Méthodes numériques ROCK4 pour la résolution d'équations différentielles et étude de l'algèbre pré-Lie sous-jacente

Jérémy Cochoy

1er février 2011

Résumé

TIPE réalisé sous la co-tutelle de Thierry Dumont et de Frédéric Chapoton
Jérémy Cochoy

Table des matières

Première partie
Méthode ROCK4

1 Introduction

Le calcul différentiel, et par conséquent les équations différentielles, sont aujourd'hui extrêmement présents dans de nombreux domaines scientifiques, et ont donné lieu à une abondante bibliographie. Les prémisses de ces concepts existaient déjà du temps d'Archimède, avec la méthode d'exhaustion qui, pour calculer l'aire d'une surface, consiste à l'encadrer entre deux suites de surfaces dont on sait calculer l'aire et qui tendent vers la même limite. L'exemple le plus élémentaire est le calcul de l'aire d'un disque en l'encadrant avec une suite de polygones inscrits et circonscrits à ce dernier.

Plus récemment, durant la première moitié du XVII^{e} siècle, *Bonaventura Francesco Cavalieri* développe la méthode des indivisibles qui constitue une nouvelle alternative aux anciennes méthodes d'*Archimède*. Les volumes sont subdivisés en une superposition de surfaces, et les surfaces en une juxtaposition de lignes.

On conviendra toutefois que le calcul différentiel est apparu aux alentours de la fin du XVII^{e} siècle avec les travaux d'*Isaac Newton* sur les rapports de quantités infinitésimales, appelant fluxion ce que l'on définit comme dérivée d'une fonction, c'est à dire la 'limite de son taux d'accroissement', aujourd'hui noté :

$$\dot{x} = \lim_{h \to x} \frac{f(x) - f(h)}{x - h}$$

et les notations introduites par *Gottfried Wilhelm von Leibniz*. Ce dernier conçoit des quantités infinitésimales, qu'il note sous la forme dx, ainsi que l'intégrale sous la forme d'une 'somme infinie de quantités infinitésimales'. Ce sont ces nouvelles notations qui permettront le développement de cet outil et ses applications aux problèmes de l'époque.

Des équations différentielles apparaissaient alors, sous la forme du **problème inverse des tangentes**. Étant donné des relations déterminant les fluxions, retrouver les fluentes, ou encore ; étant donné "les tangentes", déterminer la courbe, la surface, le volume qu'elles engendrent.

Un des problèmes les plus délicats est **le problème des trois corps** ; si l'on considère trois corps célestes, et en appliquant les théorèmes de mécanique pour décrire leurs interactions, on parvient à établir les relations entre les dérivées de leurs positions. On sait aujourd'hui qu'il n'existe pas de solution analytique à ce problème, si l'on s'abstient de négliger l'interaction entre les corps du système.

Les questions sur l'existence et l'unicité des solutions ne se sont posées que plus tard, aux alentours du début du XIX^{e} siècle quand Cauchy critiqua le bien-fondé des méthodes employées. En effet, des méthodes comme l'utilisation de séries ou encore de ce que nous appellerions aujourd'hui la **méthode de variation de la constante** étaient déjà employées pour résoudre certaines équations différentielles. De même, le fait qu'une équation différentielle homogène à coefficients constants de degré n possède pour solution une combinaison linéaire de n exponentielles de la forme e^{rx} était connu bien avant d'être démontré.

Ces outils mûrirent, et l'on chercha à résoudre de nouveaux problèmes, qui ne manquèrent pas. Et pour cause, les équations différentielles interviennent dans une grande variété de domaines,

comme l'électromagnétisme, la thermodynamique, la biologie, la chimie, la mécanique quantique. Nous pouvons citer pour exemple, provenant de la mécanique, l'oscillation libre non-amortie d'un pendule, dont l'équation différentielle, non-linéaire, associée, est :

$$\ddot{\theta} + \frac{g}{l}\sin(\theta) = 0$$

Rappelons-nous aussi l'équation de D'Alembert, décrivant la propagation d'une onde dans une corde si l'on reste dans des amplitudes raisonnables, aussi appelée équation d'onde :

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0$$

qui nous apporte un magnifique exemple d'équation aux dérivées partielles simple liant temps et espace.

Les théories relatives aux équations différentielles sont nombreuses, et l'on peut trouver des volumes entiers consacrés à la résolution de cas particuliers. Malheureusement, nombre de ces équations, si présentes dans la plupart des problèmes de modélisation, ne sont pas toujours résolubles de façon analytique, ou encore numériquement dans un temps raisonnable. Une alternative est apparue et avec l'avènement du calcul scientifique, qui consiste à utiliser des méthodes de résolution numériques.

Il existe de nombreuses méthodes numériques adaptées à la résolution de différents problèmes. En effet, pour deux équations différentielles données, une même méthode pourra fournir une approximation de la solution convenable pour la première, et des valeurs complètement erronées présentant une erreur colossale. De plus, toutes ces méthodes ne sont pas comparables dans leur utilisation ; certaines ont un coût de calcul bien plus élevé que d'autres.

Nous nous intéresserons ici à une méthode particulière, développée pour résoudre des équations différentielles raides ; ROCK4. Il nous sera nécessaire d'introduire quelques définitions, et exemples affin de mieux cerner le contexte relatif à son utilisation. Nous présenterons donc les méthodes dites de Runge-Kutta dont les premières versions *Carl Runge* et *Martin Wilhelm Kutta*, puis nous établirons un théorème sur une caractéristique de l'erreur introduite par ces méthodes. Disposant alors des pré-requis nécessaires, nous étudierons en détail la construction de ROCK2, fortement similaire à ROCK4, avant de conclure sur la construction de cette dernière méthode. Revenant sur la présence récurrente de certaines structures dans certaines démonstrations, nous chercherons à établir un important théorème qui justifie ce comportement. Il sera à nouveau nécessaire d'introduire quelques définitions et exemples. Nous démontrerons par la suite quelques comportement spécifiques avant de discuter un important théorème justifiant l'existence de ces structures. Enfin, nous prendrons plaisir à déterminer l'origine de certains coefficients particuliers et de démontrer certaines de leurs propriétés.

2 Motivation

La notion de *raideur* d'un problème est une notion qualitative difficile à définir formellement. On trouve généralement des définitions de la forme ([EH91]) :

> The most pragmatical opinion is also historically the first one (Curtiss & Hirschfelder 1952): *stiff equations are equations where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones.*
> (Hairer & Wanner 1991)

On définira plus en détail les notions de méthodes “explicite” et “implicite”. Retenons simplement que les premières nécessités, pour obtenir la valeur à l'instant t, d'effectuer un calcul de la forme $X_t = \mathcal{F}(X_{t-1})$ où X_{t-1} correspond aux données connues à l'instant précédent. A contrario, les méthodes implicites nécessitent la résolution de système (souvent non linéaire) de la forme $X_t = \mathcal{F}(X_t, X_{t-1})$.

Les équations différentielle raides sont donc celles qui se prêtent difficilement à une résolution numérique par des outils simples comme la méthode d'Euler(cf: 3.1) explicite et, plus généralement, par des méthodes explicites (cf: 3.2).

Dans la suite de ce document, nous retiendrons la définition suivante :

Définition (Equation différentielle raide) : Si l'on considère un problème de la forme

$$\frac{\mathrm{d}u}{\mathrm{d}t} = F(u)$$

et que l'on note $\mathrm{D}F(u)\ in\mathbb{R} \times \mathbb{R}$ le jacobien de $F(u)$, alors l'équation différentielle est raide si

$$\exists \lambda_1, \lambda_2 \in Sp(\mathrm{D}F(u)) |\lambda_1| << |\lambda_2|$$

Exemple : Si l'on prend le problème

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \begin{pmatrix} -10^8 & 0 \\ 0 & -1 \end{pmatrix} X$$

dont on sait que la solution est

$$X(t) = \begin{pmatrix} e^{-10^8 t} \\ e^{-t} \end{pmatrix}$$

avec pour condition initiale

$$X(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

On constate imédiatement en appliquer la première étape de la méthode d'Euler explicite qui est $X_{t=h} = X_{t=0} + hAX_{t=0}$ avec un pas de temps h supérieur à 10^{-8} fait apparaître des valeurs négatives. Cela est un réel handicap si l'on s'intéresse à l'évolution pour un ordre de grandeur temporel plus élevé.

Une équation raide peut souvent décrire un phénomène dont l'évolution est globalement « lente »issu d'un équilibre entre des phénomènes extrêmement « rapides ». Cette appréciation est essentiellement issue du domaine de la chimie où l'on peut trouver des réactions globalement lentes résultant d'une combinaison de réactions extrêmement rapides.

Afin de mieux cerner la raideur d'une équation différentielle, nous présenterons trois exemples représentatifs, dont deux issus de la chimie. Nous mettrons en évidence les facteurs liés à cette raideur pour chacun de ces problèmes.

2.1 Le problème de Curtiss-Hirschfelder

C. F. Curtiss et *J. O. Hirschfelder* proposent un ensemble d'exemples d'équations différentielles raides. Ce sont les équations de la forme :

$$\dot{y} = -k(y - \Phi)$$

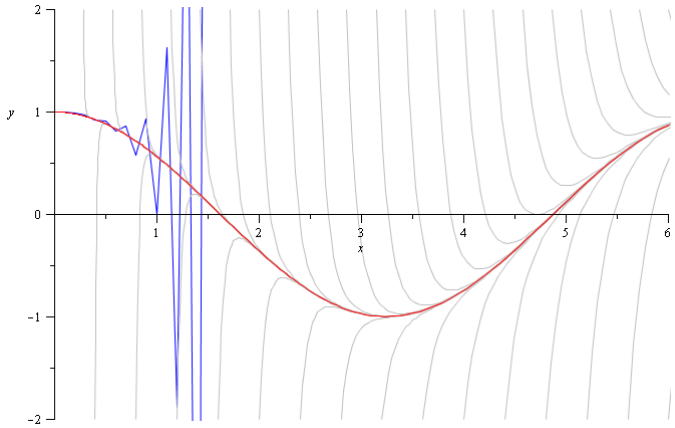
où $k \in \mathbb{R}^+$ et $\Phi(t)$ est une fonction de t. Ces équations possèdent la caractéristique commune de présenter deux échelles de temps. On trouve un comportement global proche de la fonction Φ mais, sur un faible intervalle de temps, une progression de la forme e^{-kt} qui ramène très rapidement toute valeur éloignée de la courbe de Φ vers cette dernière.

Le cas particulier que nous allons traiter correspond à l'équation :

$$\dot{y} = -50(y - \cos(t))$$

La figure 1 présente une solution particulière à ce problème, en rouge, associée à la condition initiale $y(0) = 1$ ainsi que quelques lignes de champ, en gris. On constate immédiatement que cette solution suit bien la courbe de la fonction $\cos(t)$ et que, partant d'une condition initiale $y(t_0) = y_0$, la solution décrit une trajectoire de la forme de e^{-50t} avant de rejoindre ce cosinus.

FIG. 1 – Problème de Curtiss-Hirschfelder



Toutefois, si l'on essaie de résoudre numériquement cette équation avec la méthode d'Euler explicite, que nous rappellerons plus loin dans ce document, nous sommes contraints de prendre un intervalle de temps extrêmement faible pour assurer la stabilité de la méthode, et bien plus encore pour que l'on puisse qualifier cette approximation de satisfaisante, du fait de la présence de fortes oscillations autour de la solution. On peut observer sur ce graphe (en bleu) que l'utilisation de la méthode d'Euler pour un pas de temps $h = 0.1$, ce qui est inférieur au pas que l'on souhaiterait utiliser, la méthode diverge et la solution explose très rapidement, et ceci alors que nous avons volontairement choisi une condition initiale positionnée sur la courbe du cosinus.

Qualitativement, nous pouvons remarquer que la raideur de ce problème provient de la constante $k = 50$ qui, pour un pas h donné, engendre une variation de $-ky_{t_i} * h$ tendent à éloigner y de la solution dès que h s'éloigne de 0, et provoque de si mauvais résultats avec les méthodes de résolution les plus simples.

2.2 Réaction de Belousov-Zhabotinsky : l'Oregonator

La réaction de Belousov-Zhabotinsky est une réaction chimique oscillante. Durant plus d'une centaine de périodes, le potentiel d'oxydo-réduction de la solution oscille avant de finalement se stabiliser après consommation des réactifs limitants.

Elle fut découverte par *Boris Pavlovich Belousov*, chimiste de l'union soviétique durant les années 50. Il chercha deux fois à publier sa découverte qui fut, à deux reprises, refusée. Il reçut comme réponse à sa première demande de publication que sa découverte était « impossible »([Win84]) car contraire au second principe de la thermodynamique[1].

C'est en 1961 qu'*Anatol Zhabothinsky*, étudiant en biophysique à l'Université de Moscou, consacra sa thèse, dirigée par *S. E. Schnoll*, à l'étude approfondie de cette réaction. Conseillé par ce dernier, il remplaça l'un des réactifs, l'acide citrique, par un autre, l'acide malonique, ce qui accrut l'amplitude des oscillations.

Étant donné le contexte politique de l'époque, cette découverte ne parvint en Europe de l'Ouest et aux États-Unis qu'à la conférence qu'*Anatol Zhabothinsky* donna à Prague en 1968.

La réalisation de cette réaction nécessite les quatre réactifs suivants :
- $NaBrO_3$: Bromate de sodium
- H_2SO_4 : Acide sulfurique
- $C_3H_4O_4$: Acide malonique
- NaBR : Bromure de sodium

On les préparera dans deux solutions contenant chacune les réactifs deux à deux, les paires correspondant à l'ordre de leur citation, avant de les réunir. Pour visualiser l'évolution du potentiel d'oxydo-réduction, on ajoutera une solution de ferroïne qui matérialisera le potentiel par une coloration de la solution.

