Approche numérique à l'usage du physicien pour résoudre les équations différentielles ordinaires. I. Cas des équations du premier ordre.

Hubert Baty, Observatoire Astronomique de Strasbourg.

July 10, 2018

Abstract

Ce document constitue la première partie d'un ensemble qui aurait du voir le jour sous forme d'un livre et qui pour diverses raisons sera finalement publié au détail sur HAL, chapitre par chapitre. Le contenu sert (et a servi) de support de cours pour un enseignement dit de calcul scientifique à l'usage du physicien de niveau Licence et Master1.

La motivation de cet ouvrage provient d'un double constat que j'ai fait à l'issue de mon expérience d'une trentaine d'années passées à enseigner diverses matières scientifiques à l'université. Ce constat est particulièrement vrai pour les filières de sciences physiques. Les mathématiques pures ne sont pas vraiment concernées car de véritables cours d'analyse numérique existent dans le but de former de vrais numériciens. Tout d'abord, il n'y a pas véritablement en France de cours dédié à la physique (ou plus généralement science) *computationnelle* au niveau Licence (du L1 au L3). On peut bien sûr trouver des enseignements d'informatique dont le but est d'apprendre la programmation, en utilisant tel ou tel langage. On trouve aussi parfois des cours de calcul scientifique (souvent en option suivant les filières), avec pour but d'initier les étudiants à l'utilisation de techniques numériques alternatives aux méthodes analytiques pour réaliser des opérations mathématiques (comme par exemple inverser une matrice). Dans ce dernier cas, l'enseignement peut parfois se résumer à savoir *lancer* des sous-programmes déjà écrits et disponibles dans une bibliothèque dédiée. La conséquence naturelle pour les étudiants poursuivant leurs études dans les Masters de physique, est de conduire à d'énormes difficultés dans la mise en oeuvre de la résolution de problèmes concrets. Par exemple, les étudiants sont incapables d'avoir un regard critique sur les résultats et ont tendance à avoir une confiance aveugle dans le programme informatique. Ceci rend alors impossible l'amélioration de leurs algorithmes voire même les corrections d'éventuelles erreurs. Deuxièmement, la recherche moderne en science est basée sur l'utilisation croissante de l'outil numérique. Il est ainsi d'autant plus nécéssaire que les étudiants en physique soient formés en amont afin d'éviter une utilisation de type boîte noire. C'est une situation que l'on retrouve hélas trop souvent car elle permet d'obtenir des résultats rapidement. Cet ouvrage s'adresse donc d'abord à des étudiants de Licence en sciences qui souhaitent s'initier aux méthodes numériques utilisées pour résoudre des problèmes standards de la physique. L'accent est mis particulièrement sur les spécificités inhérentes au calcul numérique par rapport à une démarche analytique par exemple. J'ai délibérément choisi de me restreindre à la résolution de problèmes de type dynamique.

Dans ce premier chapitre, nous introduirons uniquement deux méthodes numériques d'intégration que sont les schémas d'Euler et de Runge-Kutta. Ainsi, dans un premier temps, les propriétés (avantages et inconvénients) de ces méthodes seront illustrées pour intégrer des problèmes simples de type relaxation, qui se traduisent par des équations différentielles aux dérivées ordinaires du premier ordre. Dans un deuxième temps, nous comparerons ces méthodes pour la résolution de deux sytèmes dynamiques plus complexes et bien connus

que sont, le sytème de Lotka-Volterra, et le sytème chaotique de Lorenz. Nous verrons ainsi que le choix de la méthode numérique s'avère crucial en terme de stabilité, précision, mais aussi pour préserver certaines propriétés physiques importantes du sytème. En Annexe, le lecteur pourra trouver les différentes formules utilisées pour les schémas numériques ainsi que quelques calculs/rappels complémentaires utiles.

1 Equation de relaxation

Nous considérons d'abord un cas simple, qui est celui d'une équation différentielle linéaire dont la solution analytique est connue. Cela permettra de tester facilement les méthodes numériques en évaluant simplement les erreurs par différence entre la solution numérique et la solution théorique attendue.

1.1 Equation normalisée

l'équation de relaxation est un modèle physique de base pour représenter des mécanismes de retour à l'équilibre d'un état après avoir été excité. On peut citer par exemple le cas de la charge/décharge dans un circuit RC pour le courant électrique, ou encore la décroissance statistique du nombre de noyaux radioactifs au cours du temps en physique nucléaire.

On s'intéresse pour l'instant à l'équation différentielle:

$$\dot{y} + y/\tau = 0, \tag{1}$$

avec τ une constante de temps. $y(t)$ est la solution (par exemple une population de noyaux radioactifs) recherchée à l'instant t. On utilise la notation habituelle $\dot{y}$ pour la dérivée temporelle première $\frac{dy}{dt}$. L'équation précédente peut s'écrire en renormalisant le temps t avec $\tilde{t} = t/\tau$, conduisant alors à l'équation normalisée:

$$\dot{y} + y = 0, \tag{2}$$

où la dérivée temporelle $\dot{y}$ est maintenant par rapport à $\tilde{t}$. On a en effet utilisé $\frac{d}{d\tilde{t}} = \tau \frac{d}{dt}$. Cette renormalisation revient à exprimer le temps $\tilde{t}$ en unités de τ.

Par souci de simplification, on reprendra directement la notation t pour le temps normalisé dans la suite (ce qui revient aussi à prendre $\tau = 1$ dans l'équation originelle).

1.2 Solution analytique

La solution générale de l'équation précédente peut s'écrire simplement,

$$y(t) = y_0 \exp(-t), \tag{3}$$

avec la condition initiale donnée pour la solution $y(0) = y_0$ à l'instant initial $t = 0$. Il est important pour la physique comme pour le traitement numérique de trouver le(les) temp(s) caractéristique(s) du problème. Ici, la solution est simple, et conduit à une exponentielle décroissante sur une échelle de temps t_c caractéristique égale à 1 (voir figure 1). L'échelle de temps t_c est en effet définie comme l'inverse de la pente de la droite tangente à la solution à l'origine ($t = 0$), ou plus précisément,

$$t_c = -[\frac{1}{y}\frac{dy}{dt}]^{-1}, \tag{4}$$

pris à l'instant initial $t = 0$. Le signe 'moins' est introduit car la solution est décroissante avec le temps. Ce qui signifie que c'est aussi τ (comme attendu) pour l'équation initiale. Parfois on

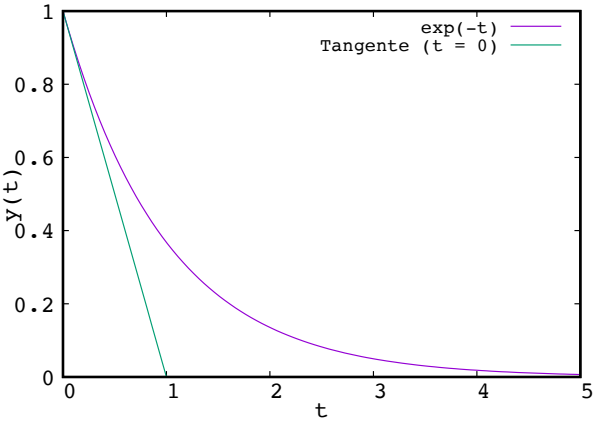

Figure 1: Solution de l'équation de relaxation normalisée en fonction du temps normalisé, pour une condition initiale $y(0) = 1$. La droite tangente à la solution à l'origine ($t = 0$) intersecte l'axe des abscisses pour un temps égal au temps caractéristique donc égal à 1 ($t_c = \tau = 1$).

utilise aussi le temps de demi-vie, $t_{1/2}$, le temps au bout duquel $y(t) = y_0/2$. On obtient alors $t_{1/2} = \tau Ln(2)$.

1.3 Une méthode numérique de base: le schéma d'Euler explicite (EE)

L'idée de base est d'approximer la dérive temporelle $\dot{y}$ par son expression dite aux différences finies, $[y(t + \Delta t) - y(t)]/\Delta t$. L'approximation sera d'autant meilleure que le pas de temps Δt sera petit (par définition de la dérivée). Nous pourrons alors construire un schéma numérique de type itératif:

$$y(t_{n+1}) = y(t_n) + \Delta t[f(y_n, t_n)], \tag{5}$$

avec $t_n = n\Delta t$ (n étant un indice entier), pour résoudre une équation générale du type $dy/dt = f[y(t), t]$. Ce schéma est la méthode d'Euler de type explicite (car la fonction f est évaluée explicitement au temps t_n). Ainsi, le but du jeu est de construire de proche en proche (tous les Δt, supposé constant pour simplifier) une solution numérique en partant de la condition initiale $y(t_0) = y_0$ connue. Cela revient à extrapoler linéairement la solution au temps t_{n+1}, $y(t_{n+1})$, à partir de la solution au temps précédent $y(t_n)$ supposée connue (voir figure 2).

La fonction étant $f(y) = -y$ pour notre équation de relaxation, le schéma EE devient alors simplement,

$$y_{n+1} = (1 - \Delta t)y_n, \tag{6}$$

pour $n = 0, 1, ...$, partant de $y_0 = y(0)$ donné, et avec la notation $y_n = y(t_n)$.

1.3.1 Mise en route et test du schéma EE

La mise en 'musique' est ultra-simple, en utilisant par exemple un programme informatique basé sur une boucle itérative. Partant de la condition initiale $y_0 = 1$, la solution ainsi obtenue y_n est

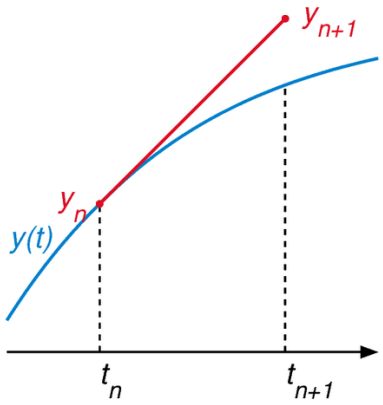

Figure 2: Principe de la méthode d'Euler explicite, basée sur une extrapolation linéaire utilisant la pente de la solution $f(y)$ au temps t_n supposée connue. La solution numérique obtenue y_{n+1} se situe au-dessus de la 'vraie' solution (trait bleu) pour une solution croissante avec le temps. (schéma tiré du site wikipédia)

visible sur la figure 3 pour deux pas de temps, ainsi que l'erreur absolue $e_n = \tilde{y}_n - y_n$ sur la figure 4, avec $\tilde{y}_n = y_0 \exp(-t_n)$ pour la solution exacte qui est évaluée en $t = t_n$.

On constate que la solution numérique se situe toujours en dessous (l'erreur est positive) de la solution attendue, et est donc 'en avance' en quelque sorte sur la solution qui tend asymptotiquement vers zéro. On peut dire que la solution 'va plus vite que la musique' par cette méthode, ou encore que **le schéma EE tend à amplifier la solution.** On constate aussi que l'erreur absolue maximum (obtenue pour $t = 1$) est divisée par 2 environ quand le pas de temps est divisé par 2. Ce qui signifie que l'erreur varie comme Δt, ou encore $e_n \propto \Delta t$. Ceci est en accord avec l'ordre attribué à la méthode d'Euler dans la litérature, qui est d'ordre 1. Une analyse mathématique plus fine est faite dans la suite. Il pourrait sembler étrange que l'erreur absolue re-diminue pour $t > 1$. En fait, l'erreur relative bien évidemment continue à augmenter continuellement avec le temps, et l'erreur absolue ne re-diminue que parce que la solution elle même tend vers zéro (panneau de droite de la figure 4).

Poussons la méthode dans ses derniers retranchements en explorant une gamme la plus large possible de pas de temps. On s'attend à des résultats d'autant meilleurs que le pas tend vers zéro et inversement quand le pas est proche de 1. Les résultats de l'erreur maximum pour une intégration jusqu'à un temps final $t_f = 1$ sont reportés sur la figure 5. Ainsi, l'ordre 1 de la méthode est parfaitement restitué. Par curiosité 'saine' (il faut être curieux dans l'utilisation du numérique), voyons maintenant le comportement pour un pas supérieur à 1, même si à priori la méthode à été développée pour un pas tendant vers zéro. Ceci est important pour bien comprendre les particularités de notre schéma. Les résultats obtenus avec un pas $1 < \Delta t < 2$ conduisent à une solution décroissante exponentiellement (en module) mais de signe alterné, alors que pour $\Delta t > 2$ la solution devient croissante en module (toujours de signe alterné) et tend carrément vers l'infini (figure 6). On dit que **le schéma EE devient instable.** Nous verrons plus loin plus en détails cette notion de stabilité.

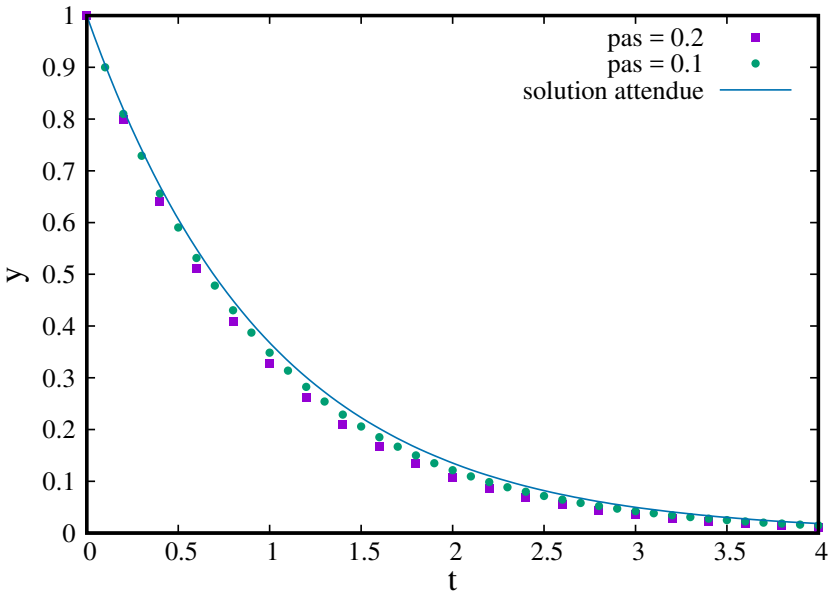

Figure 3: Solution numérique obtenue par le schéma EE pour deux pas de temps, $\Delta t = 0.2$, et 0.1, et comparée à la solution théorique attendue (en trait plein) en fonction du temps.

