méthode d'Euler utilisée)

Remarque : Lorsqu'un problème est numériquement mal conditionné, la solution numérique n'est pas satisfaisante, même pour un coût important : ce sont des problèmes raides (la quantité e^{KT} peut être très grande).
Exemple : $y' = 100(\sin t - y)$, $y_0 = 0$ sur $[0, 3]$ avec une arithmétique standard flottante (méthode de Runge-Kutta utilisée).

2.2 La méthode d'Euler

$y_{k+1} = y_k + hf(t_k, y_k) \quad k = 0, ..., n-1$ avec y_0 donné
On a $\Phi(t_k, y_k, h) = f(t_k, y_k)$ et on approxime la solution au voisinage de t_k par sa tangente en t_k.



Cette méthode est stable et consistante. Elle est d'ordre 1 : si la solution est suffisamment dérivable on a :

$$y(t_k + h) = y(t_k) + hy'(t_k) + h^2 y''(t_k + \theta h) \qquad 0 \leq \theta < 1$$

soit :

$$y(t_k + h) = y(t_k) + hf(t_k, y(t_k)) + h^2 y''(t_k + \theta h)$$

Par définition de la méthode :

$$\frac{1}{h}\left(y(t_{k+1}) - y(t_k)\right) - \Phi(t_k, y(t_k), h) = \frac{y(t_{k+1}) - y(t_k)}{h} - f(t_k, y(t_k))$$

donc :

$$\frac{1}{h}\left(y(t_{k+1}) - y(t_k)\right) - \Phi(t_k, y(t_k), h) = hy''(t_k + \theta h) \qquad 0 \leq \theta < 1$$

Si la dérivée seconde y'' est bornée par une constante R dans l'intervalle $I_0 = [t_0, t_0 + T]$, on a :

$$\max_{k=1,\dots,n} \left| \frac{1}{h}\left(y(t_{k+1}) - y(t_k)\right) - \Phi(t_k, y(t_k), h) \right| \leq Rh$$

2.3 Les méthodes de Runge-Kutta

Elles utilisent le calcul de la fonction $f(x, y)$ en un certain nombre de points intermédiaires de l'intervalle $[t_k, t_{k+1}]$.

2.3.1 Méthodes à un point intermédiaire

Méthodes explicites
Elles sont définies par :

$$y_{k,1} = y_k + \frac{h}{2\alpha} f(t_k, y_k)$$

$$y_{k+1} = y_k + h((1-\alpha)f(t_k, y_k) + \alpha f(t_k + \frac{h}{2\alpha}, y_{k,1}))$$

où α est un paramètre non nul :
$\alpha = 1$ (tangente améliorée ou point milieu) , $\alpha = 1/2$ (Euler-Cauchy), $\alpha = 3/4$ (Heun)
Ces méthodes appelées RK_{22} sont d'ordre 2.

2.3.2 Méthodes à trois points intermédiaires

$$y_{k,1} = y_k + \frac{h}{2} f(t_k, y_k)$$

$$y_{k,2} = y_k + \frac{h}{2} f(t_k + \frac{h}{2}, y_{k,1})$$

$$y_{k,3} = y_k + h f(t_k + \frac{h}{2}, y_{k,2})$$

$$y_{k+1} = y_k + \frac{h}{6}\left(f(t_k, y_k) + 2f(t_k + \frac{h}{2}, y_{k,1}) + 2f(t_k + \frac{h}{2}, y_{k,2}) + f(t_{k+1}, y_{k,3})\right)$$

Cette méthode appelée RK_{44} est d'ordre 4.