Les réactions mises en jeu sont complexes et nombreuses (cf: [IRE98], [W02]) : 18 réactions et 21 espèces différentes. On utilisera le modèle simplifié suivant (f est un coefficient stœchiométrique ajustable) :

$$\begin{array}{rcll} A + Y & \longrightarrow & X + P & (1) \\ A + X & \longrightarrow & 2X + 2Z & (2) \\ X + Y & \longrightarrow & 2P & (3) \\ 2X & \longrightarrow & A + P & (4) \\ B + Z & \longrightarrow & fY & (5) \end{array}$$

avec la correspondance :

A	B	X	Y	Z	P
BrO_3^-	Matière Organique	HBrO	$HBrO_2$	Br^-	Ce^{4+}

Duquel nous pouvons tirer le système d'équations différentielles :

$$\begin{array}{rcl} [\dot{X}] & = & k_1[A][Y] - k_2[X][Y] + k_3[A][X] - 2k_4[X]^2 \\ [\dot{Y}] & = & -k_1[A][Y] - k_2[X][Y] + \frac{1}{2}k_5 f[B][Z] \\ [\dot{Z}] & = & 2k_3[A][X] - k_5[B][Z] \end{array}$$

où l'on considèrera que, les espèces A et B étant en excès, les concentrations $[A]$ et $[B]$ comme constantes.

[1] Si l'on considère la réaction dans un système fermé, les oscillations forment un phénomène transitoire, qui, après un certain laps de temps, finit par disparaître. Si par contre, l'on considère un système ouvert (ie apport de réactifs pour maintenir certaines concentrations constantes), le phénomène peut se produire indéfiniment, ce que modélise l'Oregonator.

Si l'on tente de résoudre numériquement ce système, à partir des coefficients que l'on peut trouver dans [Fie07] et [EH], on constate, à nouveau, que des méthodes implicites donnent de bien meilleurs résultats.

D'après [EH], la raideur de ce système provient du produit $-k_2[X][Y]$ qui, après une brève phase après laquelle $[Y]$ est de l'ordre 10^3 et $[X]$ proche de 0, constitue une importante valeur négative.

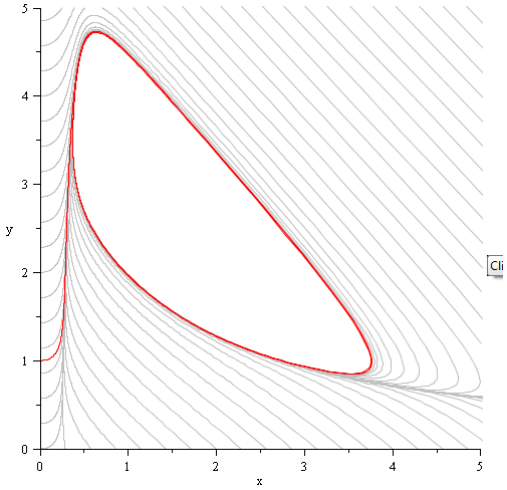
Si l'on observe les graphes, on note un pic pour $[X]$, $[Y]$ et $[Z]$ et d'une brusque chute qui lui succède immédiatement pour $[X]$, à laquelle succède une longue période où l'évolution est lente, avant que ne recommence un cycle. On trouve donc des réactions rapides qui, pourtant, induisent une évolution du système globalement lente, sous réserve d'exclure la phase d'apparition des pics de concentration.

Dans l'exemple étudié dans [EH], résolut avec une méthode à pas de temps h variable, il est nécessaire de faire varier ce dernier entre les ordres de grandeur 10 et 10^{-3}, ce qui est révélateur de la raideur de ce système.

2.3 Réaction-diffusion : Modèle du Brusselator

Il est commun, non seulement dans le domaine de la chimie mais aussi de la physique, de la biologie, et bien d'autres, d'observer une « réaction »accompagnée d'un phénomène de diffusion. On peut d'ailleurs, observer cela avec la réaction de Belousov-Zhabotinsky en versant moins d'un millimètre de solution dans une boîte de Petri. Se forment alors des ondes spirales de potentiel d'oxydo-réduction qui se propagent à la surface du récipient(cf: [IRE98]).

FIG. 2 – Brusselator en l'abscence de diffusion (ie. $\alpha = 0$)



Nous nous intéressons à un problème particulier, qui modélise une réaction autocatalytique : le Brusselator. Ce modèle, qui fut proposé par *Ilya Romanovich Prigozhin*, conduit à un système oscillant. Il met en jeu quatre réactions :

$$\begin{array}{rcll} \mathrm{A} & \longrightarrow & \mathrm{X} & (1) \\ 2\mathrm{X}+\mathrm{Y} & \longrightarrow & 3\mathrm{X} & (2) \\ \mathrm{B}+\mathrm{X} & \longrightarrow & \mathrm{Y}+\mathrm{D} & (3) \\ \mathrm{X} & \longrightarrow & \mathrm{E} & (4) \end{array}$$

Si l'on construit alors le système en ajoutant à tout ceci la diffusion, nous obtenons les équations aux dérivés partielles :

$$\begin{aligned} [\dot{X}] &= \alpha\Delta[X] + k_1[A] + k_2[X]^2[Y] - k_3[B][X] - k_4[X] \\ [\dot{Y}] &= \alpha\Delta[Y] - k_2[X]^2[Y] + k_3[B][X] \end{aligned}$$

où Δ est l'opérateur laplacien[2].

On peut étudier ce système dans le cas particulier où $[A]$ et $[B]$ sont constantes et valent respectivement 1 et 3, et toutes les constantes du systèmes k_i valent 1. C'est-à-dire :

$$\begin{aligned} \dot{x} &= \alpha\Delta x + 1 + x^2y - 4x \\ \dot{y} &= \alpha\Delta y - x^2y + 3x \end{aligned}$$

On reconnaît ici un problème parabolique non linéaire, dont la résolution numérique nous amène à une discrétisation du laplacien. C'est ce dernier point qui est la source de la raideur de ce système. En étudiant le cas d'une seul dimension spaciale, que l'on subdivise avec un pas constant en N valeurs x_i, on peut observer la matrice du laplacien. C'est une matrice tridiagonale de valeurs (1, -2, 1) dont toutes les valeurs propres sont négatives et comprises entre 0 et -4. Dans une telle configuration, nous réécrivons le système de la façon suivante :

$$\begin{aligned} \dot{x_i} &= \alpha.\delta^2 x_i + 1 + x^2y - 4x \\ \dot{y_i} &= \alpha.\delta^2 y_i - x^2y + 3x \end{aligned}$$

Avec $\delta^2 x_i = \frac{(x_{i+1} - 2x_i + x_{i-1})}{\delta x^2}$ et $\delta x = \frac{1}{N}$. Ainsi, la plus faible valeur propre de la diffusion sera proche de $\alpha.\frac{-4}{\delta x^2} = -4\alpha N^2$. On constate que plus la discrétisation est fine, plus le modèle est raide. Nous pouvons prendre un cas particulier où $N = 50$ et $\alpha = \frac{1}{50}$. La plus faible valeur propre de la diffusion est $\lambda \approx -200$, alors que, dans le cas où l'on applique des conditions aux limites de type *Neumann*, la plus grande valeur propre est proche de -1. La raideur est de l'ordre de 200, ce que l'on peut considérer comme moyennement raide.

On peut donc s'attendre à ce que résoudre ce problème nécessite une méthode implicite, ce qui impliquerait de résoudre des systèmes non-linéaires. Il s'avère que la méthode de résolution ROCK4 est particulièrement efficace pour la résolution de systèmes "modérément" raides où les valeurs propres sont réelles et négatives, ce qui correspond au problème posé par la présence de la diffusion, à condition que le coefficient α reste "relativement faible". On utilisera donc cette méthode pour des problèmes similaires.

[2] Dans $\mathbb{R}^3$, il est définit par $\Delta = \frac{\partial^2}{\partial_x^2} + \frac{\partial^2}{\partial_y^2} + \frac{\partial^2}{\partial_z^2}$.

3 Méthodes de Runge-Kutta

Dans la suite de ce chapitre, nous considérerons une équation différentielle résolue du premier ordre. C'est-à-dire le problème :

$$\dot{y} = f(y, t)$$

où f est une fonction de $\mathbb{R}^n \times \mathbb{R} \to \mathbb{C}^n$ et $y \in \{\mathbb{R} \to \mathbb{C}^n\}$.

3.1 Quelques méthodes les plus simples

méthode d'Euler explicite La méthode d'Euler (explicite) est l'une des plus ancienne méthode de résolution des équations différentielles. Elle fut introduite par *Leonhard Paul Euler* au XVIII^{e} siècle. Elle consiste, en pratique, pour une valeur initiale donnée, à tracer un graphe à partir d'une approximation successive des dérivées, en utilisant la fonction f.

En choisissant un pas[3] h, et une valeur initiale y_{t_0} à un temps t_0, on peut approximer la dérivée en ce point par $f(y_{t_0}, t_0)$ et ainsi le prochain point de la courbe au temps $h + t_0$ par $y_{t_0} + hf(y_{t_0}, t_0)$. Cela revient à tracer un segment de la tangente en ce point. En répétant ce processus, on obtient :

$$y_{t_{i+1}} = y_{t_i} + hf(y_{t_i}, t_i) = y_i + hf(y_{t_i}, t_0 + hi)$$

Elle correspond exactement au développement de Taylor à l'ordre 1 de la fonction $y(t)$, c'est à dire :

$$y(t) = y_{t_0} + h\dot{y_{t_0}} + O(h^2) = y_{t_0} + hf(y_{t_0}) + O(h^2)$$

et converge donc bien vers la solution pour $h \to 0$. L'erreur locale est en $O(h^2)$ et l'erreur globale, c'est-à-dire après sommation d'un grand nombre de pas, est donc en $O(h)$. On dira alors que cette méthode est d'ordre 1.

méthode d'Euler implicite (Euler rétrograde) De la même façon, on peut aproximer la dérivée en t_{i+1} par $f(y_{t_{i+1}}, t_{i+1})$ ce qui nous conduit à la méthode :

$$y_{t_{i+1}} = y_{t_i} + hf(y_{t_{i+1}}, t_{i+1})$$

qui nécessite la résolution d'un système, souvent non-linéaire.

Là aussi, cela correspond, pour $h \to 0$, c'est-à-dire $y_{i+1} \to y_i$, au développement de Taylor à l'ordre 1.

La grande différence entre Euler explicite et Euler rétrograde est le domaine de stabilité de ces deux méthodes. (cf 3.3)

3.2 Méthodes de Runge-Kutta

Les méthodes d'Euler sont relativement imprécises. Pour en trouver de plus précises, nous allons nous intéresser à une famille de méthodes particulières.

[3] On le considèrera comme constant, mais on peut tout à fait imaginer le faire varier entre deux pas[4] de la méthode. On posera alors simplement $h = t_{i+1} - t_i$.

Définition 1.1 : Une méthode de *Runge-Kutta* d'ordre p à s étages est une méthode dont l'erreur locale est en $O(h^{p+1})$ définie par les expressions :

$$\forall i \in \{1,\dots,s\} : Y_i = y_{t_n} + h\Sigma_{j=1}^{s} a_{ij} f(Y_j, t_n + c_j h) \quad y_{t_{n+1}} = y_n + h\Sigma_{j=1}^{s} b_j f(Y_i, t_n + c_j h)$$

ainsi que ses coefficients présentés sous la forme d'un tableau de Butcher :

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

Une méthode est donc explicite si et seulement si $\exists (i,j) \in \{1,\dots,s\}^2, j \geq i \mid a_{ij} \neq 0$.

De plus, on parlera de méthode diagonalement implicite pour une méthode implicite où $\forall (i,j) \in \{1,\dots,s\}^2 \mid j > i \Rightarrow a_{ij} = 0$.

Définition 1.2 : La méthode RK4 classique est une méthode de Runge-Kutta, datant du début du XX^{e} siècle, dont le tableau de Butcher est

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

Remarque : L'écriture de ces méthodes sous la forme d'un tableau de Butcher est relativement récente. On peut tout à fait définir RK4 comme une sucession de plusieurs calculs, et la notation, là aussi, varie. Dans ce document, nous ferons d'ailleur usage de deux notations différentes, en utilisant la plus adaptée à la situation, après l'avoir rappelée.

On vérifiera par la suite que cette méthode est bien d'ordre 4.

Remarque: Il apparaît donc que les méthodes d'Euler et Euler rétrograde sont des méthodes de Runge-Kutta à 1 étage ayant pour tableau, respectivement $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$ et $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$.

3.3 Les domaines de stabilité

L'ordre n'est pas la seule condition importante entrant en jeu dans le choix d'une méthode de Runge-Kutta. La seconde condition importante est le domaine de stabilité de la méthode. Cela correspond aux points du plan complexe z pour lesquels la méthode reste stable, c'est-à-dire que les résultats restent bornés.

Nous allons étudier un certain critère de stabilité d'une méthode, appelé *critère de Dahlquist*. On considère le problème linéaire :

$$\dot{y} = \lambda y$$

Cherchons alors à exprimer le résultat d'un pas de cette méthode comme une fraction rationnelle en λh[5] :

[5] Dans le cas d'une méthode implicite, nous aurons affaire à une fraction rationelle, alors que dans le cas d'une fonction explicite, nous aurons un polynôme. Ceci peut se déduire de l'expression sous la forme d'une somme d'une méthode de Runge-Kutta.

$$y_{t_{i+1}} = R(z)y_{t_i} \quad \text{où} \quad z = \lambda h \in \mathbb{C}$$

où $R(z)$ est appelée fonction de stabilité.

La fonction de stabilité de la méthode d'Euler (explicite) est donc

$$R(z) = 1 + z \quad \text{car} \quad y_{t_{i+1}} = (1+z)y_{t_i}$$

ce qui correspond aux deux premiers termes de la série exponentielle, solution de ce problème. De même, fonction de stabilité de la méthode d'Euler rétrograde (Euler implicite) est :

$$R(z) = \tfrac{1}{1-z} \quad \text{provenant de} \quad y_{t_{i+1}} = y_{t_i} + z y_{t_{i+1}}$$

Le domaine de stabilité d'une méthode est la partie du plan complexe pour laquelle la fonction de stabilité respecte la condition $|R(z)| \leq 1$. Cela correspond aux valeurs pour lesquelles la méthode ne diverge pas[6].

Pour Euler explicite, le domaine de stabilité est le cercle de centre $z = -1$ et de rayon $|z| = 1$.

Pour Euler rétrograde, le domaine de stabilité est la totalité du plan complexe privé du disque ouvert de centre 1. Puisque cette méthode contient le demi-plan à valeurs réelles négatives, on dit que cette méthode est A-stable. Mieux, du fait que $R(-\infty) = 0$ on dit que cette méthode est L-stable. Cela signifie qu'elle est particulièrement stable pour des valeurs propres négatives, même très faibles (i.e. élevées en module).