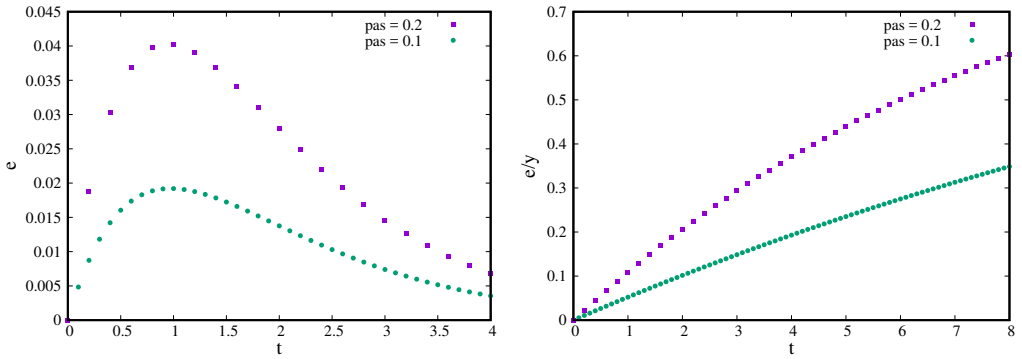

Figure 4: Erreur absolue $e_n = \tilde{y}_n - y_n$ obtenue sur la solution par le schéma EE pour deux pas de temps, $\Delta t = 0.2$, et 0.1, en fonction du temps (panneau de gauche). Erreur relative e_n/y_n correspondante (panneau de droite).

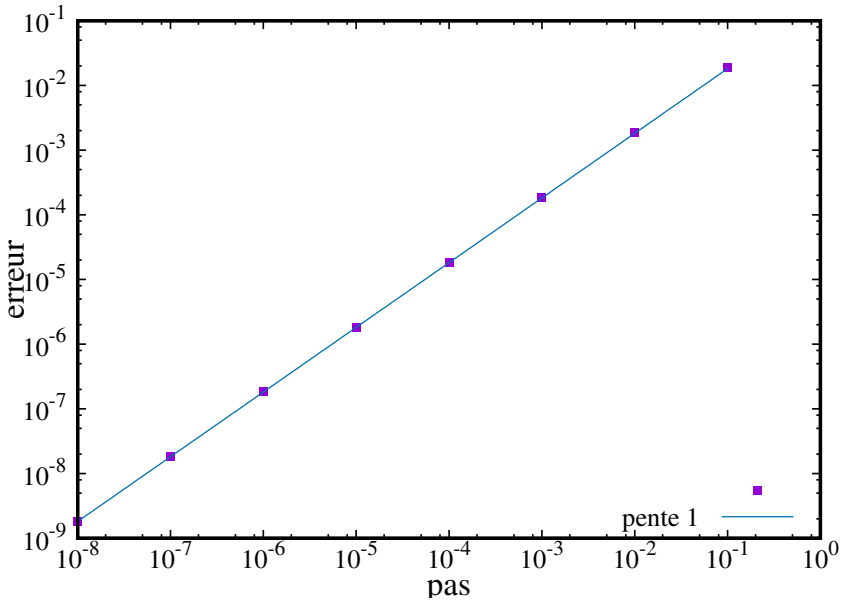

Figure 5: Erreur absolue mesurée à $t = 1$ pour la solution numérique obtenue avec le schéma EE, en fonction du pas de temps Δt. Une dépendance linéaire (droite de pente 1) est tracée par comparaison.

1.3.2 Analyse de la précision et de la stabilité du schéma EE

Essayons de comprendre les comportements observés ci-dessus, en utilisant un peu de mathématique (mais pas trop quand même). Nous avons vu que le schéma EE revenait à multiplier la solution à t_n par le facteur $1 - \Delta t$ pour obtenir la solution au pas suivant, c'est à dire y_{n+1} à t_{n+1} (voir équation 6). Ainsi, nous pouvons définir un facteur d'amplification (le terme va de soi sémantiquement),

$$k_{EE} = 1 - \Delta t, \tag{7}$$

associé au schéma EE. Nous pouvons aussi définir un facteur d'amplification théorique, $k_* = \exp(-\Delta t)$, obtenu simplement en injectant la solution attendue (équation 3) pour $\tilde{y}_{n+1}$ et $\tilde{y}_n$, conduisant alors à $\tilde{y}_{n+1} = \tilde{y}_n \exp(-\Delta t)$. Nous pouvons maintenant comparer ces deux coefficients.

Faisons d'abord un développement limité (DL) de k_* (voir annexe),

$$k_* = 1 - \Delta t + \frac{1}{2}\Delta t^2 - \frac{1}{6}\Delta t^3 + \frac{1}{24}\Delta t^4 + O(\Delta t^5), \tag{8}$$

conduisant alors à l'égalité,

$$k_{EE} = k_* + O(\Delta t^2). \tag{9}$$

Le schéma EE donne donc une erreur absolue d'ordre 2 et non pas d'ordre 1 ! Cherchez l'erreur (sans faire de jeux de mots). L'explication est simple, car l'expression ci-dessus signifie en effet que l'erreur sur le premier pas de temps (on part de la solution exacte $y_0 = \tilde{y}_0$), $e_1 = \tilde{y}_1 - y_1$ s'obtient de la façon suivante, $e_1 = (k_* - k_{EE})y_0 = O(\Delta t^2)y_0$. Ainsi nous obtenons une erreur dite locale d'ordre effectivement 2. Mais l'erreur obtenue dans les tests précédents est une erreur

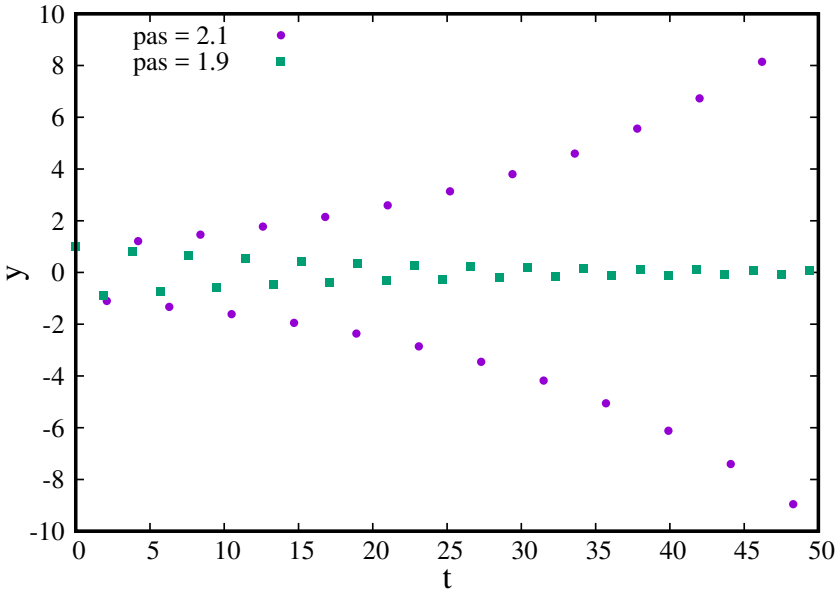

Figure 6: Solution numérique obtenue par le schéma EE pour deux pas de temps, $\Delta t = 2.1$, et 1.9. Les deux solutions sont oscillantes et la première est en plus instable.

cumulée sur un grand nombre de pas (on ne part plus de la bonne solution pour l'intégration sur un pas de temps quelconque). En effet, $y_n = y_0 k_{EE}^n = y_0(1-\Delta t)^n = y_0[1 - n\Delta t + O(\Delta t^2)]$, et $\tilde{y}_n = y_0 k_*^n = y_0 \exp(-n\Delta t)$, conduisent à $\tilde{y}_n - y_n = O(\Delta t^1)$. Pour ce faire, il faut utiliser le DL de $(1-\Delta t)^n$ et comparer au DL de $\exp(-n\Delta t)$ (voir formules en Annexe), en utilisant $n = t_n/\Delta t$. Le résultat final conduit alors à $e_n \approx y_0 \Delta t(t_n/2)$. On parle alors d'erreur globale qui est un ordre plus bas que l'erreur locale. On pouvait s'attendre à ce que la méthode d'Euler (la plus basique) soit d'ordre 1, car on remplace les dérivées par des droites (ordre le plus bas) pour extrapoler la solution (voir figure 2).

Avec cet éclairage nouveau, nous pouvons aussi comprendre le comportement obtenu pour $\Delta t > 1$. En effet, tout d'abord $\Delta t = 1$ donne un coefficient d'amplification k_{EE} nul, et pour $1 < \Delta t < 2$ le coefficient est négatif inversant donc la solution à chaque pas mais restant de module inférieur à 1 . Enfin, pour $\Delta t > 2$, le coefficient devient plus grand que 1 en module ($k_{EE} < -1$) rendant la solution numérique croissante exponentiellement donc instable.

1.4 Une autre méthode de base: le schéma d'Euler implicite (EI)

Une alternative évidente au schéma EE consiste à évaluer la fonction f (dans l'équation 5) au temps t_{n+1} au lieu de t_n. Cela revient à utiliser la pente $f(y(t), t)$ au temps t_{n+1} (temps auquel on cherche justement la solution), pour extrapoler la solution y_{n+1} (voir figure 2), d'où le terme implicite car la pente dépend implicitement de la solution cherchée.

Le schéma d'Euler implicite (EI) appliqué à notre équation de relaxation devient alors,

$$y_{n+1} = y_n/(1+\Delta t). \tag{10}$$

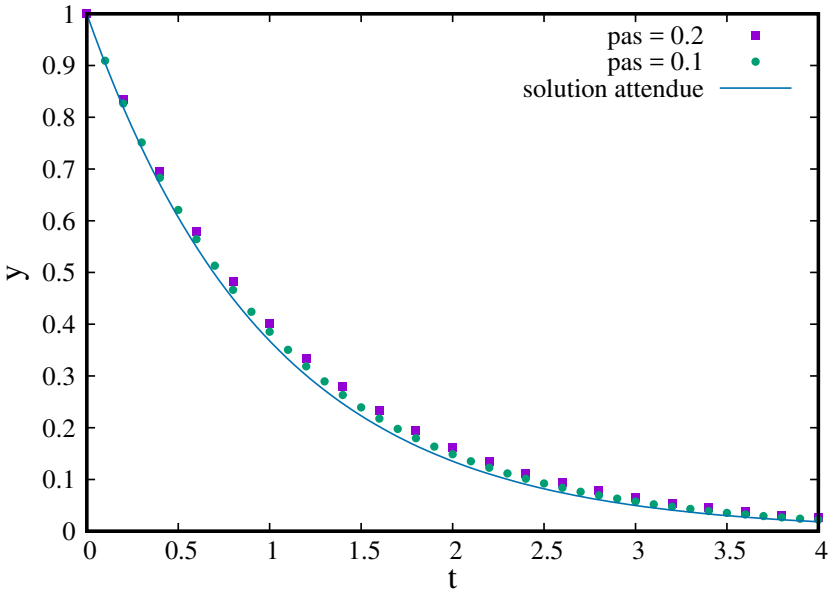

Figure 7: Solution numérique obtenue par le schéma EI pour deux pas de temps, $\Delta t = 0.2$, et 0.1, et comparée à la solution théorique attendue (en trait plein) en fonction du temps.

1.4.1 Tests du schéma EI

Nous pouvons faire les mêmes tests que ceux effectués pour le schéma EE. Les résultats sont visibles sur les figures 7-8. Maintenant, on peut dire que la solution numérique est 'en retard' sur la solution théorique, et donc elle 'va moins vite que la musique' mais avec une amplitude d'erreur comparable au schéma EE.**Le schéma EI est aussi dit amortissant.** L'erreur étant divisée par 2 quand le pas est divisé par 2, le schéma EI est donc aussi du premier ordre.

1.4.2 Précision et stabilité du schéma EI

De la même façon que pour le schéma EE, il est facile de vérifier l'ordre du schéma sur une large gamme de valeurs du pas de temps Δt (non montré), et pouvant s'étendre maintenant à des valeurs plus grandes que 1. Le schéma EI est en effet inconditionnellement stable contrairement au schéma EE. Ceci se comprend aisément en introduisant le coefficient d'amplification correspondant, $k_{EI} = 1/(1 + \Delta t)$. Ainsi (utilisant un développement limité visible en Annexe),

$$k_{EI} = 1 - \Delta t + \Delta t^2 + O(\Delta t^3), \tag{11}$$

permet d'obtenir $k_{EI} = k_* + O(\Delta t^2)$ (car le terme en Δt^2 n'est pas correct) et donc de vérifier un ordre 2 local (ordre 1 global). On comprend aussi aisément que le module de k_{EI} étant toujours inférieur à 1, la solution converge toujours vers zéro. On dit que **le schéma EI est inconditionellement stable**, contrairement au schéma EE qui est conditionnellement stable (pas de temps suffisamment petit).