3.4 Méthode des directions alternées

La méthode des directions alternées est historiquement une solution à la résolution d'équations paraboliques et elliptiques alors que les ordinateurs disposaient de mémoire et capacité de calcul faibles.

Considérons l'équation de la chaleur en dimension deux

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \Delta u \tag{1}$$

On peut alors chercher à résoudre les problèmes

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial^2 x} \qquad\qquad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial^2 y} \tag{2a}$$

On traite alors les problèmes en discrétisant les laplaciens, c'est-à-dire en considérant une grille de valeurs u_{ij}, puis en posant $\frac{\partial^2 u}{\partial^2 x} = u_{(i+1)j} - 2u_{ij} + u_{(i-1)j}$ et $\frac{\partial^2 u}{\partial^2 y} = u_{i(j+1)} - 2u_{ij} + u_{i(j-1)}$, ce que nous pouvons résoudre avec des méthodes numériques.

Si l'on condidère alors les flots[7]de ces problèmes, où $\phi_t^{[1]}(y)$ et $\phi^{[2]}(t)$ sont ceux des deux problèmes, et $\phi_t(y)$ le flot du problème 1, on peut alors chercher à approximer $\phi_t(y)$ par $(\phi_t^{[1]} \circ \phi_t^{[2]})(t)$. Cette solution est exacte dans le cas où $\phi_t^{[1]}$ et $\phi_t^{[2]}$ commutent, ce qui est le cas du laplacien.

[6] Ainsi, puisqu'un polynôme n'est pas borné, une méthode explicite ne peut qu'avoir un domaine de stabilité fini.

[7] Le flot d'une équation différentielle $\dot{y} = f(y)$ et d'une condition initiale $y(t_0) = y_0$ est l'application $\phi_t(y_0) : y \mapsto \phi_t(y_0)$ où $\phi_t(y_0)$ est la valeur de la solution du problème pour la condition initiale donné à l'instant t.

Elle est aujourd'hui utilisée pour subdiviser des problèmes (notamment les problèmes de réaction/diffusion) et résoudre ceux-ci avec des méthodes spécifiques. Par exemple, un problème de la forme $\dot{y} = D(y) + R(y)$ où D est la diffusion et R une réaction raide, pourra être résolue en utilisant $\Phi^{[1]}$ et $\Phi^{[2]}$ des méthodes spécifiques, puis la solution générale obtenue par $\Phi^{[2]}_{t/2} \circ \Phi^{[1]}_t \circ \Phi^{[2]}_{t/2}$. Cette approximation en trois termes introduit une erreur absolue en $O(h^3)$ et la précédente, à deux termes, en $O(h^2)$.

4 Ordre d'une méthode de Runge-Kutta

Dans la suite de ce chapitre, nous considérerons le problème autonome

$$\dot{y} = f(y) \quad y(t_0) = y_0 \quad f : \mathbb{R}^n \to \mathbb{R}^n$$

On rappelle qu'un problème de la forme $\dot{y} = f(y, t)$ peut se ramener au problème précédent en ajoutant l'équation $\dot{t} = 1$ au système.

Nous allons chercher à trouver une condition nécessaire, suffisante et élégante pour qu'une méthode de Runge-Kutta, dont le tableau de Butcher est donné, soit d'ordre p. Pour ce faire, il nous faudra introduire quelques notations et formalisme qui nous permettront d'exprimer cette condition.

4.1 Le formalisme des arbres enracinés

Notation 2.1 (Formes p-linéaires) : On notera $f'(y)$ le jacobien de $f(y)$, $f''(y)$ la dérivée seconde et $f^{(p)}(y)$ la dérivée p$^{\text{ième}}$ de f. Enfin, on notera chacun des produits formés de ces dérivées p$^{\text{ième}}$ ainsi que d'autres termes comme des applications p-linéaires. Par exemple, on notera $f''(y)(a, b)$ le produit $f''(y)ab$. Ceci est possible grâce aux propriétés de symétrie des dérivées partielles.

Il devient alors possible d'exprimer les dérivées temporelles de y comme combinaisons linéaires de ces applications p-linéaires en utilisant les propriétés sur la dérivée de fonctions composées[8] ainsi que la dérivation d'un produit[9].

$$\begin{array}{lcl}
\dot{y} & = & f(y) \\
\ddot{y} & = & f'(y)(\dot{y}) \\
y^{(3)} & = & f''(y)(\dot{y}, \dot{y}) + f'(y)(\ddot{y}) \\
y^{(4)} & = & f^{(3)}(y)(\dot{y}, \dot{y}, \dot{y}) + 3f''(y)(\ddot{y}, \dot{y}) + f'(y)(y^{(3)}) \\
y^{(5)} & = & f^{(4)}(y)(\dot{y}, \dot{y}, \dot{y}, \dot{y}) + 6f^{(3)}(y)(\ddot{y}, \dot{y}, \dot{y}) + 4f''(y)(y^{(3)}, \dot{y}) + 3f''(y)(\ddot{y}, \ddot{y}) + f'(y)(y^{(4)})
\end{array}$$

[8] Dérivation d'une fonction composée :

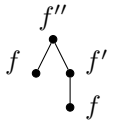
$$\frac{\mathrm{d}f'(y)}{\mathrm{d}t} = \frac{\mathrm{d}f'(y)}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}t}$$

[9] Dérivation d'un produit de fonctions :

$$\frac{\mathrm{d}\prod_{i\in I} f_i(t)}{\mathrm{d}t} = \sum_{i\in I} \frac{\mathrm{d}f_i(t))}{\mathrm{d}t} \prod_{j\in I, j\neq i} f_j(t)$$

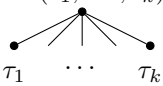
Nous pouvons alors, en itérant sur le nombre de dérivations de y par rapport à t, exprimer chaque $y^{(i)}$ en fonction des dérivées p$^{\text{ièmes}}$ et de $f(y)$. En allégeant quelque peu la notation par l'omission du paramètre (y), nous obtenons finalement :

$$\begin{array}{lcl}
\dot{y} & = & f \\
\ddot{y} & = & f'(f) \\
y^{(3)} & = & f''(f,f) + f'(f'(f)) \\
y^{(4)} & = & f^{(3)}(f,f,f) + 3f''(f'(f),f) + f'(f''(f,f)) + f'(f'(f'(f))) \\
y^{(5)} & = & f^{(4)}(y)(f,f,f,f) + 6f^{(3)}(y)(f'(f),f,f) + 4f''(f''(f,f)) + 4f''(f'(f'(f)),f) + 3f''(y)(f'(f),f'(f)) \\
 & & +f'(f^{(3)}(f,f,f)) + 3f'(f''(f'(f),f)) + f'(f'(f''(f,f))) + f'(f'(f'(f'(f))))
\end{array}$$

On peut alors schématiser chaque terme de cette combinaison linéaire sous la forme d'un arbre enraciné, où chaque noeud, représentant une dérivée q$^{\text{ième}}$, a q branches rattachées aux q opérandes. Par exemple, l'expression $f''(f, f'(f))$ sera représentée par l'arbre 

. Ceci nous amène donc à la série de définitions suivantes.

Définition 2.2 (Arbres) : L'ensemble des arbres enracinés $\mathbb{T}$ est défini de façon récursive comme suit :

- L'arbre $\bullet$, avec un unique sommet appellé racine, appartient à l'ensemble $\mathbb{T}$
- Soit $(\tau_1, \ldots, \tau_k) \in \mathbb{T}^k$, alors l'arbre $[\tau_1, \ldots, \tau_k]$ est un élément de $\mathbb{T}$, correspondant à l'arbre 

formé d'une racine à laquelle est reliée chacun des sous-arbres τ_i. On rappelle que l'ordre dans lequel figurent les sous-arbres τ_i n'a aucune importance, c'est-à-dire que $[\tau_1, \ldots, \tau_i, \ldots, \tau_j, \ldots \tau_k] = [\tau_1, \ldots, \tau_j, \ldots, \tau_i, \ldots \tau_k]$

Définition 2.3 (Différentielles élémentaires) : On définit l'application $F : \mathbb{T} \to \mathbb{R}^{n\mathbb{R}^n}$ récursivement par :

- $F(\bullet) = f = y \mapsto f(y)$
- Soit $\tau = [\tau_1, \ldots, \tau_k] \in \mathbb{T}$, alors $F(\tau) = f^{(k)}(F(\tau_1), F(\ldots), F(\tau_k))$

On appelle alors chacune des expressions $F(\tau)$ une **différentielle élementaire**.

Notation 2.4 (Arbres) : On notera :

- $|\tau|$ le nombre de sommets de l'arbre $\tau \in \mathbb{T}$.
- $\alpha(\tau) \in \mathbb{N}$ le coefficient qui apparaît devant l'expression $F(\tau), \tau \in \mathbb{T}$. Ce dernier est bien défini car les expressions apparaissant dans une dérivée p$^{\text{ième}}$ de y est décrite par un arbre à $|\tau|$ sommets, ce qui assure l'unicité de ce coefficient.

De ces notations et définitions découle alors notre premier théorème.

Théorème 2.5 : La dérivée p$^{\text{ième}}$ de $y(t)$ est donnée par l'expression

$$y^{(p)}(t_0) = \sum_{|\tau|=p} \alpha(\tau) F(\tau)(y_0)$$

Démonstration : Il s'agit d'une récurrence sur p ; on démontre que simplement que les dérivées s'expriment avec des sommes à p termes constitué de dérivées de f, ce qui revient à dire que l'on peut l'écrire comme une somme d'arbres à p sommets.

4.2 Identification des coefficients d'une méthode de Runge-Kutta

Nous allons maintenant effectuer un travail similaire, non plus avec une solution exacte, mais avec une méthode de Runge-Kutta à q étages, dont le tableau de Butcher est donné. Si l'on note alors

$$g_i = hf(u_i) \quad (1) \tag{3}$$

nous avons (avec $(i,j) \in \{1,\dots,q\}^2$)

$$u_i = y_0 + \textstyle\sum_j a_{ij} g_j \quad y_{t_0+h} = y_0 + \sum_i b_i g_i \tag{4}$$

où la première expression est une étape intermédiaire, précédemment notée Y_i, de la méthode et la seconde est le résultat de la méthode appliquée avec un pas h. Les expressions u_i, g_i et y_{t_0+h} sont des fonctions du pas de temps h. Nous allons donc nous intéresser à leurs dérivées évaluées en $h = 0$.

Rappelons la formule de *Leibniz*

$$(hf)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} \; h^{(k)} \; f^{(n-k)}$$

où $\binom{n}{k}$ est le coefficient binomial défini par

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Si nous appliquons cette expression à g_i en tenant compte que $\forall k \geq 2, h^{(k)} = 0$ nous obtenons

$$g_i^{(k)} = h[f(u_i)]^{(k)} + k[f(u_i)]^{(n-1)}$$

En faisant tendre $h \to 0$ on obtient alors

$$g_i^{(k)}\Big|_{h=0} = k[f(u_i)]^{(n-1)} \tag{5}$$

En utilisant les mêmes règles qu'au paragraphe précédent nous parvenons aux même expressions obtenues par la solution exacte, à ceci près qu'apparaît maintenant un facteur entier supplémentaire. Si nous calculons les premières expressions en $h = 0$, nous obtenons :

$$\begin{aligned}
\dot{g}_i &= 1.f(y_0) \\
\ddot{g}_i &= 2.f'(y_0)(\dot{u}_i) \\
g_i^{(3)} &= 3.\left(f''(y_0)(\dot{u}_i,\dot{u}_i) + f'(y)(\ddot{u}_i)\right) \\
g_i^{(4)} &= 4.\left(f^{(3)}(y_0)(\dot{u}_i,\dot{u}_i,\dot{u}_i) + 3f''(y_0)(\ddot{u}_i,\dot{u}_i) + f'(y_0)(u_i^{(3)})\right) \\
g_i^{(5)} &= 5.\left(f^{(4)}(y_0)(\dot{u}_i,\dot{u}_i,\dot{u}_i,\dot{u}_i) + 6f^{(3)}(y_0)(\ddot{u}_i,\dot{u}_i,\dot{u}_i) + 4f''(y_0)(u_i^{(3)},\dot{u}_i) + 3f''(y_0)(\ddot{u}_i,\ddot{u}_i) + f'(y_0)(u_i^{(4)})\right)
\end{aligned}$$

Intéressons-nous aux dérivées des u_i. Nous obtenons à partir de (2) l'équation

$$u_i^{(k)} = \sum_j a_{ij} g_j^{(k)} \tag{6}$$

Il nous est donc possible de remplacer chaque dérivée des u_i figurant dans les g_i par leur expression, et nous obtenons

$$\begin{array}{lcllcl}
\dot{g_i} & = & 1.f & \dot{u_i} & = & 1.\left(\sum_j a_{ij}\right) f \\
\ddot{g_i} & = & (1.2)\left(\sum_j a_{ij}\right) f'(f) & \ddot{u_i} & = & (1.2)\left(\sum_{jk} a_{ij}a_{jk}\right) f'(f) \\
g_i^{(3)} & = & (1.3)\left(\sum_{jk} a_{ij}a_{ik}\right) f''(f,f) & u_i^{(3)} & = & (1.3)\left(\sum_{jkl} a_{ij}a_{jk}a_{jl}\right) f''(f,f) \\
 & & +(1.2.3)\left(\sum_{jk} a_{ij}a_{jk}\right) f'(f'(f)) & & & +(1.2.3)\left(\sum_{jkl} a_{ij}a_{jk}a_{kl}\right) f'(f'(f))
\end{array}$$

Définition 2.6 (Facteurs entiers) : On définit l'application $\gamma : \mathbb{T} \to \mathbb{N}$ récursivement par
- $\gamma(\bullet) = 1$
- $\gamma(\tau) = |\tau|\gamma(\tau_1)\dots\gamma(\tau_k)$ où $\tau = [\tau_1, \dots, \tau_k]$

On appellera $\gamma(\tau)$ un **facteur entier**.

Définition 2.7 (Facteurs somme) : On définit l'application $\mathbf{g}_i(\tau) : \mathbb{T} \to \mathbb{R}$ récursivement par
- $\mathbf{g}_i(\bullet) = 1$
- $\mathbf{g}_i(\tau) = \mathbf{u}_i(\tau_1)\dots\mathbf{u}_i(\tau_k)$

On définit aussi l'application $(u)_i(\tau) : \mathbb{T} \to \mathbb{R}$ par

$$\mathbf{u}_i(\tau) = \sum_j a_{ij}\mathbf{g}_j(\tau) = \sum_j a_{ij}\mathbf{u}_i(\tau_1)\dots\mathbf{u}_i(\tau_k)$$

Lemme 2.8 (Expression des dérivées) : Les dérivées p$^{\text{ième}}$ de g_i et u_i en $h =$ sont exactement :

$$g_i^{(p)}\Big|_{h=0} = \sum_{|\tau|=p} \gamma(\tau).\mathbf{g}_i(\tau).\alpha(\tau).F(\tau)(y_0)\ u_i^{(p)}\Big|_{h=0} = \sum_{|\tau|=p} \gamma(\tau).\mathbf{u}_i(\tau).\alpha(\tau).F(\tau)(y_0)$$

Démonstration : Ces formules sont exactes pour $i = 1$. De plus, de par le fait que les expressions ne diffèrent que par des constantes avec les expressions précédentes, et que l'on peut toujours les factoriser, il est immédiat que l'on retrouve l'expression $\alpha(\tau).F(\tau)(y_0)$. On effectue alors une récurrence sur p, avec $|\tau| = p+1$, $\tau = [\tau_1, \dots, \tau_k]$ et $l \in \{1, \dots, k\}$. En appliquant 5 à $u_i^{(p)}\Big|_{h=0}$ on constate, dans un premier temps, que l'on multiplie l'expression par $p+1 = |\tau|$ ce qui correspond bien au facteur entier $\gamma(\tau)$. Dans un second, on remarque que pour chaque sous-arbre τ_l de τ on à un facteur additionnel $\mathbf{u}_i(\tau_l)$, et par factorisation en retrouve le produit qui définit $\mathbf{g}_i(\tau)$. Enfin, 6 conclut la démonstration en nous donnant immédiatement $\mathbf{g}_i(\tau)$

Définition 2.9 (poids élémentaires) : On définit l'application $\phi : \mathbb{T} \to \mathbb{R}$ par

$$\phi(\tau) = \sum_i b_i\mathbf{g}_i(\tau) \tag{7}$$

Remarque : On peut facilement calculer ce coefficient pour un arbre τ donné, en attachant à la racine de τ la lettre i, puis les suivantes à chaque autre sommet. Le coefficient $\phi(\tau)$ est alors la somme des produits regroupant les termes bi, et a_{jk} où j est un sommet de τ et k la racine d'un sous-arbre de τ. Cette propriété se démontre par récurrence sur le nombre de sommets et

la valence de la racine, en découpant chaque arbre en ses sous-arbres, et en calquant le processus de construction des g_i.

Ces résultats nous permettent finalement d'obtenir un théorème sur la forme de la solution numérique d'une méthode de Runge-Kutta.

Théorème 2.10 : La dérivée p$^{\text{ième}}$ de la solution issue d'une méthode de Runge-Kutta à q étages, prise en $h = 0$ est

$$y^{(p)}_{t_0+h}\Big|_{h=0} = \sum_{|\tau|=p} \gamma(\tau)\phi(\tau)\alpha(\tau)F(\tau)(y_0) \tag{8}$$

4.3 Condition nécessaire et suffisante sur l'ordre

Théorème 2.11 : Une méthode de Runge-Kutta est d'ordre p si et seulement si

$$\forall |\tau| \leq p \quad \theta(\tau) = \tfrac{1}{\gamma(\tau)} \tag{9}$$

Démonstration : Si la condition est vérifiée, alors d'après les Théorèmes 2.10 et 2.5 toutes les dérivées jusqu'à l'ordre p de la méthode sont égales aux dérivées de la solution exacte pour $h \to 0$. Elle est donc bien suffisante. De plus, les différentielles élémentaires $F(\tau)$ sont indépendantes (ie. Une différentielle à k termes n'apparaît que dans l'expression de la dérivée k$^{\text{ième}}$ et pour chaque expression d'une dérivée k$^{\text{ième}}$ on factorise pour ne laisser apparaître qu'une unique fois chaque différentielle élémentaire) ce qui permet d'affirmer la nécessité de cette condition.

Un récapitulatif des coefficients, différentielles élémentaires et arbres jusqu'à l'ordre 4 se trouve dans le tableau 1.

Exemple 2.12 : Pour l'arbre  nous obtenons les conditions

$$\sum_{ijklmn} b_i a_{ij} a_{ik} a_{il} a_{lm} a_{mn} = \frac{1}{6.3.2} = \frac{1}{36}$$

Remarque : Il est bon de préciser que la condition pour qu'une méthode de Runge-Kutta soit d'ordre 2 est identique à la condition imposée par un développement de Taylor dans le où f est linéaire (ie. $f(y) = \lambda y$), à savoir $\sum_i b_i = 1$ et $\sum_{ij} b_i a_{ij} = \frac{1}{2}$.

5 ROCK 2

La méthode ROCK2 est une méthode de Runge-Kutta **explicite** d'ordre 2 qui a pour particularité de réunir les conditions suivantes :

- Un intervalle de stabilité quasi-optimal
- Une relation de récurrence à trois termes
- Une erreur absolue en $O(h^3)$

TAB. 1 – Coefficients, différentielles élémentaires et arbres

$\|\tau\|$	τ	*Graph.*	$\alpha(\tau)$	$F(\tau)$	$\gamma(\tau)$	$\phi(\tau)$
1	$\bullet$		1	f	1	$\sum_i b_i$
2	$[\bullet]$		1	$f'f$	2	$\sum_{ij} b_i a_{ij}$
3	$[\bullet,\bullet]$		1	$f''(f,f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$
3	$[[\bullet]]$		1	$f'f'f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$
4	$[\bullet,\bullet,\bullet]$		1	$f'''(f,f,f)$	4	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{il}$
4	$[[\bullet],\bullet]$		3	$f''(f'f,f)$	8	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{jl}$
4	$[[\bullet,\bullet]]$		1	$f'f''(f,f)$	12	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{jl}$
4	$[[[\bullet]]]$		1	$f'f'f'f$	24	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{kl}$

La présence d'intervalles de stabilité quasi-optimaux et d'une relation de récurrence permettent d'obtenir de bonnes approximations d'une solution d'une équation différentielle moyennement raide avec un pas de temps important. C'est en voulant réunir ces trois conditions que *Alexei A. Medovikov* et *Assyr Abdulle* construisirent cette méthode (cf: **??**).

Dans cette partie, nous décrirons la méthode employée pour construire cet algorithme. Nous admettrons quelques théorèmes qui seront rappelés aux moments opportuns.

5.1 Fonction de stabilité et polynômes orthogonaux

Considérons les polynômes optimaux pour une méthode de Runge-Kutta explicite d'ordre 2. On notera $R_s(z)$ la fonction de stabilité optimale formée d'un polynôme de degré s. On aura donc $\forall s \in \mathbb{N}, z \in \mathbb{C} e^z - R_s(z) = O(h^3)$. Les conditions d'optimalité sont :

$$\overline{R}_s(z) = 1 + z + \frac{z^2}{2} + \sum_{i=3}^{s} \alpha_{i,s} z^i, \alpha_{i,s} \in \mathbb{R} \tag{10}$$

$$|\overline{R}_s(z)| \leq 1, \forall z \in [-\bar{l}_s, 0] \tag{11}$$

où $\bar{l}_s$ est maximal.

Nous admettrons([Sze59]) que ces polynômes existent et sont uniques, pour tout ordre de méthode et tout degré donné. De plus, nous admettrons que ces polynômes sont **equi-oscillants**

en $s-1$ points, i.e.

$$\begin{array}{c} \exists -\bar{l}_s < z_1 < \cdots < z_{s-2} < 0| \\ \forall i \in 0,\ldots,s-3, \overline{R}_s(z_i) = -\overline{(z_{i+1})} \\ \forall i \in 0,\ldots,s-2, |\overline{R}_s(z_i)| = 1 \end{array} \quad (12)$$

Nous admettrons aussi que ces polynômes possèdent deux racines complexes, et ainsi $s-2$ racines réelles.

Nous allons légèrement modifier la condition 11 par $|\overline{R}_s(z)| \leq \eta < 1$, et nous prendrons $\eta = 0.95$. Cela a pour effet de réduire légèrement l'intervalle de stabilité, mais permettra de tenir compte de possibles erreurs d'arrondi.

Remarque 1 : Aucune solution analytique explicite n'est connue pour les polynômes optimaux d'ordre 2. Ils sont, en pratique, calculés par des méthodes numériques.

Remarque 2 : Les polynômes optimaux pour les méthodes d'ordre 1 sont les polynômes de Tchebychev T_s, appliqué à un polynôme en z. Plus exactement, $\overline{R}_s(z) = T_s(1+\frac{z}{s^2})$. Ils respectent des conditions similaires à 10, 11 et 12. De plus, $\bar{l}_s = 2s^2$. On rappelle que les polynômes de Tchebychev sont orthogonaux[10] pour la fonction de poids $1/\sqrt{(1-x^2)}$. On rappelle aussi que s'il existe une suite de polynômes orthogonaux pour un poids donné, alors tout autre suite de polynômes orthogonaux est constituée des mêmes polynômes(à un facteur multiplicatif près)[11].

Une propriété remarquable de toute suite de polynômes orthogonaux est :

Lemme 3.1 (Relation de récurrence des polynômes orthogonaux) : Toute suite de polynômes $p_n(x)$ orthogonaux pour le produit scalaire $\langle .,. \rangle$ présente une relation de récurrence à trois termes

$$p_{n+1} = (a_n x + b_n)p_n - c_n p_{n-1}$$

Démonstration : Puisque p_n est de degré n, alors $(p_k)_{k=0}^n$ forme une base de $\mathbb{C}_{n-1}[x]$. De plus, $\exists a_n, b_n$ tq. $P(x) = (a_n x + b_n)p_n - p_{n+1}$ et $deg(P) < 0$. En effet, si $p_n = e_n z^n + e_{n-1} z^{n-1} + \ldots$ et $p_{n+1} = h_{n+1} z^{n+1} + h_n z^n + \ldots$ il suffit de prendre $a_n = \frac{h_{n+1}}{e_n}$ et $b_n = \frac{h_n}{a_n}$. On a alors

$$(a_n x + b_n)p_n - p_{n+1} = \sum_{j=0}^{n-1} \mu_{n,j} p_j$$

En appliquant $\langle .,. \rangle$ à cette expression, nous obtenons :

$$\langle p_j, \sum_{j=0}^{n-1} \mu_{n,j} p_j \rangle = \langle (a_n x + b_n)p_n - p_{n+1}, p_j \rangle = \langle a_n x p_n, p_j \rangle + \langle b_n p_n, p_j \rangle - \langle p_{n+1}, p_j \rangle$$

Par orthogonalité, il vient

$$\forall 1 \leq j \leq n-1, \langle p_j, p_j \rangle \mu_{n,j} = a_n \langle x p_n, p_j \rangle$$

[10] Soit un produit scalaire défini par $\langle f, g \rangle = \int_a^b f(x)g(x)\omega(x)\mathrm{d}x$ où $\omega(x)$ est appelé fonction de poids. Deux polynômes $P(x)$ et $Q(x)$ sont dit orthogonaux si et seulement si $\langle P, Q \rangle = 0$ On appelle $[a, b]$ l'intervalle d'orthogonalité

[11] En effet, considérons deux suites de polynômes (a_n) et (b_n) pour un poids $w(x)$. Soit $m \in \mathbb{N}$. Le polynôme a_m peut s'écrire comme combinaison linéaire des polynômes b_i pour $i \in 0,\ldots,m$. De plus, si l'on calcule le produit scalaire $\langle b_m, a_m \rangle = \sum_i \alpha_i \langle b_m, b_i \rangle$ de par l'orthogonalité des polynômes de la suite (b_n), seul le terme $\langle b_m, b_i \rangle$ subsiste. On en déduit que $a_m = C b_m$.

De plus, de par la définition du produit scalaire, nous avons $\langle xp_n, p_j\rangle = \langle p_n, xp_j\rangle$. Pour $j < n-1$, ce produit est nul, donc :

$$c_n := \langle p_j, p_j\rangle \mu_{n,j} = a_n \langle xp_n, p_{n-1}\rangle$$

d'où l'on conclut

$$(a_n x + b_n)p_n - p_{n+1} = c_n p_{n-1}$$

Nous allons chercher à construire une suite de polynômes orthogonaux qui approximent les polynômes optimaux. Pour nous faciliter la tâche, nous travaillerons dans l'intervalle $[-1, 1]$ en posant $x = 1 + \frac{2z}{\overline{l}_s}$. Posons

$$\overline{R}_s(x) = \overline{\omega}(x)\overline{P}_{s-2}(x) \tag{13}$$

où $\overline{w}$ est le polynôme de degré 2 possédant les deux racines complexes de $\overline{R}_s(x)$. Tenant compte du fait que $\overline{R}_s(1) = 1$ nous pouvons supposer, sans perte de généralité, que $\overline{\omega}(1) = 1 = \overline{P}_{s-2}(x)$. Nous pouvons réécrire les conditions 10 et 11 sous la forme

$$\overline{R}_s(1) = 1, \overline{R}'_s(1) = \overline{d}_s, \overline{R}''_s(1) = \overline{d}_s^2 \tag{14}$$

où $\overline{d}_s = \frac{\overline{l}_s}{2}$, et

$$\max_{x\in[-1,1]} |\overline{R}_s(x)| = 1 \tag{15}$$

Nous admettrons(**??**, **??**) le théorème suivant :

Théorème 3.2 : Soit $q(x)$ une fonction de poids positive sur $[-1, 1]$, tel que $0 < \lambda < q(x) < \Lambda$ et $|q(x+\delta) - q(x)||\ln\delta^{1+\epsilon} < L$ (où $\lambda, \Lambda, \epsilon, L$ sont des constantes positives). Alors, les polynômes othogonaux P_n associé à la fonction de poids $q(x)/(\sqrt{1-x^2})$ respectent, $\forall x \in [-1, 1]$:

$$q(x)^{1/2}P_n(x) = cos(n\theta + \psi(\theta)) + O(ln(n)^{-\epsilon}), \theta := arccos(x) \tag{16}$$

où ψ, appelée fonction de phase de Szegö-Bernstein, est continue et positive tel que $\psi(0) = \psi(\pi) = 0$.