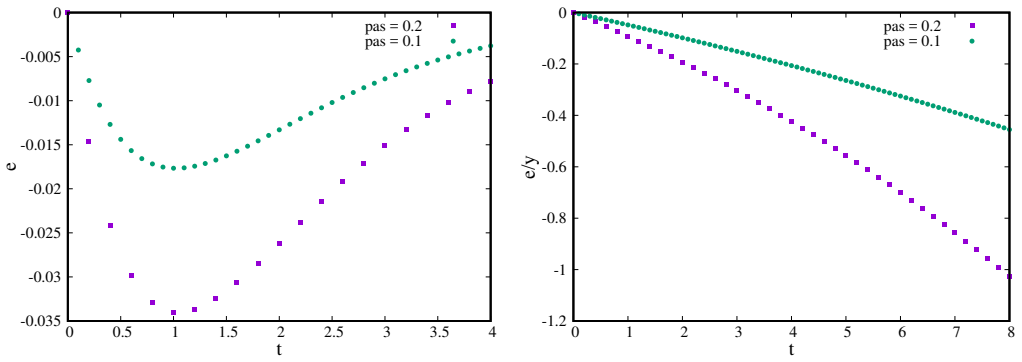

Figure 8: Erreur absolue e obtenue sur la solution par le schéma EI pour deux pas de temps, $\Delta t = 0.2$, et 0.1 en fonction du temps (panneau de gauche). Erreur relative e/y correspondante (panneau de droite).

1.5 Une méthode alternative efficace: le schéma d'Euler centré (EC)

Dans la continuation de la philosophie précédente, nous pouvons utiliser une moyenne $(f_n + f_{n+1})/2$ pour évaluer la pente qui permet d'extrapoler la solution (figure 2), expliquant ainsi le terme centré. Cette méthode est aussi parfois appellée formule trapezoidale, ou méthode du point-milieu. Ce centrage devrait ainsi améliorer la précision, ce que nous verrons plus loin.

Le schéma EC appliqué à notre équation de relaxation devient alors,

$$y_{n+1} = y_n \frac{1 - \Delta t/2}{1 + \Delta t/2}. \tag{12}$$

1.5.1 Tests du schéma EC

Nous pouvons faire les mêmes tests que ceux effectués pour les schéma EE et EI. Les résultats sont visibles sur les figures 9-10. Maintenant, nous pouvons constater que l'erreur absolue (maximum) est divisée par 4 environ lorsque le pas est divisée par 2 (figure 9), conduisant à conclure que ce dernier schéma est d'ordre 2 (on parle d'ordre global, car l'ordre local sera alors de 3 selon la discussion tenue pour le schéma EE). Ceci peut se vérifier en testant sur une large gamme de Δt compris entre 10^{-5} et 10^{-1}(voir figure 10). Cependant, pour un pas plus petit que 10^{-5}, l'erreur semble suivre une autre loi d'échelle sur laquelle nous reviendrons plus loin (voir méthodes de type Runge-Kutta).

Pour des pas de temps supérieurs à 2, la solution numérique devient alternée à chaque pas tout en restant stable comme pour le schéma EI.

1.5.2 Précision et stabilité du schéma EC

Le coefficient d'amplification pour le schéma EC est, $k_{EC} = \frac{1-\Delta t/2}{1+\Delta t/2}$, conduisant à,

$$k_{EC} = 1 - \Delta t + \frac{1}{2}\Delta t^2 - \frac{1}{4}\Delta t^3 + O(\Delta t^4), \tag{13}$$

et $k_{EC} = k_* + O(\Delta t^3)$. Ce qui confirme l'odre local de 3 et global de 2 pour le schéma EC. On comprend aussi aisément que le module de k_{EC} étant toujours inférieur à 1, **cette méthode**

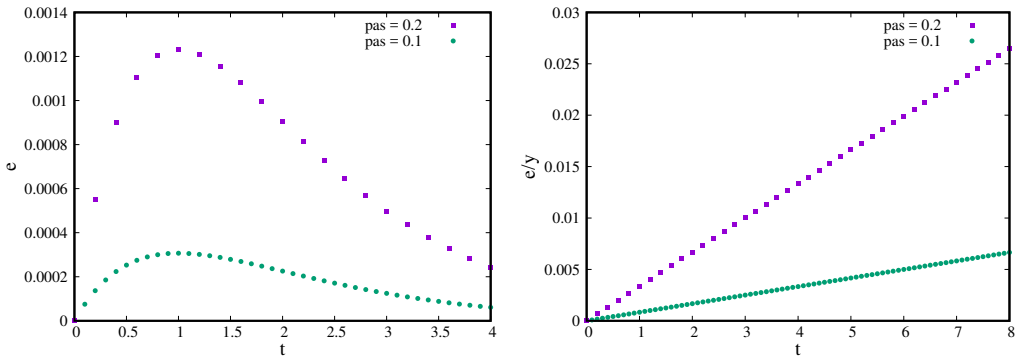

Figure 9: Erreur absolue e obtenue sur la solution par le schéma EC pour deux pas de temps, $\Delta t = 0.2$, et 0.1, en fonction du temps (panneau de gauche). Erreur relative e/y correspondante (panneau de droite).

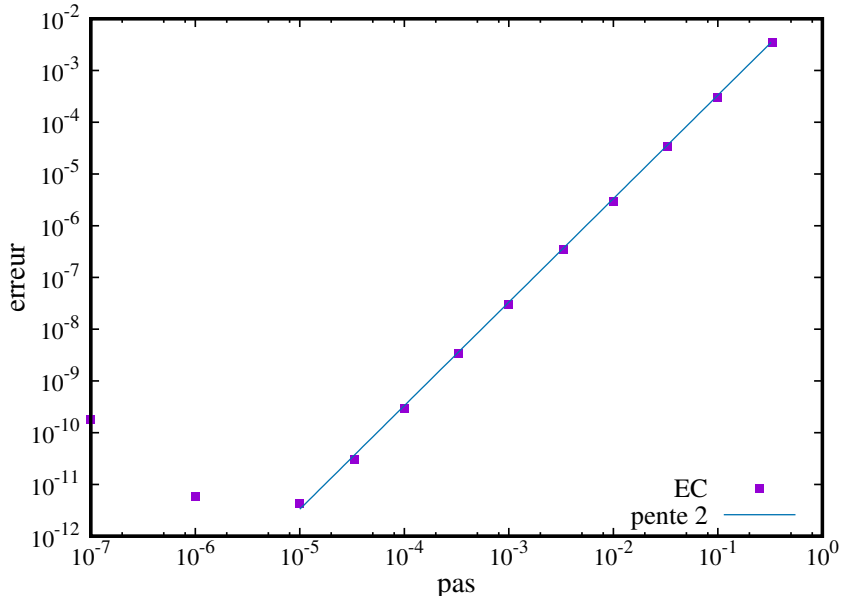

Figure 10: Erreur absolue maximum mesurée à $t = 1$ pour la solution numérique obtenue avec le schéma EC, en fonction du pas de temps Δt. Une dépendance parabolique (droite de pente 2) est tracée par comparaison.

est inconditionnellement stable comme pour le schéma implicite. Ce résultat n'est pas entièrement surprenant, vu le terme du coefficient d'amplification en $1/(1+\Delta t/2)$ traduisant le caractère partiellement implicite et conduisant ainsi à un amortissement partiel.

1.6 Les méthodes de type Runge-Kutta (RK)

Les méthodes d'Euler de type partiellement ou totalement implicites (EI et EC) ne sont pas faciles à implémenter lorsque $f(y)$ n'est pas linéaire par rapport à y, car on ne peut pas alors déduire une formule simple donnant y_{n+1} en fonction de y_n (comme c'était le cas pour les schémas EE et EI). On préfère alors **des méthodes de type Runge-Kutta qui sont pour la plupart conditionnellement stables** mais qui ont l'avantage d'avoir des ordres élevés. Il ne s'agit pas ici de faire le 'listing' de toutes les méthodes de type Runge-Kutta existantes dans la litérature, mais de comprendre la philosophie en utilisant les deux méthodes classiques RK les plus connues.

1.6.1 Principe des méthodes de Runge-Kutta

Nous avons vu que pour les méthodes d'Euler, il fallait estimer la pente $f[y(t),t]$ pour extrapoler la solution en t_{n+1} en partant de la solution en t_n (figure 2). Les méthodes de Runge-Kutta partent du même principe, mais l'évaluation de la pente est faite différemment. En effet, la pente est calculée à partir de plusieurs estimations (le nombre dépendant de l'ordre voulu) de la solution à des temps intermédiaires entre t_n et t_{n+1}. On parle de **méthode à pas unique** contrairement à d'autres méthodes dont on ne parlera pas ici qui sont des méthodes à pas multiples.

Par exemple, pour la méthode classique RK d'ordre 2 (RK2), on évalue une solution intermédiaire en $t_{n+\frac{1}{2}} = t_n + \frac{\Delta t}{2}$, par $y_{n+\frac{1}{2}} = y_n + \frac{\Delta t}{2} f(y_n, t_n)$, qui est une formule d'Euler explicite obtenue avec un demi-pas $\Delta t/2$. La pente obtenue en $t_{n+\frac{1}{2}}$ permet alors d'extrapoler la solution en t_{n+1} par,

$$y_{n+1} = y_n + \Delta t f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}). \tag{14}$$

La plus connue est sans conteste la méthode RK d'ordre 4 (RK4). En pratique, il existe un grand nombre de formules de Runge-Kutta d'ordre 4, mais on se limitera ici à la méthode classique. On a besoin de trois évaluations intermédiaires (deux en $t_{n+\frac{1}{2}}$ et une en t_{n+1}), qui sont $y_{n+\frac{1}{2},1}$, $y_{n+\frac{1}{2},2}$, et $y_{n+1,1}$, en plus de y_n. Celles ci sont obtenues avec,

$$y_{n+\frac{1}{2},1} = y_n + \frac{\Delta t}{2} f(y_n, t_n), \tag{15}$$

$$y_{n+\frac{1}{2},2} = y_n + \frac{\Delta t}{2} f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}), \tag{16}$$

$$y_{n+1,1} = y_n + \Delta t f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}). \tag{17}$$

La formule d'extrapolation finale est alors,

$$y_{n+1} = y_n + \frac{\Delta t}{6}[f(y_n,t_n) + 2f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}) + 2f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}) + f(y_{n+1,1}, t_{n+1})]. \tag{18}$$

On a ainsi une pente moyenne calculée comme la moyenne de 4 pentes évaluées à 4 instants différents et avec des poids statistiques différents. On remarquera que 4 évaluations intermédiaires sont nécéssaires pour la méthode RK4 et 2 pour la RK2. Plus d'informations sont disponibles en Annexe pour le lecteur qui le souhaite.

1.6.2 Tests de RK2 et RK4 sur l'équation de relaxation

Contrairement au pragmatisme affiché pour les méthodes d'Euler (tests avant d'analyser le comportement), commençons d'abord par examiner l'ordre et la stabilité. L'application de la formule de RK2 donne,

$$y_{n+1} = (1 - \Delta t + \Delta t^2/2)y_n, \tag{19}$$

qui permet ainsi d'obtenir le coefficient d'amplification correspondant,

$$k_{RK2} = 1 - \Delta t + \Delta t^2/2 = k_* + O(\Delta t^3). \tag{20}$$

L'ordre local de RK2 sera donc 3 et l'ordre global sera 2 comme attendu. On vérifie aussi que pour RK4,

$$k_{RK4} = 1 - \Delta t + \Delta t^2/2 - \Delta t^3/6 + \Delta t^4/24, \tag{21}$$

et que donc, $k_{RK4} = k_* + O(\Delta t^5)$. L'ordre local de RK4 sera donc 5 et l'ordre global sera 4 comme attendus.

On peut d'abord s'assurer que l'ordre attendu pour chaque schéma est atteint. L'erreur absolue pour deux pas de temps ($\Delta t = 0.2$ et $\Delta t = 0.1$) est visible sur la figure 11 pour RK2 (panneau de gauche) er RK4 (panneau de droite). Nous pouvons aussi explorer l'erreur pour une large gamme de valeurs pour le pas (figure 12). Revenons sur le comportement surprenant lorsque le pas est trop petit (voir figure 10 pour $\Delta t < 10^{-5}$) pour le schéma EC, et que nous retrouvons clairement avec RK4 pour $\Delta t < 10^{-3}$ sur la figure 12. L'erreur augmente approximativement linéairement quand le pas diminue. Cet effet n'est pas lié à la méthode numérique mais aux erreurs d'arrondis. En effet, chaque nombre est entaché d'une erreur (dite erreur d'arrondi) associé au codage binaire. Les opérations sur ces nombres conduisent alors à une dégradation de l'erreur de base qui donne une erreur sur le résultat final proportionnelle au nombre d'opérations. Les erreurs d'arrondis deviennent alors dominantes sur les erreurs du schéma (dites de troncature) quand le pas devient suffisamment petit. Les erreurs d'arrondis ont aussi une allure très particulière (moins régulières) que les erreurs de troncature en fonction du temps (voir figure 13). On voit très bien sur cette figure comment la transition d'erreur de troncature dominante à erreur d'arrondi dominante se fait progressivement quand le pas diminue. Bien sûr, il est possible d'abaisser le niveau minimum des erreurs d'arrondi (ici 10^{-15} environ) en choisissant un codage binaire sur plus de chiffres significatifs (voir aussi en Annexe).