Appliquons ce théorème en posant $q(x) = \overline{\omega}(x)^2$, et $n = s - 2$. D'après 16, et d'après la continuité de $n\theta + \psi(\theta)$, $\overline{\omega}\overline{P}_{s-2}$ oscille, à $O(ln(n)^{-\epsilon})$ près, entre -1 et 1 en $s - 1$ points sur $[-1, 1]$.

Nous allons donc utiliser le polynôme P_{s-2} plutôt que $\overline{P}_{s-2}$. Notre problème devient alors :
- Trouver un polynôme $\omega(x)$ de degré 2 (qui peut dépenre de $s - 2$)
- Construire P_{s-2} associé à la fonction de poids $\omega(x)^2/(\sqrt{1-x^2})$ de façon à ce que

$$R_s(x) = \omega(x)P_{s-2}(x) \tag{17}$$

soit le polynôme d'une fonction de stabilité d'ordre 2, tout en cherchant à maximiser l_s

5.2 Construction des coefficients dont dépend R_s

Nous allons chercher à construire le polynôme de stabilité $R_s(x)$. Dans un premier temps, posons

$$\omega(x) = (x - (\alpha + i\beta))(x - (\alpha - i\beta))$$

. Pour $s \geq 3$, P_{s-2} est le polynône orthogonal associé à la fonction de poids $\omega(x)^2/(\sqrt{1-x^2})$. Si nous normalisons $P_{s-2}(x)$ de façon à ce que $|\omega(x)P_{s-2}(x)| \leq 1$ pour $x \in [-1,1]$, il est peu probable que $\omega(1)P_{s-2}(1) = 1$, c'est pourquoi nous allons poser

$$R_s(x) = \frac{\omega(x)P_{s-2}(x)}{\omega(a)P_{s-2}(a)} \tag{18}$$

où $a \in [-1,1]$ est proche de 1. Nous travaillerons par la suite sur $[-1,a]$ et poserons les conditions d'ordre en a.

On notera parfois $R_s(x,\alpha,\beta)$ pour faire ressortir le fait que les coefficients de ce polynôme dépendent de ces variables. De plus, de par la structure du polynôme ω, nous pouvons sans perte de généralité imposer $\beta \geq 0$.

Nous posons alors le problème suivant :

$$R'_s(a,\alpha,\beta) = d, R''_s(a,\alpha,\beta) = d^2 \tag{19}$$

$$\forall x \in [-1,a], |R_s(x)| \leq 1 \tag{20}$$

$$l = (1+a)d \text{ aussi grand que possible} \tag{21}$$

Notons que par le changement de variable $z = (x-a)d$ et en posant $\hat{R}_s(z) = R_s(a + \frac{z}{d})$ nous obtenons les conditions

$$\hat{R}_s(0) = 1, \hat{R}'_s(0) = 1, \hat{R}''_s(0) = 1 \tag{22}$$

$$\forall z \in [-l,0], |\hat{R}_s(z)| \leq 1 \tag{23}$$

Pour construire les coefficients a, d, α, β, les auteurs proposent un algorithme en deux étapes.

5.2.1 Etape 1 de l'algorithme

Pour des valeurs a et d donné, cherchons à déterminer α_{nv} et β_{nv} tel que la condition 19 soit respectée. Nous devons donc résoudre le système d'équations (non linéaire) suivant :

$$\begin{array}{c} R'_s(a,\alpha_{nv},\beta_{nv}) = d \\ R''_s(a,\alpha_{nv},\beta_{nv}) = d^2 \end{array} \tag{24}$$

Il existe des formules explicites, bien que d'un esthétisme douteux, pour calculer les valeurs de $R_s(x)$ et ses dérivées en x, ce qui permet de résoudre numériquement le système d'équations.

5.2.2 Etape 2 de l'algorithme

Pour $R_s(x)$ donné (i.e. a, d, α, β donné) nous allons déterminer a_{nv} respectant 20 et l_{nv} s'approchant de 21.

i) Calcul de a_{nv}

Après la résolution du système 24, il est probable que la condition

$$\max_{x\in[-1,a-\epsilon]} |R_s(x)| = 1 \text{où } \epsilon > 0 \text{ est le plus petit réel tel que } R_s(x) \text{ ait un extremum local en } a-\epsilon \tag{25}$$

ne soit pas respectée. Pour déterminer a_{nv} respectant cette condition, on pourra procéder comme suit :

Soit $m = \max_{x\in[-1,a-\epsilon]} |R_s(x)|$
- Si $m < 1$, on prendra $a_{nv} < a$ tel que $R_s(a_{nv}) = m$
- Si $m > 1$, on prendra $a_{nv} > a$ tel que $R_s(a_{nv}) = m$

Nous pouvons affirmer l'existence d'une solution a_{nv} par le théorème des valeurs intermédiaires puisque, en notant γ_{s-2} la plus grande racine réelle, $R(x)$ augmente pour $x > \max(\alpha, \gamma_{s-2})$ et diminue si x tend vers γ_{s-2}

Nous redéfinissons alors $R_s(x)$ par

$$R_s(x) = \frac{\omega(x)P_{s-2}(x)}{\omega(a_{nv})P_{s-2}(a_{nv})} \tag{26}$$

ii) Calcul de l_{nv}

Nous allons chercher une nouvelle valeur pour l, et donc par la même occasion de $d = \frac{l}{1+a}$. Cette valeur est nécessairement inférieure à $\bar{l}_s$, valeur des polynômes optimaux que nous cherchons à approcher. Puisque les polynômes optimaux sont equi-oscillants avec pour valeur 1 en $s-1$ points, nous allons chercher à déterminer l_{nv} de façon à ce que les extrema locaux de $R_s(x)$ restent proche de 1.

Nous posons

$$\forall i \in \{1,\dots,s-2\}, \mu_i = \max_{x\in[\gamma_i,\gamma_i]} |R_s(x)|, \text{ et } \mu_{min} = \min_i \mu_i \tag{27}$$

$$l_{nv} = +\zeta(1-\mu_{min}) \tag{28}$$

où γ_i sont les racines réelles de $R_s(x)$, $\gamma_{s-1} = a-\epsilon$ est le plus proche extremum local de a et ζ une constante positive (Les auteurs de la méthode proposent $\zeta = 0.5$). Si $l > \bar{l}$ ou que $1-\mu_{min} < t$ où $t > 0$ est une constante, on conserve l (i.e. $l_{nv} = l$). Puisque $l < \bar{l}$, les auteurs de la méthode proposent de prendre pour valeur initiale de l une fraction de $\bar{l}$, comme $\frac{4}{5}\bar{l}$.

Résumé de l'algorithme :

1. Calcul de α et β par la résolution du système 24
2. Calcul de a_{nv} et l_{nv}
3. Recommencer la première étape, tant que $|a_{nv}-a_{pre}| < q$ et $|l_{nv}-l_{pre} < q$ où $q > 0$ est une constante proche de 0, déterminant à partir de quel instant les solutions sont considérés comme relativement stables.

Nous avions évoqué l'utilisation d'une constante positive η afin d'éviter les problèmes d'arrondi. En pratique, il faudrait changer certaines conditions par $\max_{x\in[-1,a-\epsilon]} |R_s(x)| = \eta$ et $l_{nv} = l + \zeta(\eta - \mu_{min})$.

On remarque, expérimentalement, que dans le cas des polynômes optimaux d'ordre 2, l'intervalle de stabilité maximal est donnée par $\bar{l} = \bar{c}_2(s)s^2$ et que $\bar{c}_2(s) \to 0.82$, très rapidement.

6 Existence de R_s

Notons

$$\forall s \geq 3, P_{s-2}(x) = \prod_{i=1}^{s-2}(x-\gamma_i), -1 < \gamma_1 < \cdots < \gamma_{s-2} < 1$$

nos polynômes orthogonaux. On posera aussi $\tilde{R}_s(x) = \omega(x)P_{s-2}(x)$.

Nous allons utiliser l'hypothèse suivante, vérifiée par tous les polynômes optimaux d'ordre pair :

$$\text{(H) } \exists \xi_1, \xi_2 \text{ avec } \gamma_{s-2} < \xi_1 < \xi_2 < \alpha \text{ tq. } \tilde{R}'_s(\xi_i) = 0$$

Commençons par démontrer qu'elle est vérifiée pour nos polynômes $\tilde{R}_s$.

Lemme 3.3 : Si l'on note γ_{s-2} la plus grande racine réelle de $\tilde{R}_s(x)$, alors $\exists \delta > 0$ tq. $\forall -\delta \leq 1-\alpha < \delta, 0 \leq \beta \leq \delta$, alors $\exists \gamma_{s-2} < \xi_1 < \xi_2 < \alpha$ tq.

$$\forall i \in 1, 2, \tilde{R}'_s(\xi_i) = 0$$

Démonstration : Posons

$$g(x) = \frac{\tilde{R}'_s(x)}{\prod_{i=1}^{s-2}(x-\gamma_i)} = 2(x-a) + ((x-a)^2 + b^2)\sum_{i=1}^{-2}\frac{1}{x-\gamma_i}$$

Nous pouvons supposer que $\alpha = 1$ et $\beta = 0$. Posons aussi $\epsilon = (1-\gamma_{s-2})/2 > 0$. Nous avons alors, concernant le signe de $g(x)$

- $g(1+\frac{\epsilon}{s-2}) > 0$: Il suffit de remplacer dans l'expression pour constater que les termes sont tous positifs
- $g(1-\frac{\epsilon}{s-2}) > 0$: Il suffit de constater que

$$\sum_{i=1}^{s-2}\frac{1}{1-\frac{e}{s-2}-\gamma_i} \leq \frac{s-2}{1-\frac{e}{s-2}-\gamma_{s-2}} \leq \frac{s-2}{\epsilon}$$

- $g(x) \to 0$ quand $x \to \xi_{s-2}$ car $\tilde{R}_s(1) > 0$ et $\tilde{R}_s(x)$ s'annule en γ_{s-2}.

Puisque $g(x)$ dépend de façon continue de α et β, ces trois résultats restent vrais dans un voisinage de 1. On en déduit immédiatement que la fonction s'annule donc en deux points, ξ_1 et ξ_2.

Nous allons maintenant chercher à démontrer, pour α et β respectant (H), qu'il existe a et d tel que notre système 19 admette une solution. De cette façon nous aurons démontré l'existence d'un tel polynôme.

Théorème 3.4 : Soit $\tilde{R}_s(x)$ respectant (H) avec $\beta \neq 0$. Posons

$$R_s(x) = \frac{\tilde{R}_s(x)}{\tilde{R}_s(a)} = \frac{\omega(x)P_{s-2}(x)}{\omega(x)P_{s-2}(a)}$$

Alors, $\exists a > \xi \text{et} d$ tel que $R_s(x)$ respecte les conditions

$$R'_s(a,\alpha,\beta) = d, R''_s(a,\alpha,\beta) = d^2 \tag{29}$$

Démonstration : La condition $\beta \neq 0$ nous permet d'affirmer que $\tilde{R}_s(x)$ a exactement $s-2$ racines réelles. De plus, puisque $\tilde{R}_s(x)$ s'annule $s-2$ fois, il existe donc au moins $s-3$ racines réelles pour $x < \xi_1$. (Application du théorème de Rolle) De plus, ξ_1 et ξ_2 sont racines de $\tilde{R}'_s(x)$ ce qui nous permet d'affirmer(du fait que $\deg \tilde{R}_s(x) = s-1$) que $\tilde{R}'_s(x)$ possède exactement $s-1$ racines réelles.

Ce résultat implique que $\tilde{R}''_s(x)$ s'annule entre ξ_1 et ξ_2. De plus, pour $x > \xi_2$, $\tilde{R}'(x) > 0$ car $\tilde{R}(x) > 0$. En effet, si $\tilde{R}'(x) < 0$, alors $\tilde{R}(x)$ décroît et serait donc négatif. Puisque le dernier zéro de $\tilde{R}'(x)$ est ξ_2, $\tilde{R}(x)$ continue de décroître et on conclut que $\tilde{R}(x) < 0$, ce qui est absurde. De même, puisque $\tilde{R}'(x) > 0$, nous en déduisons $\tilde{R}''(x) > 0$.

Posons

$$p(x) = (\tilde{R}'_s(x))^2 - \tilde{R}_s(x)\tilde{R}''_s(x)$$

Puisque $\tilde{R}''_s(\xi_2) > 0$, $\tilde{R}_s(\xi_2) > 0$ et $\tilde{R}'_s(\xi_2) = 0$, nous pouvons affirmer $p(\xi_2) < 0$.

Or :

$$\tilde{R}_s(x) = x^s + O(x^{s-1})$$
$$\tilde{R}'_s(x) = sx^{s-1} + O(x^{s-2})$$
$$\tilde{R}''_s(x) = s(s-1)x^{s-2} + O(x^{s-3})$$

Donc, il vient

$$p(x) = s^2x^{s-2} - s(s-1)x^{2s-2} + O(x^{2s-3})$$

Donc $\exists y > \xi_2 | p(y) > 0$. Ainsi, $\exists a > \xi_2 | p(a) = 0$. Or, en utilisant l'égalité $p(x) = 0$, on obtient

$$\left(\frac{\tilde{R}'_s(a)}{\tilde{R}_s(a)}\right)^2 = \frac{\tilde{R}''_s(a)}{\tilde{R}_s(a)}$$

Il suffit alors de poser $R_s(x) = \frac{\tilde{R}_s(x)}{\tilde{R}_s(a)}$ et $d = R'_s(a)$.

7 Constante d'erreur

Soit $R_s(x)$ obtenu par le Théorème 3.4. On pose alors

$$\hat{R}_s(z) = Rs(a + \frac{z}{d}) =_0 1 + z + \frac{z^2}{2!} + a_3z^3 + o(z^3) \tag{30}$$

On appellera *Constante d'erreur* la constante

$$C = \frac{1}{3!} - a_3 \tag{31}$$

qui est le coefficient du terme z^3 de degré 3 entre la solution exacte $\sum_n^{+\infty} \frac{z^n}{n!}$ et la solution approchée $\hat{R}_s(z)$. De plus, on remarque que

$$R'''_s(a) = a_3 3! d^3 = a_3 3! (R'_s(a))^3$$

ce qui nous donne une expression pour le coefficient a_3, qui de plus est positive (cf: dernier argument sur les signes des dérivées dans le Théorème 3.4). Nous allons déterminer un encadrement pour C(ce qui revient à en déterminer un pour a_3).

Lemme 3.5 (Encadrement de la constante d'erreur) : Soit Q(z) un polynôme d'ordre 2 (i.e. $Q(z) = 1 + z + \frac{z^2}{2!} + \dots$) dont le polynôme dérivé $Q'(z)$ n'a que des racines réelles. Alors, la constante d'erreur $C = \frac{1}{3!} - a_3$ appartient à $]0, \frac{1}{6}[$.

Démonstration : Nous avons

$$Q'(z) = \prod_{i=1}^{s-1}(1 + \lambda_i z) = 1 + z + 3a_3 z^2 + \dots$$

où les λ_i sont tous réels. De plus, en développant le produit, et par identification des coefficients, on obtient

$$s_1 = \sum_{i=1}^{s-1} \lambda_i = 1 \text{ et } s_2 = \sum i < j^{j=s-1} \lambda_i \lambda_j = 3a_3 \tag{32}$$

Enfin, on rappelle la propriété suivante

$$(\sum_{i=1}^{s-1} \lambda_i)^2 = \left(\sum_{i=1}^{s-1} \lambda_i^2\right) + \left(\sum_{i<j}^{j=1} 2\lambda_i \lambda_j\right)$$

qui se démontre par une récurrence triviale sur le nombre de termes.

On en déduit que $s_1^2 - 2s_2$ est une somme de carrés, et est donc positif. La racine étant distinctes, la somme est strictement positive. Il vient alors $1 - 2 * 3a_3 > 0 \Rightarrow \frac{1}{6} > a_3$.

Ainsi, $a_3 \in [0, \frac{1}{6}]$ et $C \in]0, \frac{1}{6}[$.

Théorème 3.6 : Soit $R_s(x)$ un polynôme issu du théorème 3.4 respectant (H). La constante d'erreur du polynôme $R_s(a + \frac{z}{d}$ est $C \in]0, \frac{1}{6}[$.

Démontration: Application du lemme précédent à $R_s(x)$.

7.1 Formule explicite pour les polynômes orthogonaux P_{s-2}

Nous avions parlé de la possibilité de calculer $P_n(x_0)$. Nous alons ici présenter la formule explicite. Il existe des théorèmes bien plus généraux, et la situation traitée ici correspond à un cas particulier d'un théorème de [Sze59].

Théorème 3.7 : Soit $s \geq 3 \in \mathbb{N}$ et soit $n \leq s-2 \in \mathbb{N}$. Soit P_n le polynôme orthogonal de degré n pour la fonction $\omega(x) = (x - \alpha - i\beta)(x - \alpha + i\beta) = (x - x_1)(x - x_2)$ construit aux chapitres précédents. On a alors l'égalité :

$$\omega(x)^2 P_n(x) = CQ(x) \tag{33}$$

avec

$$Q(x) = \begin{vmatrix} T_n(x_1) & T_n(x_2) & T'_n(x_1) & T'_n(x_2) & T_n(x) \\ T_{n+1}(x_1) & T_{n+1}(x_2) & T'_{n+1}(x_1) & T'_{n+1}(x_2) & T_{n+1}(x) \\ T_{n+2}(x_1) & T_{n+2}(x_2) & T'_{n+2}(x_1) & T'_{n+2}(x_2) & T_{n+2}(x) \\ T_{n+3}(x_1) & T_{n+3}(x_2) & T'_{n+3}(x_1) & T'_{n+3}(x_2) & T_{n+3}(x) \\ T_{n+4}(x_1) & T_{n+4}(x_2) & T'_{n+4}(x_1) & T'_{n+4}(x_2) & T_{n+4}(x) \end{vmatrix} \tag{34}$$

où $T_n(x)$ est le polynôme de Tchebychev de degré n.

Remarque : Il existe un résultat plus général, sur lequel est calquée cette démonstration, qui permet d'exprimer tout polynôme orthogonal "inconnu" sous une forme de déterminant "connu".

Demonstration : D'après les propriétés sur les déterminants, nous avons $Q(x_i) = 0$. De plus, on constate que dérivées Q revient à dériver la dernière colonne du déterminant, et finalement il vient $Q'(x_i) = 0$. On en déduit donc immédiatement que les x_i sont racines doubles, i.e. $\exists S_n(x) \in \mathbb{C}[x] | Q(x) = C(x-x_1)^2(x-x_2)^2 S_n(x)$. De plus, $\deg S_n \leq n$. Nous cherchons alors à prouver que $S_n(x)$ est le polynôme orthogonal de degré n pour le poids $\omega(x)^2/\sqrt{(1-x^2)}$.

Soit $p(x)$ un polynôme orthogonal pour le poids $\omega(x)^2/\sqrt{(1-x^2)}$ de degré $\deg p(x) < n$. Nous savons que $\exists b_i | Q(x) = \sum_{i=0}^{4} b_i T_{n+i}(x)$. Il vient alors

$$\int_{-1}^{1} p(x) S_n(x) \frac{\omega(x)^2}{\sqrt{1-x^2}} \mathrm{d}x = \int_{-1}^{1} p(x) \sum_{i=0}^{4} T_{n+i}(x) \frac{1}{\sqrt{1-x^2}} \mathrm{d}x =$$

Or, comme $p(x)$ peut s'écrire comme combinaison linéaire des polynômes de Tchebychev de degré strictement inférieur à n, on obtient une somme de produits scalaires

$$\sum_{i=0}^{4} \sum_{k=0}^{n} \alpha_k < T_k, T_{n+i} >= 0$$

et l'on en déduit donc que S_n est orthogonal pour le poids $\omega(x)^2/\sqrt{(1-x^2)}$.

Enfin, pour conclure, il reste à démontrer que $S_n(x)$ n'est pas le polynôme nul. Pour ce faire, il suffit de démontrer que le coefficient de $T_{n+3}(x)$ est non nul. Par l'absurde, supposons que

$$\begin{vmatrix} T_n(x_1) & T_n(x_2) & T'_n(x_1) & T'_n(x_2) \\ T_{n+1}(x_1) & T_{n+1}(x_2) & T'_{n+1}(x_1) & T'_{n+1}(x_2) \\ T_{n+2}(x_1) & T_{n+2}(x_2) & T'_{n+2}(x_1) & T'_{n+2}(x_2) \\ T_{n+3}(x_1) & T_{n+3}(x_2) & T'_{n+3}(x_1) & T'_{n+3}(x_2) \end{vmatrix} = 0$$

Alors, de par les propriétés sur la non-liberté des colonnes, $\exists b_i$, non tous nuls, tel que

$$L(x) = b_0 T_n(x) + b_1 T_{n+1}(x) + b_2 T_{n+2}(x) + b_3 T_{n+3}(x) + b_4 T_{n+4}(x)$$

ait x_1 et x_2 comme racines doubles. Ceci implique que $L(x) = \omega(x)^2 G(x)$ où $G(x)$ est de degré $\deg G(x) \leq n-1$. Puisque $G(x)$ peut s'écrire comme une combinaison linéaire de polynômes $T_q(x)$ de degré $q \leq n-1$, il vient

$$\int_{-1}^{1} \omega(x)^2 G(x) G(x) \frac{1}{\sqrt{1-x^2}} \mathrm{d}x = \int_{-1}^{1} L(x) G(x) \frac{1}{\sqrt{1-x^2}} \mathrm{d}x$$

puis

$$\int_{-1}^{1} L(x) G(x) \frac{1}{\sqrt{1-x^2}} \mathrm{d}x = \int_{-1}^{1} \left(\sum_{i=0}^{4} b_i T_{n+i}(x) \right) \left(\sum_{q=0}^{n-1} T_q(x) \right)$$

Ce qui implique que $G(x) = 0$, et contredit notre hypothèse sur les b_i.

Utilisation de ce théorème :
Il est possible d'utiliser ce théorème afin de déterminer les coefficients de la relation de récurrence entre nos polynômes orthogonaux. Notons $\hat{P}_n(z) = P_n(a+\frac{z}{d})$ pour travailler sur $[-l, 0]$. Nommons les coefficients de récurrence par :

$$\hat{P}_n(z) = (\mu_n z - \nu_n)\hat{P}_{n-1}(z) - \kappa_n \hat{P}_{n-2}(z) \tag{35}$$

En utilisant trois valeurs distinctes r_1, r_2 et r_3, respectant la condition

$$\begin{vmatrix} r_1\hat{P}_{n-1}(r_1) & \hat{P}_{n-1}(r_1) & \hat{P}_{n-2}(r_1) \\ r_1\hat{P}_{n-1}(r_2) & \hat{P}_{n-1}(r_2) & \hat{P}_{n-2}(r_2) \\ r_1\hat{P}_{n-1}(r_3) & \hat{P}_{n-1}(r_3) & \hat{P}_{n-2}(r_3) \end{vmatrix} \neq 0 \tag{36}$$

nous sommes ramenés à résoudre le système

$$\forall i \in \{1,2,3\}\ \mu_n r_i \hat{P}_{n-1}(r_i) - \nu_n \hat{P}_{n-1}(r_i) - \kappa_n \hat{P}_{n-2}(r_i) = \hat{P}_n(r_i) \tag{37}$$

dont la matrice est inversible.

7.2 Construction de la méthode ROCK2

La détermination des coefficients de la relation de récurrence des polynômes orthogonaux va nous permettre d'obtenir facilement, en utilisant cette relation, les différentes étapes de calcul de l'algorithme de résolution, qui possèdera donc un domaine de stabilité quasi-optimal.

Le polynôme de stabilité de notre méthode doit être exactement $\hat{R}_s(z) = \hat{\omega}(z)\hat{P}_s(z)$, et comme les conditions pour que la méthode soit d'ordre deux se résument à celles imposées par notre choix de la fonction de stabilité, nous pouvons nous contenter de calculer ce polynôme, de la façon dont nous le désirons.

La méthode sera alors définie, pour ses premières étapes, par

$$\begin{array}{rcl} g_0 & := & y_0 \\ g_1 & := & y_0 + \mu_1 h f(g_0) \\ g_i & := & \mu_i h f(g_{i-1}) - \nu_i g_{i-1} - \kappa_i g_{i-1} \end{array} \tag{38}$$

ce qui, si l'on pose $f(x) = \lambda x$ nous donne, après calcul, $g_i = \hat{P}_i(h\lambda)g_0$.

Il nous reste alors à introduire trois étapes de finition, pour introduire le polynôme $\omega(x) = 1 + 2\sigma + \tau z^2$ par

$$\begin{array}{rcl} g_{s-1} & := & g_{s-2} + \sigma h f(g_{s-2}) \\ g_s^* & := & g_{s-1} + \sigma h f(g_{s-1}) \\ g_s & := & g_s^* - \sigma(1 - \frac{\tau^2}{\sigma^2})h(f(g_{s-1}) - f(g_{s-2})) \end{array} \tag{39}$$

À nouveau, si l'on pose $f(x) = \lambda x$ il vient $g_s = \hat{R}_s(z)g_0$. L'intérêt d'insérer l'étape intermédiaire g_s^* est de fournir une méthode embarquée, c'est-à-dire une méthode d'ordre inférieur qui permet ainsi d'estimer l'erreur commise par la méthode, et d'ainsi de réajuster le pas de temps h entre deux itérations.

8 ROCK 4

La méthode ROCK4 est, à l'exception de l'étape de construction de la méthode, calquée sur ROCK2, et il n'y a donc pas de réel intérêt à répéter ce qui a déjà été dit. Notons tout de même les quelques détails qui différent.

Dans un premier temps, il est nécessaire de déterminer les coefficients des quatres racines complexe du polynôme $\omega_4(x) = (x-(\alpha_1 - i\beta_1))(x-(\alpha_1 + i\beta_1))(x-(\alpha_2 - i\beta_2))(x-(\alpha_1 + i\beta_2))$. Cela se fait par un algorithme quasi-identique à celui de ROCK2, à ceci près que l'on dispose maintenant des quatres équations :

$$\begin{array}{c} R'_s(a,\alpha_1,\beta_1,\alpha_2,\beta_2) = d \\ R''_s(a,\alpha_1,\beta_1,\alpha_2,\beta_2) = d^2 \\ R'''_s(a,\alpha_1,\beta_1,\alpha_2,\beta_2) = d^3 \\ R^{(4)}_s(a,\alpha_1,\beta_1,\alpha_2,\beta_2) = d^4 \end{array} \tag{40}$$

Sur le même modèle que ROCK2, une suite de polynômes orthogonaux quasi-optimaux pour le poids $\omega_4(z)^2/\sqrt{1-z^2}$ peuvent être construits. L'on déterminera les coefficients de la relation de récurrence par la même méthode que celle employée dans ROCK2, issue du cas général traité dans [Sze59].

8.1 Spécificités de la construction

La construction de la méthode ROCK4 est légèrement différente, puisqu'il ne suffit plus de respecter les deux mêmes conditions que celles imposées par le choix de la fonction de stabilité. Dans un premier temps, il est nécessaire de déterminer les coefficients du tableau de Butcher en fonction des coefficients de la relation de récurrence. Cela se fait facilement en remarquant que, si l'on note $k_i = f(y_0 + h\sum_{j=1}^{i-1} \tilde{a}_{ij}k_j)$, nous avons[12] :

$$g_i = y_0 + \sum_{j=1}^{i} \tilde{a}_{i+1,j} h\kappa_j$$

Une fois ces coefficients déterminés, en utilisant les théorèmes relatifs aux groupes de Butcher et à la composition de méthodes, et en cherchant à déterminer les coefficients d'une méthode à trois étages tel que la composition de la première et de celle ci soit d'ordre quatre, on obtient un système d'équation à deux degrés de liberté. En s'imposant deux conditions supplémentaires, et en résolvant le système, on parvient à déterminer les coefficients des trois derniers étages de la méthode.

Deuxième partie
Étude des algèbres pré-Lie

Dans la suite de ce document, nous considérerons le problème :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x)\ x(0) = x_0$$

[12] Il peut être aussi utile de remarquer que $P_j(0) = 1$ et donc que $(-\nu_j - \kappa_j) = 1$

dans des conditions où la série de Taylor de $x(t)$ converge.