Examinons maintenant la stabilité des deux schémas. Il suffit d'écrire les conditions requises pour les coefficients d'amplifications (k_{RK2} et k_{RK4}) qui doivent avoir leur module inférieur à l'unité. **Ainsi, on peut vérifier que les deux schémas sont conditionnellement stables pour, $\Delta t < 2$ et $\Delta t < 2.785$ pour RK2 et RK4 respectivement.**

1.7 Comparatif des méthodes par leurs coefficients d'amplification

On peut résumer les différentes méthodes introduites en utilisant les coefficients d'amplification calculés pour l'équation de relaxation et en les comparant au coefficient idéal. Les résultats sont traçés sur la figure 14. Le lecteur peut légitimement se poser la question de l'intérêt de déterminer les conditions de stabilité des méthodes, puisque l'on utilisera de toute façon des pas qui sont toujours très petits devant les pas limites pour des raisons de précisions minimales requises. Ceci est vrai pour l'exemple choisi (cas d'une équation simple scalaire), mais moins évident pour un cas multi-dimensionnel avec plusieurs échelles de temps caracteristiques. C'est pourquoi dans la suite nous allons prendre des cas plus complexes que le sytème de relaxation mono-dimensionnelle. Enfin, nous noterons que les schémas conditionnellement stables sont assujettis a une condition de type $\Delta t < \tau$, avec τ un temps caractéristique de la solution, qui vaut 2 pour le schéma EE par exemple.

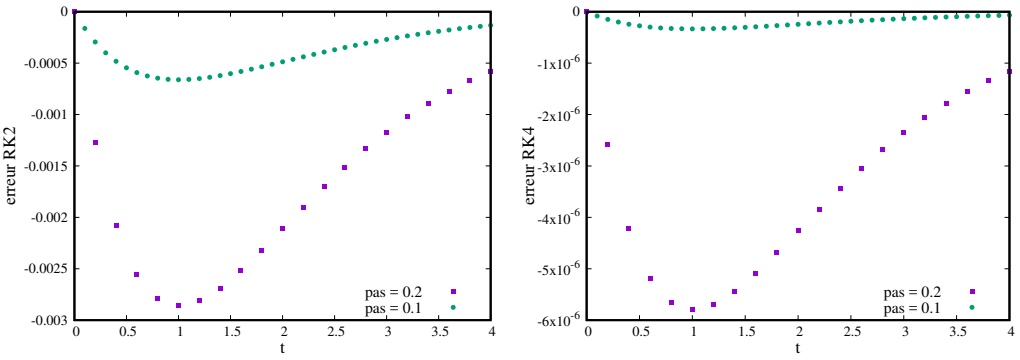

Figure 11: Erreur absolue e obtenue sur la solution par le schéma RK2 pour deux pas de temps, $\Delta t = 0.2$, et 0.1 en fonction du temps (panneau de gauche), et par le schéma RK4 (panneau de droite).

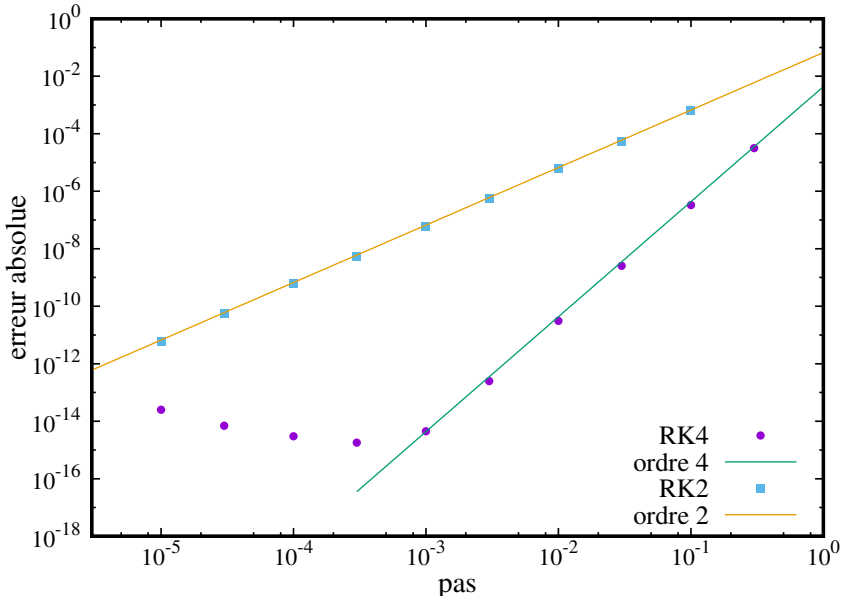

Figure 12: Erreur absolue e maximum (mesurée à $t = 1$) obtenue en fonction du pas de temps, en utilisant les schémas RK4 et RK2. Les dépendances d'ordre 2 et 4 sont tracées en trait plein.

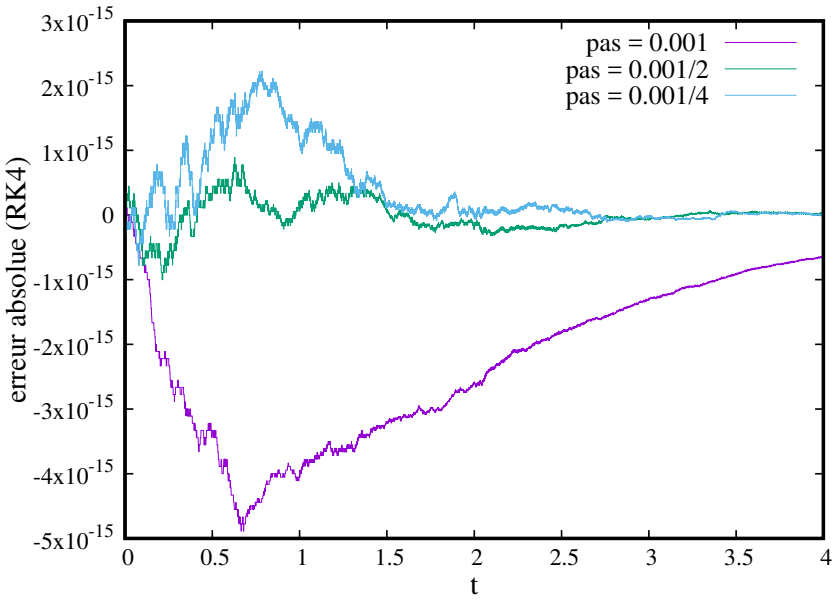

Figure 13: Erreur absolue e en fonction du temps en utilisant le schéma RK4, avec trois pas de temps relativement petits, $\Delta t = 10^{-3}$, 5×10^{-4}, et 2.25×10^{-4}.

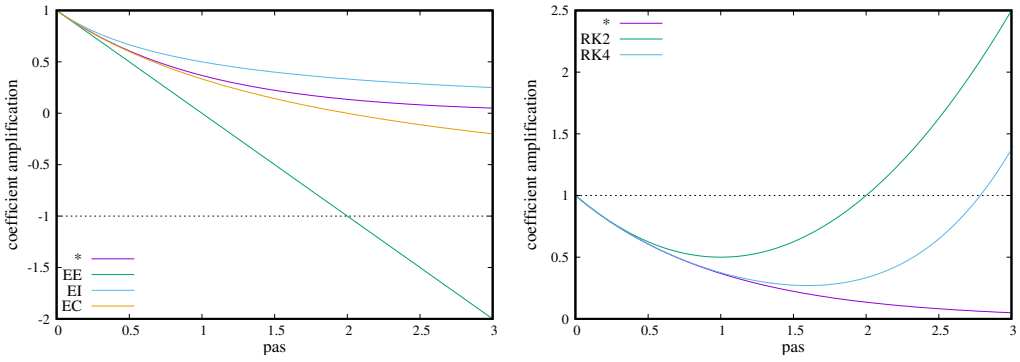

Figure 14: Coefficient d'amplification des différentes méthodes d'Euler (EE-EI-EC) sur le panneau de gauche et de Runge-Kutta (RK2-RK4) sur le panneau de droite. Le coefficient théorique ($*$) est aussi indiqué pour comparaison, et un trait horizontal discontinu indique la limite de stabilité (coefficient d'amplification égal à l'unité en module).

2 Equation de relaxation bi-dimensionnelle

Nous considérons maintenant un système de deux équations différentielles couplées de type relaxation,

$$\dot{y} = ay + bz, \tag{22}$$

$$\dot{z} = cy + dz, \tag{23}$$

dont nous cherchons les solutions $y(t)$ et $z(t)$ à partir des conditions initiales $y(0) = y_0$ et $z(0) = z_0$ données. Les 4 coefficients (a, b, c, d) sont des nombres réels donnés. Ce système peut se mettre sous une forme plus compacte,

$$\dot{\vec{Y}} = A\vec{Y}, \tag{24}$$

avec les définitions $\vec{Y} = \begin{pmatrix} y \\ z \end{pmatrix}$, et la matrice 2×2 A définie par $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Toutes les méthodes vues auparavant peuvent s'appliquer sur ce système.

2.1 Prenons sur un exemple

Soyons pragmatique et commençons par prendre un exemple. Soit le système suivant avec $A = \begin{pmatrix} -6 & 10 \\ 2 & -5 \end{pmatrix}$, qui n'est pas choisi totalement au hasard mais de façon à obtenir des solutions de type relaxation. Nous avons donc besoin de chercher la solution analytique. Ceci est fait en Annexe, en calculant un système aux valeurs propres. Les valeurs propres de la matrice A sont -1 et -10. Et les vecteurs propres (définis à une constante multiplicative près) sont donnés par, $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ associé à -1, et $\begin{pmatrix} -5/2 \\ 1 \end{pmatrix}$ associé à -10. Ainsi, on peut vérifier qu'une solution analytique est,

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 2\exp(-t) - \frac{5}{2}\exp(-10t) \\ \exp(-t) + \exp(-10t) \end{pmatrix}, \tag{25}$$

avec comme condition initiale $\begin{pmatrix} -1/2 \\ 2 \end{pmatrix}$. Par analogie avec la solution de relaxation scalaire, nous devinons que les deux valeurs propres introduisent deux constantes de relaxation (deux temps caractéristiques) qui valent $\tau_1 = 1$ et $\tau_2 = 1/10$ respectivement.

Il ne s'agit plus de tester toutes les méthodes vues avant, mais de tester l'effet multi-dimensionnel. Nous nous limitons donc au test du schéma EE. Les résultats sont reportés sur les figures 15-16. Tout d'abord, nous vérifions que l'ordre attendu (ordre 1) est bien obtenu. Ensuite, il faut remarquer que l'erreur maximum (pour y comme pour z) est maintenant obtenue pour $t = 0.1$, qui correspond à l'échelle de temps la plus petite $\tau_2 = 0.1$. On s'attend donc à ce que les propriétés obtenues pour l'équation de relaxation scalaire restent valables pour ce cas bi-dimensionnel à condition d'exprimer le pas de temps en fonction du temps de relaxation le plus petit qui est τ_2. Et c'est le cas aussi de la condition de stabilité qui pour le schéma EE devient $\Delta t < 2\tau_2$, et donc $\Delta t < 0.2$. Ce que nous avons pu vérifié. Dans ce cas, cette condition de stabilité est en fait une condition nécéssaire (mais pas suffisante) car la matrice d'amplification n'est pas normale (voir Annexe pour plus de précisions). Nous avons aussi vérifié que les solutions sont oscillantes pour $\tau_2 < \Delta t < 2\tau_2$. **C'est donc la valeur propre la plus grande en module qui détermine les propriétés de stabilité du schéma** (ou encore le rayon spectral de la matrice d'amplification qui est détaillée en Annexe). Pour terminer, nous pouvons illustrer l'efficacité du schéma RK4 sur ce système en utilisant le même pas de temps $\Delta t = 0.03$ que précédemment avec RK2. La solution obtenue sur la figure 17 montre l'excellent accord avec la solution attendue malgré ce pas relativement grossier.

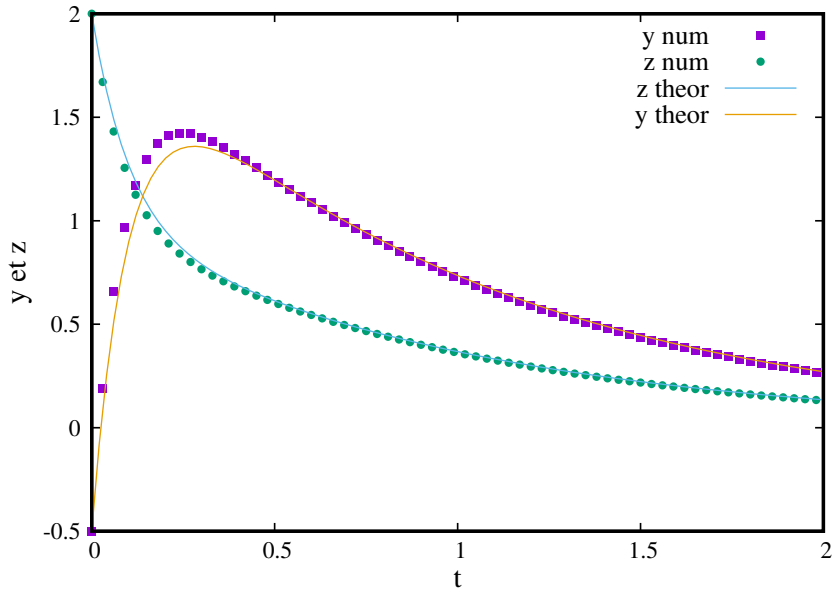

Figure 15: Solutions numériques ($y(t)$ et $z(t)$, en points carrés et ronds) obtenues par le schéma EE en fonction du temps, pour un pas de temps de $\Delta t = 0.03$, et solutions analytiques en trait plein.

2.2 Conclusion sur le choix optimal de la méthode

Pour le type de problème traité ici (relaxation), les méthodes de Runge-Kutta classiques sont parfaitement adaptées, et le schéma RK4 offre un bon compromis entre facilité de programmation, précision, et stabilité.