9 Champs vectoriels et équations différentielles

Nous pouvons écrire la solution de cette équation par sa série de Taylor, qui est de la forme :

$$x(t) = x_0 + c_1(x_0)t + c_2(x_0)\frac{t^2}{2} + c_3(x_0)\frac{t^3}{6} + \cdots + c_i(x_0)\frac{t^i}{i!} + \ldots$$

où nous savons, d'après 4.1 que

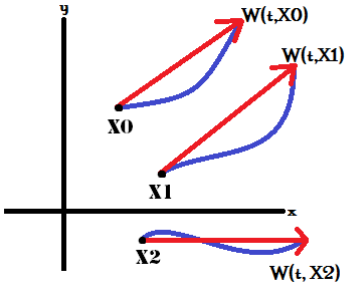
$$c_i(x_0) = \sum_{|\tau|=p, \tau\in\mathbb{T}} \alpha(\tau)F(\tau)(x_0)$$

Nous pouvons alors construire la fonction de $\mathbb{R}^p \to \mathbb{R}^p$ suivante :

$$W_t(X) = x(t) - X = \sum_{n=1}^{+\infty} c_n(X)\frac{t^n}{n!}$$

C'est une fonction qui associe à un instant donné et à une condition initiale X donnée le vecteur qui relie le point X au point de la solution à l'instant t.

FIG. 3 – Quelques valeurs de W_t



Pour poursuivre, introduisons quelques définitions relatives aux champs de vecteurs.

Définition 4.1 (Champ de vecteurs réel) : Un champ de vecteurs réel[13] est une application $\mathbb{R}^p \to \mathbb{R}^q$. Dans le cadre de ce document, nous ne considérerons que des champs de vecteurs infiniment différentiables, ceci afin de garantir l'existence des dérivées partielles, et de leur différentielle $n^{\text{ième}}$. *Remarque :* Dans le cas où $q = 1$ on parle de champ scalaire.

[13] On peut, bien entendu, définir des champs de vecteurs sur un autre corps commutatif.

Définition 4.2 (Opérateur ∂_{x_i}) : Soit une application linéaire f de $\mathbb{R}^p \to \mathbb{R}^q$ définie par $(x_1, \ldots, x_p) \mapsto f(x_1, \ldots, x_p)$, alors l'opérateur linéaire ∂_{x_i} est l'application

$$\begin{array}{rccc} \partial_{x_i} & : & \mathcal{L}(\mathbb{R}^p, \mathbb{R}^q) & \to & \mathcal{L}(\mathbb{R}^p, \mathbb{R}^q) \\ & & f & \mapsto & \frac{\partial f}{\partial x_i} \end{array}$$

où

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_p) - f(x_1, \ldots, x_i, \ldots, x_p)}{h}$$

Il respecte la règle de Leibniz[14].

Lemme 4.3 (Identification des champs de vecteurs et des opérateurs respectant la règle de Leibniz) : Il existe une bijection entre les champs de vecteurs et les opérateurs linéaires respectant la règle de Leibniz.

Démonstration : Soit $f(x_1, \ldots, x_p) = \sum_{i=1}^p f_i(x_1, \ldots, x_p)\vec{x_i}$ un champ de vecteurs. On peut alors lui associer l'opérateur $\mathcal{F} = \sum_{i=1}^p f_i(x_1, \ldots, x_p)\partial_{x_i}$, qui vérifie la règle de Leibniz.

À l'inverse, soit un opérateur linéaire $Q(x)$ qui vérifie la règle de Leibniz. Sa valeur est déterminée par ses valeurs sur les monômes x_i^k, et en appliquant la règle de Leibniz, on se ramène à une expression de la forme $kx_i^{k-1}Q(x) = Q(x)\partial_{x_i}x^k = Q(x_i^k)$.

Ainsi, Q est déterminé par ses valeurs sur chacune des variables x_i, qui sont de la forme $Q(x_i) = \alpha_i(x_1, \ldots, x_p)$. On en conclut que $Q = \sum_{i=1}^p \alpha_i(x_1, \ldots, x_p)\partial_{x_i}$.

Donc, tout opérateur a pour antécédent le champ de vecteurs $\sum_{i=1}^p \alpha_i(x_1, \ldots, x_p)\vec{x_i}$ et, si deux champs de vecteurs ont même image, nous avons l'égalité des fonctions $\alpha_i(x_1, \ldots, x_p)$.

Remarque : Cela revient à identifier ∂_{x_i} à $\vec{x_i}$.

Notons que notre fonction $W_t(X)$, à t fixé, est un champ de vecteurs. De plus, chacun des coefficients $x_i(X_0)$ définit aussi un champ de vecteurs. Nous pouvons même expliciter le premier de ces coefficients $c_1(X) = f(X) = \begin{pmatrix} f_1(X) \\ \vdots \\ f_p(X) \end{pmatrix}$ ou bien sous la forme de l'opérateur $c_1 = \sum_{i=1}^p f_i\partial_{x_i}$.

Nous allons alors définir un nouvel opérateur, qui va nous permettre d'exprimer facilement les coefficients c_i.

Définition 4.4 (Opérateur de greffe vectorielle) : On définit l'opérateur $\triangleleft : \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p) x \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p) \to \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$ de greffe pour les champs de vecteurs par

$$f(x) \triangleleft g(x) = g(x)f'(x)$$

Lemme 4.5 : Si l'on note G et F les opérateurs de dérivation partielle associés aux champs de vecteurs f et g, alors $f \triangleleft g = G \circ F$[15].

[14] Nous l'avions déjà rappelé ; Aussi appelée règle du produit, il s'agit de $(fg)' = f'g + fg'$, ou encore en notant δ l'opérateur, $\delta(fg) = \delta f.g + f.\delta g$

[15] On entend par ici que l'on exprime $f \triangleleft g$ sous la forme d'un opérateur de dérivation partielle. Nous nous permettrons de changer "spontanément" de notation, quand cela sera nécessaire.

Démonstration : Dans un premier temps, considérons le cas d'un champ scalaire unidimensionel (i.e. une fonction d'une seule variable de $\mathbb{R}$ dans $\mathbb{R}$). Nous avons donc $G(x) = g(x)\partial_x$ et $F(x) = f(x)\partial_x$. En explicitant le calcul, on a immédiatement l'égalité recherchée :

$$G \circ F = g(x)\partial_x f(x)\partial_x = g(x)\frac{\partial f(x)}{\partial x}\partial_x = g(x)f'(x)\partial_x$$

Observons le cas où $f : (x, y) \mapsto (\alpha_f(x, y), \beta_f(x, y))$ et $g : (x, y) \mapsto (\alpha_g(x, y), \beta_g(x, y))$ sont des champs vectoriels de $\mathbb{R}^2$ dans $\mathbb{R}^2$. Avec la même notation, nous avons :

$$G \circ F = (\alpha_g\partial_x + \beta_g\partial_y) \circ (\alpha_f\partial_x + \beta_f\partial_y) = \alpha_g\frac{\partial\alpha_f}{\partial x}\partial_x + \alpha_g\frac{\partial\beta_f}{\partial x}\partial_y + \beta_g\frac{\partial\alpha_f}{\partial y}\partial_x + \beta_g\frac{\partial\beta_f}{\partial y}\partial_y$$

et finalement

$$G \circ F = J_f \begin{pmatrix}\alpha_g \\ \beta_g\end{pmatrix}$$

où

$$J_f = \begin{pmatrix}\frac{\partial\alpha_f}{\partial x} & \frac{\partial\alpha_f}{\partial y} \\ \frac{\partial\beta_f}{\partial x} & \frac{\partial\beta_f}{\partial y}\end{pmatrix}$$

est le jacobien de f, ausi noté f'.

Le cas des fonctions $\mathbb{R}^p \to \mathbb{R}^p$ où $p > 2$ est identique au cas précédent, à ceci près que les notations sont moins élégantes. Par la suite, on emploira l'opérateur $\partial_x = \begin{pmatrix}\partial_{x_1} \\ \vdots \\ \partial_{x_p}\end{pmatrix}$ afin de pouvoir écrire plus simplement nos opérateurs de dérivées partielles. De cette façon, on retrouve la notation $f(x)\partial_x = (f_1, \ldots, f_p)\begin{pmatrix}\partial_{x_1} \\ \vdots \\ \partial_{x_p}\end{pmatrix}$ des fonctions scalaires à une variable avec des fonctions de $\mathbb{R}^p \to \mathbb{R}^p$.

Nous allons enfin pouvoir exprimer les coefficients c_i de façon élégante. D'après la définition d'une série de Taylor, nous avons la relation de récurrence

$$c_{n+1}(x(t))|_{x=X} = \left.\frac{\mathrm{d}}{\mathrm{d}t}c_n(x(t))\right|_{x=X}$$

Simplifions alors l'expression, en utilisant l'équation différentielle

$$c_{n+1}(x) = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}c_n}{\mathrm{d}x}(x) = f(x)\frac{\mathrm{d}c_n}{\mathrm{d}x}(x) = (c_n \triangleleft f)(x)$$

Armons-nous d'une dernière définition, avant d'énoncer un premier théorème.

Définition 4.6 : On notera $f^{\triangleleft n}$ l'application de l'opérateur $\triangleleft$ à f répété n fois, en parenthésant à gauche.

$$f^{\triangleleft n} = (((f \triangleleft f) \triangleleft f) \triangleleft f \ldots) \triangleleft f \text{où } f \text{ apparaît } n \text{ fois}$$

Vient alors notre premier résultat remarquable :

Théorème 4.7 :

$$W_t(X) = x(t) - X = \sum_{n\geq 1} f^{\triangleleft n}\frac{t^n}{n!}$$

Démonstration : Après les résultats que nous avons établis, la démonstration revient à une récurrence triviale sur n.

10 Morphisme de l'ensemble des arbres vers l'ensemble des champs de vecteurs

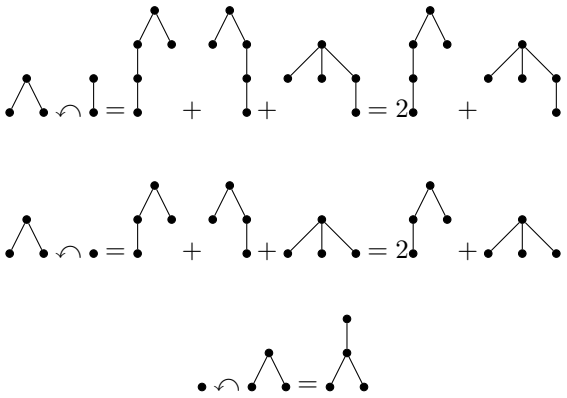
Nous pouvons chercher à établir un premier résultat qui lie les arbres aux champs de vecteurs, et à l'opérateur de greffe $\triangleleft$ que nous avons définis. Pour ce faire, nous avons besoin d'un opérateur similaire à $\triangleleft$ qui prenne ses valeurs dans l'ensemble des arbres $\mathbb{T}$.

Définition 4.8 (Opérateur de greffe sur les arbres) : On notera $\curvearrowleft$ l'opérateur, définit sur le $\mathbb{R}$-espace vectoriel des arbres, associant à un couple d'arbres $(\tau_1, \tau_2) \in \mathbb{T} \times \mathbb{T}$ la somme (formelle) :

$$\tau_1 \curvearrowleft \tau_2 = \sum_{s\in \text{Sommets}(\tau_1)} \tau_1 \circ_s \tau_2$$

où $\circ_s$ correspond à l'ajout de τ_2 comme sous-arbre de τ_1 fixé sur le sommet s.

Exemple : Il est beaucoup plus facile de se représenter cet opérateur en observant son comportement sur quelques arbres.



Théorème 4.9 : L'application linéaire Θ du $\mathbb{R}$-espace vectoriel $\mathbb{T}$ vers le $\mathbb{R}$-espace vectoriel $\text{Vec}(f(x)\partial_x)$ (L'espace vectoriel engendré par $f(x)\partial_x$ et la loi $\triangleleft$) qui à un arbre associe son expression en "produit des dérivées de f"[16], que nous avions appelé différentielle élémentaire, et défini au chapitre 4.1, est un morphisme pour les opérateurs de greffe de ces deux ensembles.

Démonstration : Soit deux arbres $(\tau_1, \tau_2) \in \mathbb{T}^2$ Appliquons Θ à $\tau_1 \curvearrowleft \tau_2$:

$$\Theta(\tau_1 \curvearrowleft \tau_2) = \Theta(\sum_{s\in \text{Sommets}(\tau 1)} \tau_1 \circ_s \tau_2)$$

[16] Ce n'est un produit de scalaire qu'en dimension 1. Pour des fonctions de $\mathbb{R}^p \to \mathbb{R}^p$, ce sont des applications n linéaires où n est le degré de différentiation.

par définition de Θ :

$$\Theta(\sum_{s\in\text{Sommets}(\tau 1)} \tau_1 \circ_s \tau_2) = \left[\sum_{s\in\text{Sommets}(\tau_1)} \Theta(\tau_1, s)\right] \Theta(\tau_2)$$

où $\Theta(\tau, s)$ représente l'expression $\Theta(\tau)$ où l'on a dérivé une fois de plus f pour le sommet s. Il vient alors

$$\left[\sum_{s\in\text{Sommets}(\tau_1)} \Theta(\tau_1, s)\right] \Theta(\tau_2) = \left[\sum_{s\in\text{Sommets}(\tau_1)} f^{(\text{valence de } s+1)} \prod_{j\neq s} f^{(\text{valence de } j)}\right] \Theta(\tau_2)$$

et finalement

$$\Theta(\tau_1 \curvearrowleft \tau_2) = \left(\prod_{s\in\text{Sommets}(\tau_1)} f^{(\text{valence de } s)}\right)' \Theta(\tau_2) = \Theta(\tau_1)'\Theta(\tau_2) = \Theta(\tau_1) \triangleleft \Theta(\tau_2)$$

11 Les algèbres pré-Lie

Pour mieux comprendre l'origine de cette similitude entre ces deux opérateurs, nous allons étudier une importante propriété de ces derniers. À défaut d'associativité, ou de commutativité, nous pouvons observer la différence entre deux parenthésages d'une expression à trois opérandes.

$$(a \triangleleft b) \triangleleft c - a \triangleleft (b \triangleleft c) = (ba') \triangleleft c - a \triangleleft (cb') = c(ba')' - (cb')a' = a''bc + a'b'c - a'b'c = a''bc$$

Nous remarquons alors que cette dernière expression est symétrique en b et en c, c'est-à-dire

$$(a \triangleleft b) \triangleleft c - a \triangleleft (b \triangleleft c) = (a \triangleleft c) \triangleleft b - a \triangleleft (c \triangleleft b)$$

Définition 4.10 (Algèbre pré-Lie) : Une algèbre pré-Lie sur un corps $\mathbb{K}$ est un $\mathbb{K}$-espace vectoriel E muni d'un opérateur bilinéaire $\triangleleft : E \times E \to E$ tel que

$$\forall (a, b, c) \in E^3,\ (a \triangleleft b) \triangleleft c - a \triangleleft (b \triangleleft c) = (a \triangleleft c) \triangleleft b - a \triangleleft (c \triangleleft b) \tag{41}$$

Remarque : Les champs de vecteurs munis de la loi de greffe sont donc une algèbre pré-Lie. On dira qu'une algèbre de pré-Lie est libre si la seule relation qu'elle respecte est celle-ci. On notera $\mathcal{PL}[u]$ l'algèbre pré-Lie engendrée par le singleton u et les expressions, où l'opérateur $\triangleleft : \mathcal{PL}[u] \times \mathcal{PL}[u] \to \mathcal{PL}[u]$ respecte la relation précédente, et aucune autre[17].