3 Application au système non-linéaire le Lotka-Volterra

Amusons nous un peu maintenant avec le système de Lotka-Volterra, qui représente l'évolution temporelle de deux populations, celle des proies $y(t)$ et celle des prédateurs $z(t)$. Ce modèle avait été initialement introduit pour expliquer la diminution de poissons à Trieste pendant la première guerre mondiale. Le bureau des pêches avait constaté que la population de requins avait considérablement augmenté par rapport aux poissons de type sardines. Volterra modélisa alors le système requins-sardines avec deux équations différentielles. Ce système obéit à,

$$\dot{y} = ay(1 - \frac{z}{d}), \tag{26}$$

$$\dot{z} = -bz(1 - \frac{y}{c}), \tag{27}$$

avec a le taux de croissance des proies en l'absence de prédateurs, et b le taux de décroissance des prédateurs en l'absence de proies. Le taux de croissance des prédateurs au détriment des proies est by/c, alors que le taux de destruction des proies par les prédateurs est az/d. Les quatre paramètres (a, b, c, d) sont réels et positifs.

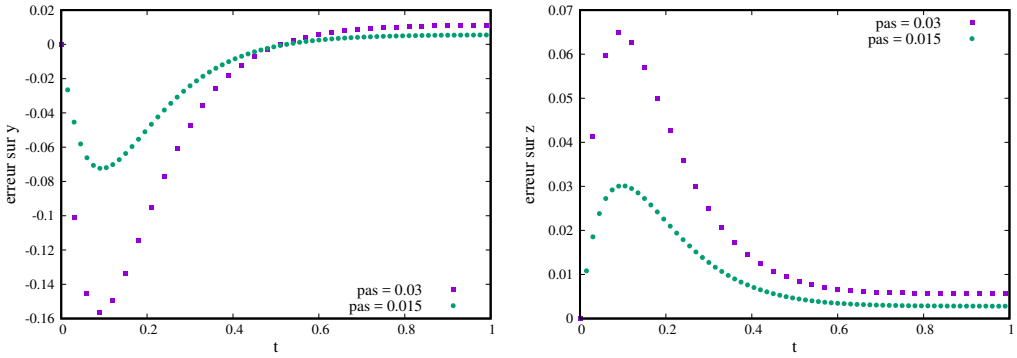

Figure 16: Erreur absolue sur y (tableau de gauche) et sur z (tableau de droite) en fonction du temps pour deux intégrations obtenues avec les deux pas de temps, $\Delta t = 0.03$, et 0.015.

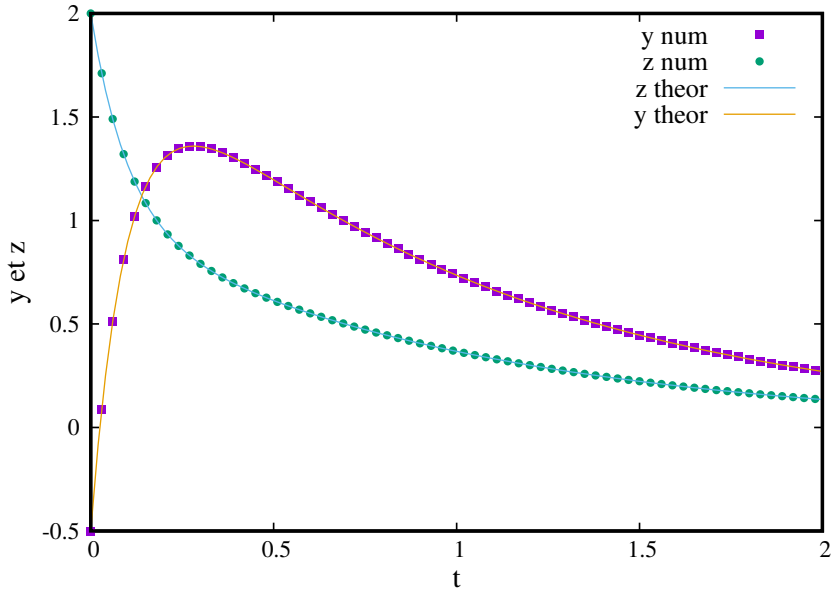

Figure 17: Solutions numériques ($y(t)$ et $z(t)$, en points carrés et ronds) obtenues par le schéma RK4 en fonction du temps, pour un pas de temps de $\Delta t = 0.03$, et solutions analytiques en trait plein.

Les solutions analytiques sont périodiques mais ne peuvent hélas pas s'exprimer simplement. Cependant, on peut montrer qu'elles possèdent une propriété importante dans l'espace (y, z) que l'on peut appeler l'espace des phases. Les trajectoires dans cet espace sont en effet fermées, en vertu de la relation existante suivante entre y et z,

$$y^b z^a \exp(-a\frac{z}{d} - b\frac{y}{c}) = constante, \tag{28}$$

qui est obtenue en faisant le rapport des deux équations différentielles et en intégrant. Ceci se traduit aussi par la **conservation du volume (la surface dans le cas présent) dans l'espace des phases**, car les trajectoires déterminées par 4 conditions initiales (définissant ainsi une surface initiale) conserveront au cours du temps la surface engendrée. On pourrait aussi relier la constante à un **invariant du mouvement** en utilisant la théorie des systèmes dynamiques (mais cela nous emmènerait trop loin de nos objectifs de ce chapitre). Ces notions seront abordées plus amplement dans le chapitre suivant. Par contre, nous allons nous intéresser à tester l'aptitude des schémas à reproduire ou pas cette propriété, qui s'avère fondamentale pour l'équilibre entre les deux populations.

3.1 Tests de plusieurs schémas

Prenons le cas $c = d = a = 1$, et $b = 0.2$ assorti des conditions initiales, $y(0) = 10$ et $z(0) = 1$. Nous avons donc au début 10 proies pour 1 prédateur. Et, commençons par utiliser le schéma le plus simple (cas EE). Quel que soit le pas de temps employé, les solutions (voir Figures 18-19) montrent des maxima qui s'amplifient continuellement (linéairement) au cours du temps et ne sont donc pas rigoureusement périodiques. De plus, la solution vue dans l'espace des phases n'est pas fermée. L'amplitude du cycle s'agrandit au cours du temps d'autant plus que le pas de temps est grand, en accord avec l'amplification citée ci-dessus. Ainsi, les solutions (et surtout la propriété principale) ne sont pas bien reproduites par notre schéma, sauf pour l'utilisation d'un pas tout petit et à condition de ne pas intégrer trop longtemps. Ce comportement sur un système de type périodique n'est pas surprenant si on se souvient que la méthode d'Euler explicite amplifie et 'va plus vite que la musique'. Nous donnerons une explication plus rigoureuse, basée sur une analyse plus fine plus loin.

Continuons par l'utilisation d'une Runge-Kutta (RK2) sur ce problème. Les résultats sont visibles sur la figure 20, en employant deux pas de temps ($\Delta t = 0.1$ et $\Delta t = 0.001$). La solution pour le pas de temps le plus petit reproduit en apparence correctement la solution attendue et servira donc de cas de référence. Nous constatons maintenant un amortissement de la solution $y(t)$ pour le pas le plus grand, qui se traduit par une trajectoire de cycle non fermée dont l'aire encerclée diminue au cours du temps dans l'espace des phases. La solution $z(t)$ suit une tendance amortissante similaire (pas montrée ici). De plus, la solution se déphase au cours du temps (voir la comparaison des solutions sur le panneau de gauche de la figure). L'emploi d'un schéma d'ordre plus élevé (RK4 par exemple) ne changera pas cette tendance même si l'ampleur sera grandement plus faible pour un pas de temps donné et à peine observable si le temps d'intégration n'est pas trop long (voir plus loin).

Les schémas EI et EC ne sont pas implémentables tels quels pour ce problème non linéaire. Il nous faut donc trouver un autre schéma. L'idée est de bénéficier de l'amortissement des schémas de type implicite tout en gardant un caractère explicite pour compenser. Ainsi, il est possible d'utiliser le schéma suivant,

$$y_{n+1} = y_n + a y_n \Delta t(1 - z_n/d), \tag{29}$$

$$z_{n+1} = z_n - b z_n \Delta t(1 - y_{n+1}/c), \tag{30}$$

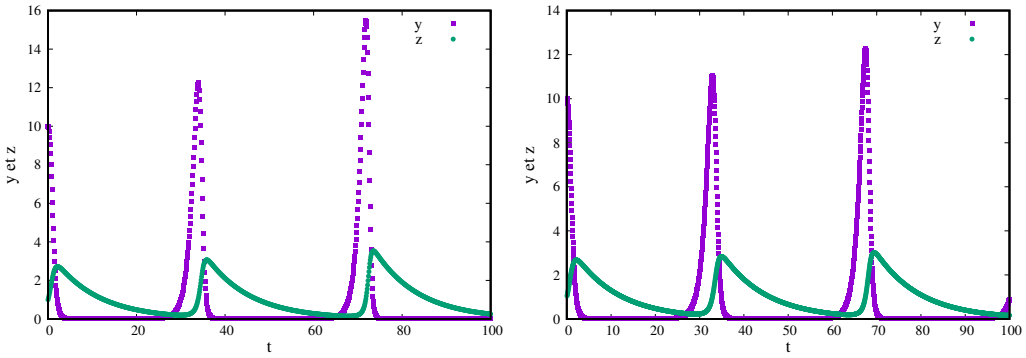

Figure 18: Evolution des populations de proies $y(t)$ et de prédateurs $z(t)$ au cours du temps obtenue par le schéma EE, avec deux pas de temps $\Delta t = 0.1$ (panneau de gauche) et $\Delta t = 0.05$ (panneau de droite).

qui utilise une formule de type Euler explicite pour la première équation et une formule de type Euler implicite (en partie avec y_{n+1} au lieu de y_n utilise pour le second membre) pour la seconde équation. Cette implicitation partielle est rendue possible car y_{n+1} a été obtenu juste avant. Nous appellerons ce schéma, **Euler semi-implicite ou ESI**. Il est aussi possible d'utiliser la variante suivante,

$$y_{n+1} = y_n + a y_{n+1} \Delta t (1 - z_n/d), \tag{31}$$

$$z_{n+1} = z_n - b z_n \Delta t (1 - y_{n+1}/c), \tag{32}$$

qui donne des résultats similaires (la différence est le traitement légèrement différent de la première équation qui doit être inversée). Les résultats obtenus en employant les mêmes paramètres que pour le cas précédent sont visibles sur la figure 21. Cette fois-ci, la trajectoire suit bien une courbe fermée dans l'espace des phases et il n'y a plus d'amplification ou d'amortissement (pour le pas le plus grand). Cette méthode possède donc une propriété supplémentaire que n'ont pas les autres méthodes y compris les Runge-Kutta classiques. Cette propriété que nous verrons plus en détails plus loin concerne la notion de **schéma symplectique**. Bien sûr il n'y a pas de méthode miracle, et l'erreur de déphasage observée précédemment est encore présente (celle ci sera diminuée avec un pas de temps plus petit). Enfin, l'emploi d'une Runge-Kutta d'ordre relativement élevé (RK4 par exemple) reste très correcte même avec un pas de temps grossier ($\Delta t = 0.1$), car la différence avec une trajectoire fermée reste très petit et n'est pas visible, comme on peut le constater sur la figure 22.

3.2 Conclusion sur le choix optimal de la méthode

Contrairement aux problèmes traités précédemment, les méthodes de Runge-Kutta classiques ne sont pas les plus adaptées lorsque la solution requiert la propriété supplémentaire de conservation du volume dans l'espace des phases. Ici, cela se traduit par un déséquilibre entre les deux populations qui se traduit alors par la disparition par exemple sur le long terme des populations (cas d'un schéma trop amortissant). Ainsi, **un schéma dit symplectique est préférable comme le schéma ESI**. On verra dans le chapitre suivant qu'il est possible d'utiliser des shéma symplectiques d'ordre élevé.

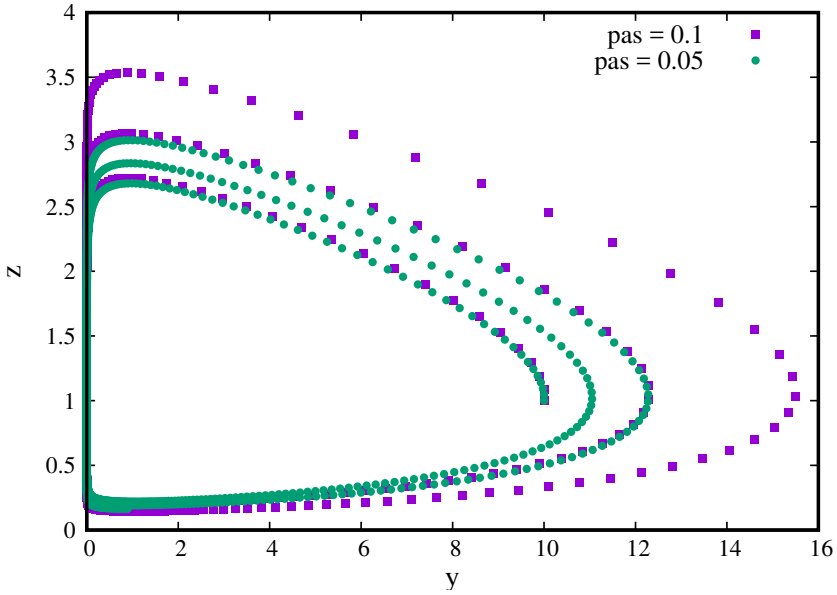

Figure 19: Evolution des populations dans l'espace des phases (y, z), obtenu pour les deux pas de temps de la figure précédente utilisant EE ($\Delta t = 0.1$ et $\Delta t = 0.05$).

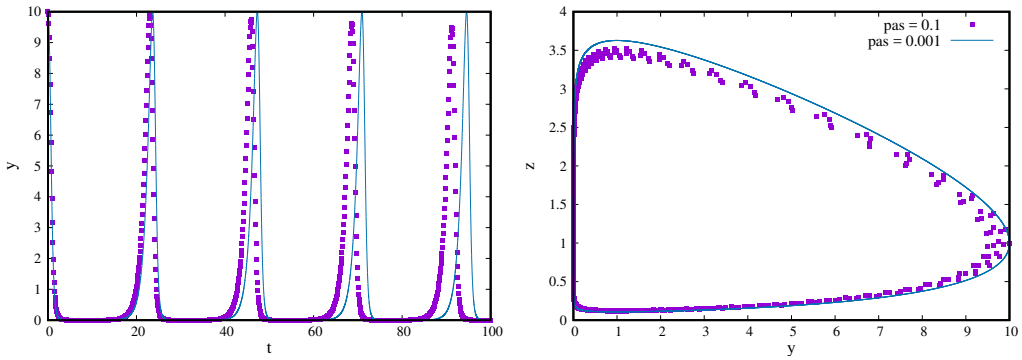

Figure 20: Evolution des populations de proies $y(t)$ au cours du temps (panneau de gauche) obtenue par le schéma RK2, avec deux pas de temps $\Delta t = 0.1$ (points) et $\Delta t = 0.001$ (trait plein). Trajectoire correspondante $y(z)$ (panneau de droite) dans l'espace des phases.

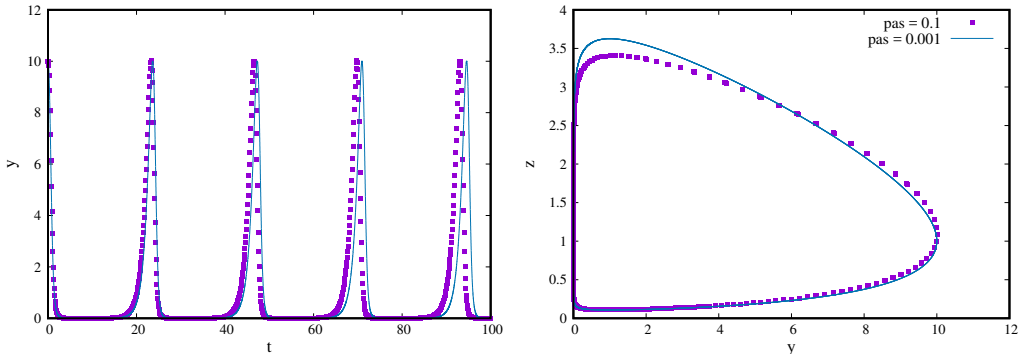

Figure 21: Evolution des populations de proies $y(t)$ au cours du temps (panneau de gauche) obtenue par le schéma ESI, avec deux pas de temps, $\Delta t = 0.1$ (points), et $\Delta t = 0.001$ (trait plein). Trajectoire correspondante $y(z)$ (panneau de droite) dans l'espace des phases.

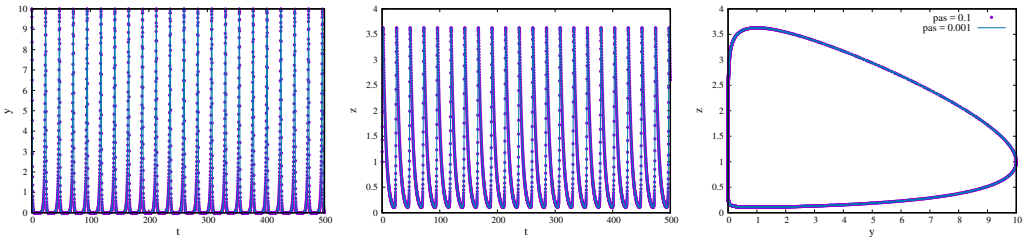

Figure 22: Evolution des populations de proies $y(t)$ au cours du temps (panneau de gauche) et de prédateurs $z(t)$ (panneau du milieu) obtenue par le schéma RK4, avec deux pas de temps, $\Delta t = 0.1$ (points), et $\Delta t = 0.001$ (trait plein). Trajectoire correspondante $y(z)$ (panneau de droite) dans l'espace des phases.

4 Application aux attracteurs étranges et au chaos dissipatif

Faisons nous maintenant un peu plus plaisir, en nous focalisant sur un système particulier de trois équations différentielles, connu pour conduire à des solutions chaotiques au cours du temps. C'est un modèle simplificateur de phénomènes physiques pour lesquels la dissipation joue un rôle important. On parle alors de chaos dissipatif, contrairement au chaos Hamiltonien pour lequel l'énergie totale est conservée (pas de dissipation) et dont nous parlerons dans les prochains chapitres. La dynamique de type chaos dissipatif requiert un minimum de trois équations différentielles du premier ordre (cela se démontre, mais nous l'admettrons ici). Les trois variables sont couplées au travers de termes non-linéaires dans les équations.

4.1 Equations du système de Lorenz

Le modèle de Lorenz a été introduit en 1963 comme un système simplificateur en mécanique des fluides pour expliquer l'existence de solutions turbulentes (chaotiques). C'est le modèle auquel on songe parfois pour évoquer le fameux effet papillon de la météo. Et par un curieux effet du hasard, les trajectoires tri-dimensionelles dans l'espace des phases (x, y, z) donnent une structure en ailes de papillon (voir plus loin). Le système de Lorenz s'écrit,

$$\frac{dx}{dt} = P_r(y - x), \tag{33}$$

$$\frac{dy}{dt} = -xz + rx - y, \tag{34}$$

$$\frac{dz}{dt} = xy - bz, \tag{35}$$

avec P_r, b, et r trois paramètres réels positifs. Les valeurs de P_r et de b sont souvent fixés ($P_r = 10$ et $b = 8/3$), et r est le paramètre variable.

4.2 Solutions obtenues pour plusieurs valeurs de r

Lançons nous dans le bain, ou plutôt dans l'attracteur en l'occurrence, et choisissons la méthode la plus performante que nous ayons programmée jusqu'à présent (Runge-Kutta d'ordre 4, RK4) du moins au niveau de la précision. Nous partons des conditions initiales $x_0 = y_0 = z_0 = 10$ et utilisons un pas $\Delta t = 0.01$. Et nous choisissons tout d'abord un cas avec $r < 1$, $r = 0.5$ en l'occurence. Les résultats sont visible sur la figure 23. Ils montrent que la solution converge rapidement vers le point origine $(0, 0, 0)$. Ceci est valable quel que soit le point de départ. Ils confirment l'analyse exposée en annexe, qui montre que l'origine est un point fixe stable pour $r < 1$, qui attire ainsi toutes les trajectoires.

Prenons ensuite un autre cas pour lequel $r = 5$, et appliquons exactement la même procédure et les mêmes conditions initiales. Les résultats sont maintenant visibles sur les figures 24-25 pour deux conditions initiales. Ils montrent que les trajectoires convergent vers deux nouveaux points fixes stables, en accord avec ce que prévoit la théorie. En effet pour $1 < r < 24.74$, le point origine est instable alors que deux nouveaux points fixe stables déterminent la dynamique. Ces deux points d'attraction ont pour coordonnées $(x_1, y_1, z_1) = (\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$, et $(x_2, y_2, z_2) = (-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$, qui sont vérifiées sur les figures. La transition pour $r = 1$ est appelée une bifurcation fourche dans la théorie des systèmes dynamiques. Le même type de comportement est observé pour une valeur de r plus élevée ($r = 20$ pour les figures 26-27), mais avec un comportement transitoire (oscillant) plus complexe et une convergence plus lente vers les points fixes.

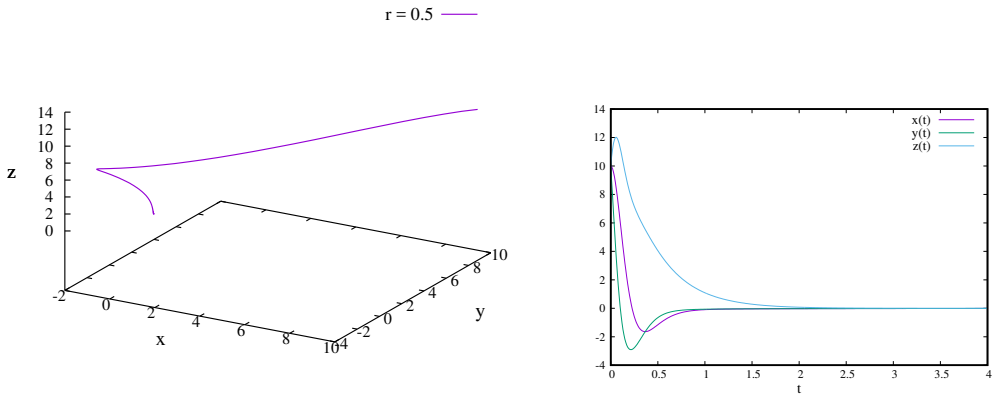

Figure 23: Solution obtenue pour le cas $r = 0.5$ (utilisant RK4) dans l'espace des phases (tableau de gauche), et en fonction du temps (tableau de droite).

Enfin, nous prenons le cas $r = 28$, qui est en général la valeur choisie pour illustrer les comportement chaotiques dans la litérature. Nous avons utilisé les conditions initiales $(x_0, y_0, z_0) = (-15.8, -17.48, 35.64)$. Maintenant, les solutions au cours du temps apparaissent fluctuer de façon stochastiques au cours du temps (figure 28). La trajectoire correspondante dans l'espace des phases décrit une pseudo-surface appelée attracteur étrange de Lorenz. En fait, les deux points fixes sont devenus instables pour $r > 24.74$ par une bifurcation dite de Hopf (quelques éléments théoriques sont dispensés en Annexe).

4.3 Effet du schéma numérique sur les solutions chaotiques

Alors que le choix de la méthode numérique pour les valeurs de $r < 24.74$ n'est pas aussi fondamental que pour le système de Lotka-Volterra (pas montré ici mais facile à vérifier), ce n'est plus le cas lorsque on s'intéresse aux solutions pour des valeurs de r plus élevées.

On se place maintenant dans le cas chaotique, $r = 28$. Il est alors instructif d'examiner les effets suivants sur la solution obtenue, comme celui du pas de temps et du schéma numérique. Nous utilisons exactement la même procédure et conditions initiales que décrites au dessus. Les résultats sont maintenant visibles sur les figures 29. Sur le panneau de gauche qui permet de comparer l'emploi de plusieurs pas de temps pour une même méthode (RK4), nous constatons que la solution commence à différer pour $t = 15$ entre le cas $\Delta t = 10^{-2}$ et les autres. La solution pour $\Delta t = 10^{-3}$ devient différente de celle pour $\Delta t = 10^{-4}$ plus tard, pour $t = 24$ plus précisément. Ensuite, la comparaison de l'emploi de deux méthodes similaires (RK4, et une variante qui est en faite une autre Runge-Kutta d'ordre 4 détaillée en annexe et appellée RK4b) montre la différentiation entre les deux solutions sur le panneau central pour $t = 25$. Quand au dernier test (panneau de droite), il montre que les solutions se séparent encore plus tard (pour $t = 29$) si on utilise deux schémas de Runge-Kutta d'ordre 5. Dans tous ces tests, la structure sur laquelle la trajectoire se déplace dans l'espace des phases reste identique (non montré).

De plus, nous avons vérifié que si nous introduisons un petit écart entre deux jeux de conditions initiales, un comportement similaire à celui décrit ci-dessus est obtenu. Par exemple une différence de 10^{-8} sur z_0 en employant le schéma RK5b conduit à la divergence de la solution à $t = 23$, alors que c'est à $t = 14$ seulement pour une une différence de 10^{-5}. Ce genre de comportement est bien sûr caractéristique des solutions chaotiques, et représente le fameux 'effet papillon'.

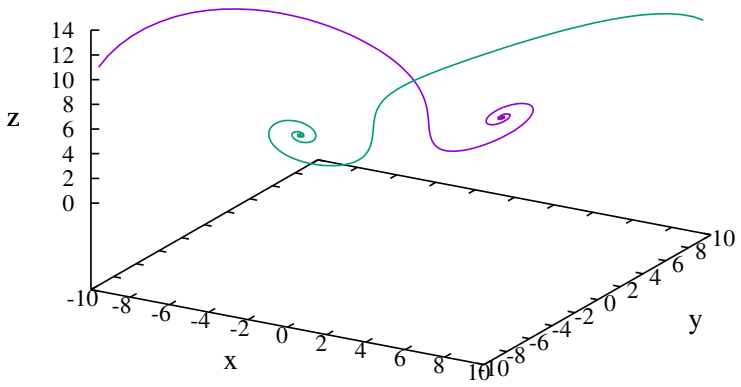

Figure 24: Trajectoires obtenue pour le cas $r = 5$ (utilisant RK4) dans l'espace des phases en partant de deux conditions initiales différentes (deux couleurs différentes).

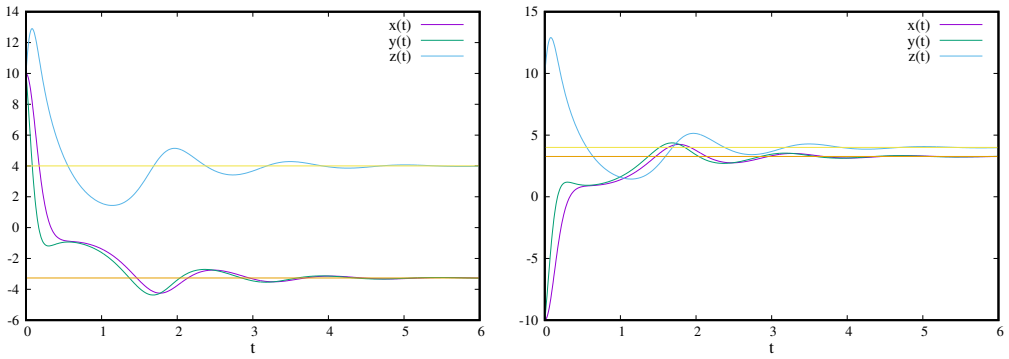

Figure 25: Solutions obtenues pour le cas $r = 5$ (utilisant RK4) en fonction du temps, pour une condition initiale $(x_0, y_0, z_0) = (10, 10, 10)$ (panneau de gauche), et $(x_0, y_0, z_0) = (-10, -10, 10)$ (panneau de droite). Les traits verticaux indiquent les coordonnées des points fixes (voir texte).

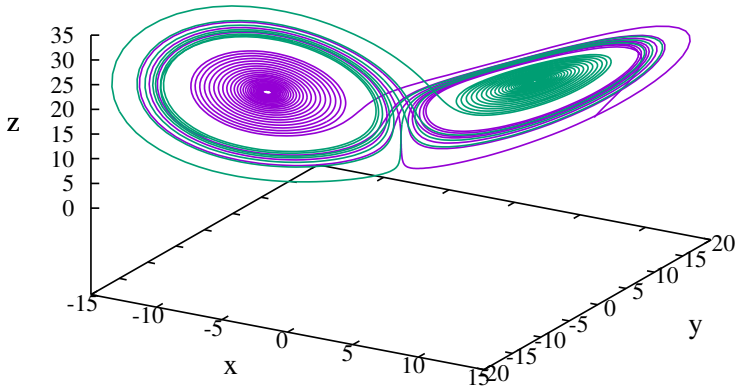

Figure 26: Trajectoires obtenue pour le cas $r = 20$ (utilisant RK4) dans l'espace des phases en partant de deux conditions initiales différentes (deux couleurs différentes).

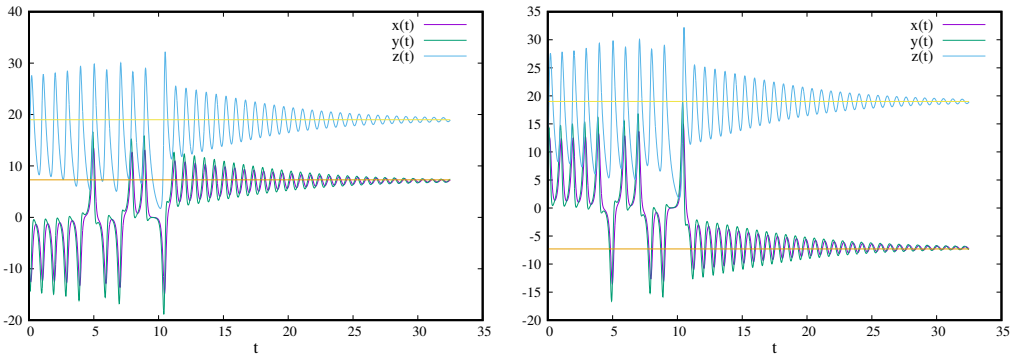

Figure 27: Solutions obtenues pour le cas $r = 20$ (utilisant RK4) en fonction du temps, pour une condition initiale $(x_0, y_0, z_0) = (-10, -10, 10)$ (tableau de gauche), et $(x_0, y_0, z_0) = (10, 10, 10)$ (tableau de droite). Les traits verticaux indiquent les coordonnées des points fixes (voir texte).

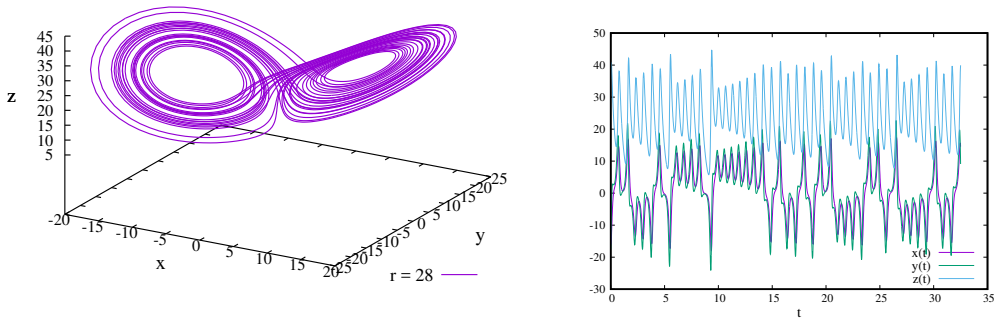

Figure 28: Solution obtenue pour le cas $r = 28$ (utilisant RK4) dans l'espace des phases (tableau de gauche), et en fonction du temps (tableau de droite).

Un petit écart introduit δ_0 sur les conditions initiales ou le schéma numérique va se traduire par une divergence exponentielle de cet écart entre les solutions en $\delta_0 \exp(\lambda t)$. Nous pouvons ainsi calculer λ en reportant cette écart au cours du temps. Les résultats mesurant sur la composante x sont reportés sur la figure 30, en utilisant RK5b pour un pas de temps $\Delta t = 10^{-3}$ et un écart initial introduit de $\delta_0 = 10^{-8}$. On obtient une variation temporelle exponentielle moyenne en accord avec $\lambda \approx 0.9$, qui est justement la valeur attendue par la théorie. Ce paramètre est appelé exposant de Lyapunov.

Ces tests montrent que les solutions obtenues par deux méthodes vont de toute façon diverger de par la nature chaotique du système (pour r dépassant le seuil de Hopf). On peut même faire l'expérience d'exécuter le même programme sur deux ordinateurs différents pour vérifier que les solutions vont diverger plus tard car les deux 'runs' correspondront à un écart introduit $\delta_0 \approx 10^{-15}$, qui est l'ordre de grandeur des erreurs de codage en double précision standard.

En conclusion, la prédiction sera d'autant meilleure que l'ordre du schéma utilisé est élevé, et les méthodes de Runge-Kutta sont particulièment performantes. Nul besoin ici (contrairement au problème de Lotka-Volterra) d'avoir recours à un schéma symplectique, car il n'y a pas conservation du volume dans l'espace des phases pour la solution attendue.

A Quelques développements limités utiles

Dans les formules suivantes, on considère ϵ comme un nombre réel petit devant 1.

$$\exp(-\epsilon) = 1 - \frac{\epsilon}{1!} + \frac{\epsilon^2}{2!} - \frac{\epsilon^3}{3!} + ... + (-1)^k \frac{\epsilon^k}{k!} + O(\epsilon^{k+1}) \tag{36}$$

$$\frac{1}{1+\epsilon} = 1 - \epsilon + \epsilon^2 - \epsilon^3 + ... + (-1)^k \epsilon^k + O(\epsilon^{k+1}) \tag{37}$$

$$\frac{1+\epsilon/2}{1-\epsilon/2} = 1 - \epsilon + \frac{\epsilon^2}{2} - \frac{\epsilon^3}{4} + ...(-1)^k \frac{\epsilon^k}{2^{k-1}} + O(\epsilon^{k+1}) \tag{38}$$

Pour les formules suivantes, n est un entier grand devant 1.

$$\exp(-n\epsilon) = 1 - \frac{n\epsilon}{1!} + \frac{n^2\epsilon^2}{2!} - \frac{n^3\epsilon^3}{3!} + ... + (-1)^k \frac{n^k \epsilon^k}{k!} + O(\epsilon^{k+1}) \tag{39}$$

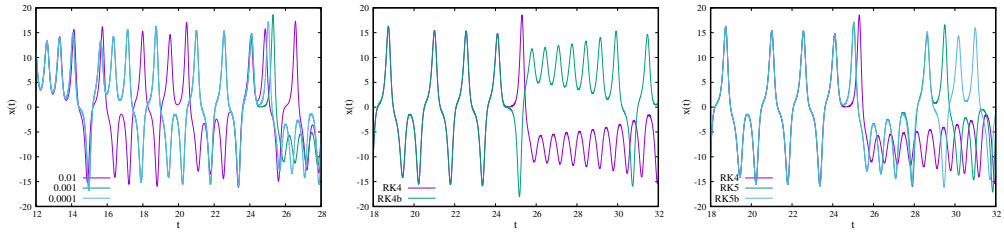

Figure 29: Panneau de gauche: solution obtenue pour $x(t)$ pour le cas $r = 28$, utilisant RK4 avec 3 pas de temps différents, ($\Delta t = 10^{-2}, 10^{-3}$,et 10^{-4}). Panneau central: idem mais avec deux schémas de type Runge-Kutta d'ordre 4 (RK4 et RK4b) et $\Delta t = 10^{-3}$. Panneau de droite: idem mais avec trois schémas de type Runge-Kutta (RK4, et deux autres schémas d'ordre 5 RK5 et RK5b) et $\Delta t = 10^{-3}$.

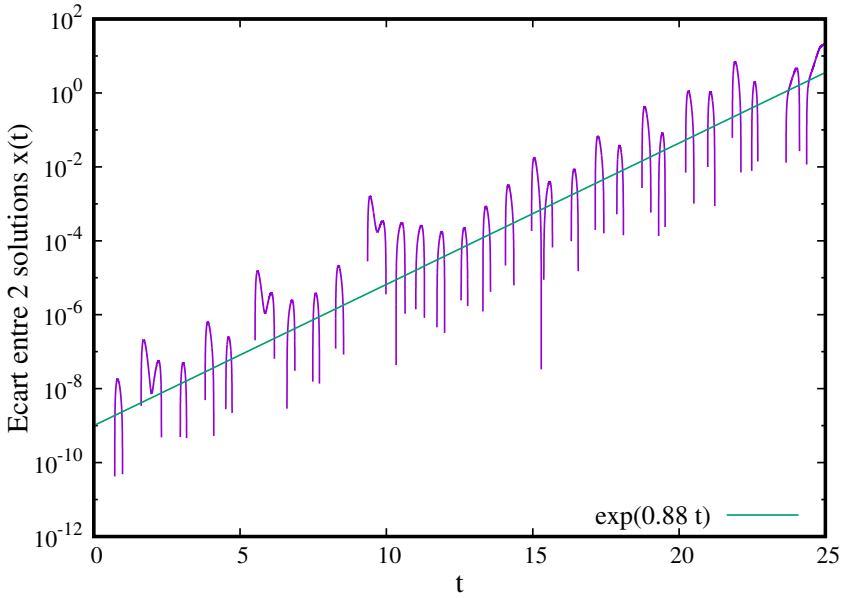

Figure 30: Ecart sur $x(t)$ mesuré entre deux solutions obtenues par le schéma RK5b et $\Delta t = 10^{-3}$ et qui diffèrrent par un écart initial de $\delta_0 = 10^{-8}$ sur les conditions initiales. Une courbe en trait plein permet d'ajuster une variation en $\exp(0.88t)$. .

$$(1-\epsilon)^n = 1-n\epsilon+\frac{n(n-1)}{2!}\epsilon^2-\frac{n(n-1)(n-2)}{3!}\epsilon^3+...+(-1)^k\frac{n(n-1)...(n-k+1)}{k!}\epsilon^k+O(\epsilon^{k+1}) \tag{40}$$

La différence entre les deux dernières expressions conduit à, $\exp(-n\epsilon)-(1-\epsilon)^n = n\epsilon^2/2+O(\epsilon^3)$. Ainsi, pour $t_n = n\epsilon$ (ϵ représentant le pas de temps), le terme dominant de cette différence devient $t_n\epsilon/2$ donc est en $O(\epsilon)$. Ce calcul sert à montrer qu'un schéma ayant un ordre local de 2 se réduit à un ordre global de 1.

B Notions sur les erreurs d'arrondis et leurs effets

La précision dite 'machine' qui entache les nombres réels utilisés sur un ordinateur dépend du nombre de bits (les 0 et les 1 en codage binaire) défini pour les coder. Ainsi, l'emploi de la double précision (format 64 bits) généralement utilisée, implique qu'un nombre réel x est égal à sa vraie valeur x_* avec une incertitude relative d'environ 2×10^{-16}, ou plus précisément $x \approx x_*(1 \pm 2 \times 10^{-16})$. Ainsi cette erreur appelée erreur d'arrondi est liée au nombre de chiffres significatifs dans la représentation des nombres en machine.

Les opérations élémentaires (addition, multiplication, ...) vont conduire alors à une dégradation de cette erreur lorsque le nombre d'opérations augmente. Plus précisément, l'erreur d'arrondi augmente linéairement avec le nombre d'opérations. Ainsi pour une intégration numérique à pas constant, l'erreur d'arrondi varie inversement proportionnellement avec le pas de temps.

C Matrices d'amplification (relaxation 2D) et stabilité

La stabilité numérique des schémas impliquant un seul scalaire (une seule équation différentielle) s'examine facilement en utilisant le coefficient d'amplification k_a si son expression est disponible analytiquement. C'est le cas des schémas abordés dans cet ouvrage et appliqués à l'équation de relaxation. Ainsi la condition nécéssaire et suffisante de stabilité d'un schéma se résume à $k_a \leq 1$.

Pour un système impliquant plusieurs scalaires (un système de plusieurs équations différentielles), la notion de coefficient d'amplification se généralise alors pour donner une matrice d'amplification. On se limitera au cas de deux scalaires, donc à une matrice 2×2. Dans le cas d'un système de type relaxation (cas linéaire), la matrice d'amplification $\tilde{A}$ peut s'obtenir simplement à partir de la matrice A définie par,

$$\dot{\vec{Y}} = A\vec{Y}. \tag{41}$$

Par exemple le schéma EE conduit à $\tilde{A}_{EE} = I + A\Delta t$, alors que $\tilde{A}_{EI} = (I - A\Delta t)^{-1}$ pour le schema EI, et $\tilde{A}_{EI} = \frac{I+A\Delta t/2}{I-A\Delta t/2}$ pour le schéma EC (I désignant la matrice unité). Il faut faire attention au signe si on compare au cas scalaire (i.e. $dy/dt = -y$, et donc A devient -1). La matrice d'amplification est en effet définie à partir de la relation,

$$\vec{Y}_{n+1} = \tilde{A}\vec{Y}_n. \tag{42}$$

Une condition nécessaire et suffisante de stabilité est alors $\|\tilde{A}\| \leq 1$. Si $\tilde{A}$ est dite normale (ou encore elle commute avec sa transposée, $\tilde{A}\tilde{A}^t = \tilde{A}^t\tilde{A}$), alors la condition se ramène à,

$$\rho(\tilde{A}) \leq 1, \tag{43}$$

avec $\rho(\tilde{A})$ le rayon spectral de $\tilde{A}$ (c'est à dire le module de la plus grande valeur propre, $Max|\lambda_i|$, pour $i = 1, 2$). Si $\tilde{A}$ n'est pas normale, $\rho(\tilde{A}) \leq 1$ est seulement une condition nécessaire mais

pas suffisante de stabilité. Par contre, $\rho(\tilde{A}) > 1$ implique une instabilité car on a toujours l'inégalité $\rho(\tilde{A}) \leq \|\tilde{A}\|$ (que $\tilde{A}$ soit normale ou pas). Par exemple, si on prend le cas de la matrice $A = \begin{pmatrix} -6 & 10 \\ 2 & -5 \end{pmatrix}$, le calcul des valeurs propres de A donne -1 et -10. Malheureusement, les matrices d'amplifications des schémas (EE, EI, et EC) ne sont pas normales et le calcul ces conditions nécessaires et suffisantes de stabilité s'avèrent très compliqués. On se contentera des conditions (suffisantes) d'instabilité, $\rho(\tilde{A}) > 1$. Ainsi, comme les deux valeurs propres de $\tilde{A}_{EE}$ sont $1 - \Delta t$ et $1 - 10\Delta t$, la condition suffisante d'instabilité est $\Delta t > 0.2$ car $(1 - 10\Delta t < -1)$.

Le calcul des solutions analytiques de relaxation pour le système avec $A = \begin{pmatrix} -6 & 10 \\ 2 & -5 \end{pmatrix}$, est le suivant. Il faut calculer les vecteurs propres associés aux deux valeurs propres, $\lambda_1 = -1$ et $\lambda_2 = -10$. Le premier vecteur propre doit vérifier $A\vec{X_1} = \lambda_1 \vec{X_1}$, conduisant alors à $\vec{X_1}^t = (2, 1)$. De même $\vec{X_2}^t = (-5/2, 1)$ pour le second vecteur propre associé à λ_2. Ainsi, la solution générale s'écrit,

$$\vec{X} = a_1 \vec{X_1} \exp(-\lambda_1 t) + a_2 \vec{X_2} \exp(-\lambda_2 t), \tag{44}$$

avec a_1 et a_2 deux constantes réelles déterminées par les conditions initiales.

D Les différents schémas utilisés: Euler et Runge-Kutta classiques

Nous récapitulons ici tous les schémas évoqués (de près ou de loin) dans le manuscrit. Pour simplifier, nous considérons leurs applications sur l'équation type,

$$dy/dt = f(y(t), t). \tag{45}$$

D.1 Les schémas d'Euler

Le schéma d'Euler explicite (EE) sécrit,

$$y_{n+1} = f(y_n, t_n). \tag{46}$$

Le schéma d'Euler implicite (EI) sécrit,

$$y_{n+1} = f(y_{n+1}, t_{n+1}). \tag{47}$$

Le schéma d'Euler centré (EC) sécrit,

$$y_{n+1} = [f(y_{n+1}, t_{n+1}) + f(y_n, t_n)]/2. \tag{48}$$

D.2 Les schémas de Runge-Kutta

D.2.1 RK2

-Le schéma de Runge-Kutta d'ordre 2 (RK2) sécrit,

$$y_{n+\frac{1}{2}} = y_n + \frac{\Delta t}{2} f(y_n, t_n), \tag{49}$$

et finalement,

$$y_{n+1} = y_n + \Delta t f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}). \tag{50}$$

D.2.2 RK3

-Le schéma de Runge-Kutta d'ordre 2 (RK3) sécrit (bien que non utilisé pour nos tests),

$$y_{n+\frac{1}{2}} = y_n + \frac{\Delta t}{2} f(y_n, t_n), \tag{51}$$

$$y_{n+1,1} = y_n + \Delta t[2f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}) - f(y_n, t_n)], \tag{52}$$

et finalement,

$$y_{n+1} = y_n + \frac{\Delta t}{6}[f(y_n, t_n) + 4f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}) + f(y_{n+1,1}, t_{n+1})]. \tag{53}$$

D.2.3 RK4

-Le schéma de Runge-Kutta d'ordre 4 (RK4) sécrit,

$$y_{n+\frac{1}{2},1} = y_n + \frac{\Delta t}{2} f(y_n, t_n), \tag{54}$$

$$y_{n+\frac{1}{2},2} = y_n + \frac{\Delta t}{2} f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}), \tag{55}$$

$$y_{n+1,1} = y_n + \Delta t f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}), \tag{56}$$

pour les 3 évaluations intermédiaires (deux en $t_{n+\frac{1}{2}}$ et une en t_{n+1}). Et, la formule d'extrapolation finale est alors,

$$y_{n+1} = y_n + \frac{\Delta t}{6}[f(y_n, t_n) + 2f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}) + 2f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}) + f(y_{n+1,1}, t_{n+1})]. \tag{57}$$

D.2.4 RK4b

-Le second schéma de Runge-Kutta d'ordre 4 (noté RK4b) sécrit,

$$y_{n+\frac{1}{3}} = y_n + \frac{\Delta t}{3} f(y_n, t_n), \tag{58}$$

$$y_{n+\frac{2}{3}} = y_n + \frac{\Delta t}{3}[3f(y_{n+\frac{1}{3}}, t_{n+\frac{1}{3}}) - f(y_n, t_n)], \tag{59}$$

$$y_{n+1,1} = y_n + \Delta t[f(y_n, t_n) - f(y_{n+\frac{1}{3}}, t_{n+\frac{1}{3}}) + f(y_{n+\frac{2}{3}}, t_{n+\frac{2}{3}})], \tag{60}$$

pour les 3 évaluations intermédiaires (une en $t_{n+\frac{1}{3}} = t_n + \frac{\Delta t}{3}$, une en $t_{n+\frac{2}{3}} = t_n + 2\frac{\Delta t}{3}$, et enfin une en t_{n+1}). Et, la formule d'extrapolation finale est alors,

$$y_{n+1} = y_n + \frac{\Delta t}{8}[f(y_n, t_n) + 3f(y_{n+\frac{1}{3}}, t_{n+\frac{1}{3}}) + 3f(y_{n+\frac{2}{3}}, t_{n+\frac{2}{3}}) + f(y_{n+1,1}, t_{n+1})]. \tag{61}$$

Cette seconde Runge-Kutta d'ordre 4 est aussi appellée méthode RK4 avec règle des 3/8.

D.2.5 RK5

-Le premier schéma de Runge-Kutta d'ordre 5 (noté RK5) sécrit,

$$y_{n+\frac{1}{2},1} = y_n + \frac{\Delta t}{2} f(y_n, t_n), \tag{62}$$

$$y_{n+\frac{1}{4}} = y_n + \frac{\Delta t}{16}[3f(y_n, t_n) + f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}})], \tag{63}$$

$$y_{n+\frac{1}{2},2} = y_n + \frac{\Delta t}{2} f(y_{n+\frac{1}{4}}, t_{n+\frac{1}{4}}) \tag{64}$$

$$y_{n+\frac{3}{4}} = y_n + \frac{\Delta t}{16}[-3f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}) + 6f(y_{n+\frac{1}{4}}, t_{n+\frac{1}{4}}) + 9f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}})], \tag{65}$$

$$y_{n+1,1} = y_n + \frac{\Delta t}{7}[f(y_n, t_n) + 4f(y_{n+\frac{1}{2},1}, t_{n+\frac{1}{2}}) + 6f(y_{n+\frac{1}{4}}, t_{n+\frac{1}{4}}) - 12f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}) + 8f(y_{n+\frac{3}{4}}, t_{n+\frac{3}{4}})]. \tag{66}$$

Et, la formule d'extrapolation finale est alors,

$$y_{n+1} = y_n + \frac{\Delta t}{90}[7f(y_n, t_n) + 32f(y_{n+\frac{1}{4}}, t_{n+\frac{1}{4}}) + 12f(y_{n+\frac{1}{2},2}, t_{n+\frac{1}{2}}) + 32f(y_{n+\frac{3}{4}}, t_{n+\frac{3}{4}}) + 7f(y_{n+1,1}, t_{n+1})]. \tag{67}$$

D.2.6 RK5b

-Le second schéma de Runge-Kutta d'ordre 5 (noté RK5) sécrit,

$$y_{n+\frac{1}{4},1} = y_n + \frac{\Delta t}{4} f(y_n, t_n), \tag{68}$$

$$y_{n+\frac{1}{4},2} = y_n + \frac{\Delta t}{8}[f(y_n, t_n) + f(y_{n+\frac{1}{4},1}, t_{n+\frac{1}{4}})], \tag{69}$$

$$y_{n+\frac{1}{2}} = y_n + \frac{\Delta t}{2}[-f(y_{n+\frac{1}{4},1}, t_{n+\frac{1}{4}}) + 2f(y_{n+\frac{1}{4},2}, t_{n+\frac{1}{4}})] \tag{70}$$

$$y_{n+\frac{3}{4}} = y_n + \frac{\Delta t}{16}[3f(y_n, t_n) + 9f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}})], \tag{71}$$

$$y_{n+1,1} = y_n + \frac{\Delta t}{7}[-3f(y_n, t_n) + 2f(y_{n+\frac{1}{4},1}, t_{n+\frac{1}{4}}) + 12f(y_{n+\frac{1}{4},2}, t_{n+\frac{1}{4}} - 12f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}) + 8f(y_{n+\frac{3}{4}}, t_{n+\frac{3}{4}})]. \tag{72}$$

Et, la formule d'extrapolation finale est alors,

$$y_{n+1} = y_n + \frac{\Delta t}{90}[7f(y_n, t_n) + 32f(y_{n+\frac{1}{4},2}, t_{n+\frac{1}{4}}) + 12f(y_{n+\frac{1}{2}}, t_{n+\frac{1}{2}}) + 32f(y_{n+\frac{3}{4}}, t_{n+\frac{3}{4}}) + 7f(y_{n+1,1}, t_{n+1})]. \tag{73}$$

E Quelques résultats théoriques sur le systeme de Lorenz

Le système de Lorenz s'écrit,

$$\frac{dx}{dt} = P_r(y - x), \tag{74}$$

$$\frac{dy}{dt} = -xz + rx - y, \tag{75}$$

$$\frac{dz}{dt} = xy - bz, \tag{76}$$

avec P_r, b, et r trois paramètres réels positifs. Les valeurs de P_r et de b sont fixés ($P_r = 10$ et $b = 8/3$), et r est le paramètre variable. La dynamique des solutions est déterminée par l'existence de points fixes (définis par les solutions de $dx/dt = dy/dt = dz/dt = 0$) dans l'espace des phases (x, y, z), ainsi que leur stabilité.

Tout d'abord, il existe le point fixe trivial, $(0, 0, 0)$, quel que soit la valeur prise par r. Ensuite, deux point fixes supplémentaires existent lorsque $r > 1$. Ces deux points sont symétriques et ont pour coordonnées $(x_1, y_1, z_1) = (\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$, et $(x_2, y_2, z_2) = (-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$.

La stabilité d'un point fixe se détermine en étudiant si des trajectoires partant de conditions initiales proches du point fixe vont s'en éloigner continuellement ou y converger définitivement. Pour ce faire, il nous faut calculer une matrice, la matrice Jacobienne J de l'opérateur non-linéaire F défini par $\frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = F\begin{pmatrix} x \\ y \\ z \end{pmatrix}$. On obtient alors,

$$J = \begin{pmatrix} -P_r & P_r & 0 \\ r - z & -1 & -x \\ y & x & -b \end{pmatrix}. \tag{77}$$

Cette matrice prise en $(0, 0, 0)$, devient $J_0 = \begin{pmatrix} -P_r & P_r & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}$. Sans rentrer dans les détails de théorie des systèmes dynamiques, on admet que c'est l'existence de valeurs propres positives qui rend le point fixe instable. Ainsi, le polynôme caractéristique (du troisième ordre) qui est, $(\lambda + b)[\lambda^2 + (1 + P_r)\lambda + P_r(1 - r)] = 0$, admet 3 valeurs propres négatives si $r < 1$ et seulement 2 négatives (1 positive) pour $r > 1$. Ceci montre que le point fixe $(0, 0, 0)$ devient instable pour $r > 1$.

La matrice Jacobienne prise aux deux autres points fixes conduit au polynôme caractéristique suivant, $\lambda^3 + (P_r + b + 1)\lambda^2 + b(r + P_r)\lambda + 2P_r b(r - 1) = 0$. Maintenant, les solutions de cette équation conduit à deux cas selon la valeur de r. Soit nous avons trois valeurs propres réelles négatives (points fixes stables), soit nous avons une valeur propres réelle négative et deux autres complexes conjuguées. Ainsi, on peut montrer que c'est lorsqu' apparaissent ces valeurs complexes qu'une bifurcation vers un système chaotique survient. La valeurs critique correspondante pour r est obtenue en portant dans le polynôme caractéristique une valeur pour λ, $\lambda = i\alpha$, avec α réel. Ceci permet d'obtenir la valeur critique pour r, $r_c = \frac{P_r(P_r+b+3)}{(P_r-b-1)} \approx 24.74$.