Comme tout lecteur peut s'y attendre, l'ensemble des arbres et de son opérateur de greffe va, de façon inattendue, lui aussi respecter cette relation.

[17] Bien que noté de la même façon que l'opérateur de greffe sur les champs de vecteurs, il ne représente pas exactement la même chose. On ne s'intéresse ici qu'à l'aspect formel des expressions. Il y a bien évidement une surjection de $\mathcal{PL}[u] \to \text{Vec}(f(x)\partial_x)$ mais rien ne garantit l'injection. On peut s'en convaincre immédiatement en observant $\mathcal{PL}[u] \to \text{Vec}(e^x\partial_x)$.

Théorème 4.11 (Les arbres forment une algèbre pré-Lie) : Le $\mathbb{R}$-espace vectoriel sur les arbres $\mathbb{T}$, muni de la loi $\curvearrowleft$ est une algèbre pré-lie.

Démonstration : Observons la différence

$$(a \curvearrowleft b) \curvearrowleft c - a \curvearrowleft (b \curvearrowleft c)$$

comme nous avons fait avec les champs de vecteurs. Observons la première expression à trois termes. On peut distinguer deux cas : le premier où b et c se greffent sur deux sommets différents de a, et le second où c se greffe sur un des sommets de b. Nous pouvons alors réécrire cette relation sous la forme

$$\left(a \curvearrowleft (b \curvearrowleft c) + \sum_{i \neq j} [\dots, \tau_i \curvearrowleft b, \dots, \tau_j \curvearrowleft c, \dots]_a \right) - a \curvearrowleft (b \curvearrowleft c)$$

où $a = [\tau_1, \dots, \tau_k]_a$. Finalement, la différence se simplifie et nous avons une expression symétrique en b et c

$$\sum_{i \neq j} [\dots, \tau_i \curvearrowleft b, \dots, \tau_j \curvearrowleft c, \dots]_a = \sum_{i \neq j} [\dots, \tau_i \curvearrowleft c, \dots, \tau_j \curvearrowleft b, \dots]_a$$

12 Arbres et algèbre pré-Lie

Nous en savons maintenant un peu plus sur les arbres et leur lien avec les différentielles élémentaires. De ce théorème et de sa démontration, nous pouvons tirer les affirmations précédentes comme des conséquences.

Théorème 4.12 (Morphisme) : Soit $(V, \triangleleft)$ une algèbre pré-Lie et v un élément de V. Alors, $\exists!$ morphisme d'algèbre pré-Lie

$$\begin{array}{ccc} (\mathbb{T}, \curvearrowleft) & \to & (V, \triangleleft) \\ \bullet & \mapsto & v \end{array}$$

Pour démontrer ce théorème, nous allons dans un premier temps chercher à construire un isomorphisme entre les algèbres de pré-Lie libre et l'espace vectoriel sur les arbres.

12.1 Isomorphisme d'une algèbre pré-Lie libre et d'un espace vectoriel sur les arbres

Nous avons précédemment défini $\mathcal{PL}[u]$ comme "l'algèbre engendrée par l'opérateur $\triangleleft$ respectant uniquement l'expression 41", mais une façon plus rigoureuse serait de parler de "l'ensemble des expressions engendrées par le singleton u et $\triangleleft$ quotienté par la relation 41".

Nous définissons alors l'application linéaire $\phi(\tau) : \mathbb{T} \to \mathcal{PL}[u]$ par récurrence double sur la valence et le nombre de sommets. En utilisant les notations précédemment introduites, nous avons :

$$\phi(\tau) = \phi([\tau_2, \dots, \tau_k]_\tau) \triangleleft \phi(\tau_1) - \sum_{j=2}^{k} \phi([\tau_2, \dots, \tau_j \triangleleft \tau_1, \dots, \tau_k]_\tau)$$

Il est question de "découper" l'arbre en sous-arbre, afin de réduire la complexité de ce dernier. À chaque application, la valence des deux termes décroît, et pour une valence de 1, le second

terme est nul. On peut appliquer alors cette définition autant de fois que nécessaire pour que n'apparaisse plus qu'une expression en $\phi(\bullet)$.

Dans un premier temps, il est nécessaire de vérifier que cette application est bien définie ; le choix du sous-arbre à découper est purement subjectif et nous devons vérifier que cela n'influence pas le résultat obtenu. Nous allons donc appliquer une seconde découpe, et vérifier que l'expression obtenue soit symétrique en τ_1 et τ_2. Le succès, réside bien sûr dans la relation 41.

$$\begin{aligned}\phi(\tau) = \quad & \Big[\phi([\tau_3,\dots,\tau_k]) \triangleleft \phi(\tau_2) - \textstyle\sum_{l=3}^{k} \phi([\tau_3,\dots,\tau_l \curvearrowleft \tau_2,\dots,\tau_k])\Big] \triangleleft \phi(\tau_1) \\ & - \Big(\textstyle\sum_{j=3}^{k} \phi([\tau_3,\dots,\tau_j \curvearrowleft \tau_1,\dots,\tau_k]) \triangleleft \phi(\tau_2)\Big) \\ & + \Big(\textstyle\sum_{j\neq l, j\neq 2} \phi(\tau_3,\dots,\tau_j \curvearrowleft \tau_1,\dots,\tau_l \curvearrowleft \tau_2,\dots,\tau_k)\Big) \\ & + \Big(\textstyle\sum_{j\neq 2} \phi(\tau_3,\dots,(\tau_j \curvearrowleft \tau_1) \curvearrowleft \tau_2,\dots,\tau_k)\Big) \\ & - (\phi([\tau_3,\dots,\tau_k) \triangleleft \phi(\tau_2 \curvearrowleft \tau_1)) \\ & + \Big(\textstyle\sum_{j=3}^{k} \phi(\tau_3,\dots,\tau_j \curvearrowleft (\tau_2 \curvearrowleft \tau_1),\dots,\tau_k)\Big)\end{aligned}$$

Chacun de ces termes mérite une explication, car leur obtention n'est pas triviale.

Le premier correspond à la découpe de $\phi([\tau_2,\dots,\tau_k]_\tau) \triangleleft \phi(\tau_1)$ sur τ_2 ; on remplace simplement ϕ par son expression. Pour la découpe de $\sum_{j=2}^{k} \phi([\tau_2,\dots,\tau_j \triangleleft \tau_1,\dots,\tau_k]_\tau)$ ce n'est pas aussi simple. Il nous faudra distinguer différents cas selon la valeur de j.

Le second terme correspond à la première partie de la découpe, c'est-à-dire $\phi([\tau_2,\dots,\tau_k]_\tau) \triangleleft \phi(\tau_1)$, d'un terme de la somme où $j \neq 2$.

Le troisième terme correspond à la deuxième partie de cette découpe (toujours pour $j \neq 2$) dans le cas où $l \neq j$. C'est donc une somme imbriquée, de la forme $\sum_j \sum_l$. On a exclu le cas où τ_1 vient se greffer au même emplacement que l. Il nous manque donc un dernier élément de cette découpe.

Le quatrième terme est la dernière partie de la découpe, correspondant à la somme des greffes de τ_2 sur la même position de τ_1. C'est une somme simple car une fois τ_1 placé, la position de τ_2 est déterminée.

Enfin, le cinquième et le sixième termes correspondent à la découpe dans le cas que nous avions exclu, c'est-à-dire où τ_1 est venu se greffer sur τ_2. C'est donc l'ensemble $\tau_1 \curvearrowleft \tau_2$ que l'on découpe et que l'on vient greffer.

Nous pourrons simplifier l'expression car deux termes vont s'annuler. Réordonnons l'expression afin de mettre en évidence la symétrie.

$$\begin{aligned}\phi(\tau) = \quad & ((\phi([\tau_3,\dots,\tau_k]) \triangleleft \phi(\tau_2)) \triangleleft \phi(\tau_1)) \\ & - (\phi([\tau_3,\dots,\tau_k) \triangleleft \phi(\tau_2 \curvearrowleft \tau_1)) \\ & \\ & - \Big(\textstyle\sum_{l=3}^{k} \phi([\tau_3,\dots,\tau_l \curvearrowleft \tau_2,\dots,\tau_k]) \triangleleft \phi(\tau_1)\Big) \\ & - \Big(\textstyle\sum_{j=3}^{k} \phi([\tau_3,\dots,\tau_j \curvearrowleft \tau_1,\dots,\tau_k]) \triangleleft \phi(\tau_2)\Big) \\ & \\ & + \Big(\textstyle\sum_{j\neq l, j\neq 2} \phi(\tau_3,\dots,\tau_j \curvearrowleft \tau_1,\dots,\tau_l \curvearrowleft \tau_2,\dots,\tau_k)\Big) \\ & \\ & + \Big(\textstyle\sum_{j=3}^{k} \phi(\tau_3,\dots,(\tau_j \curvearrowleft \tau_1) \curvearrowleft \tau_2,\dots,\tau_k)\Big) \\ & + \Big(\textstyle\sum_{j=3}^{k} \phi(\tau_3,\dots,\tau_j \curvearrowleft (\tau_2 \curvearrowleft \tau_1),\dots,\tau_k)\Big)\end{aligned} \tag{42}$$

Le cinquième terme est bien évidemment symétrique en τ_1 et τ_2. La somme du troisième et du quatrième est elle aussi symétrique. En appliquant la relation 41 on constate que les deux dernières expressions sont symétriques.

Pour conclure, nous allons avoir besoin d'un petit lemme, sur l'application ϕ.

Lemme 4.13 : ϕ respecte l'égalitée $\phi(\tau \curvearrowleft \tau') = \phi(\tau) \triangleleft \phi(\tau')$

Démonstration : La démonstration se fait par récurrence sur la valence. Pour une expression à une valence de 1, le second terme de ϕ est nul, comme nous l'avons déjà dit, et donc seul le premier terme subsiste, et l'égalité est vérifiée quelque soit le nombre de sommets. Si l'on suppose l'égalité vraie à valence n, d'après la définition de $\curvearrowleft$ il vient

$$\phi(\tau \curvearrowleft \tau') = \phi([\tau', \tau_1, \ldots, \tau_k]) + \sum_{j=1}^{k} \phi([\tau_1, \ldots, \tau_j \curvearrowleft \tau', \ldots, \tau_k])$$

puis, en appliquant la définition de ϕ à la première expression :

$$\phi(\tau \curvearrowleft \tau') = \phi(\tau) \triangleleft \phi(\tau') - \sum_{j=1}^{k} \phi([\tau_1, \ldots, \tau_j \curvearrowleft \tau', \ldots, \tau_k]) + \sum_{j=1}^{k} \phi([\tau_1, \ldots, \tau_j \curvearrowleft \tau', \ldots, \tau_k])$$

d'où

$$\phi(\tau \curvearrowleft \tau') = \phi(\tau) \triangleleft \phi(\tau')$$

En appliquant ceci à la deuxième expression de l'équation 42, ainsi que la relation 41, nous pouvons conclure que la somme des deux premières expressions est nulle, et ainsi que notre application est bien définie.

Nous pouvons maintenant définir l'application $\psi : \mathcal{PL}[u] \to \mathbb{T}$ qui associe à une expression, la même expression formée de l'opérateur $\curvearrowleft$. Cette application est évidement bien définie puisque deux éléments de la même classe d'équivalence sont liés par la relation 41, et que les images de ces éléments sont aussi liés par cette relation. Elle est, par définition, un morphisme. Nous allons montrer que cette dernière est la bijection réciproque de ϕ.

Nous pouvons démontrer simplement, par récurrence sur la valence, que $\psi \circ \phi = id_{\mathbb{T}}$. C'est évidemment vrai pour une valence de 1, en appliquant la définition de ϕ puis de ψ. De plus, en subdivisant un arbre $[\tau_1, \ldots, \tau_k]$ sous la forme $[\tau_2, \ldots, \tau_k] \curvearrowleft \tau_1 - \sum_{j=2}^{k} -[\tau_2, \ldots, \tau_i \curvearrowleft \tau_1, \ldots, \tau_k]$ et en appliquant le lemme 4.13, ainsi que l'hypothèse de récurrence, nous avons immédiatement le résultat escompté :

$$\psi \circ \phi(\tau) = \psi \circ \phi([\tau_2, \ldots, \tau_k] \curvearrowleft \tau_1 - \sum_{j=2}^{k} [\tau_2, \ldots, \tau_i \curvearrowleft \tau_1, \ldots, \tau_k]) \tag{43a}$$

$$= \psi \circ \phi([\tau_2, \ldots, \tau_k]) \triangleleft \psi \circ \phi(\tau_1) - \sum_{j=2}^{k} \psi \circ \phi([\tau_2, \ldots, \tau_i \curvearrowleft \tau_1, \ldots, \tau_k]) \tag{43b}$$

$$= \tau \tag{43c}$$

Enfin, par récurrence sur le nombre de termes dans un expression en $\triangleleft$, que l'on pourra toujours exprimer sous la forme $(\dots)_x \triangleleft (\dots)_y$, et à l'aide du lemme 4.13, nous avons

$$\psi \circ \phi((\dots)_x \triangleleft (\dots)_y) = \psi(\phi((\dots)_x) \curvearrowleft \phi((\dots)_y)) \tag{44a}$$
$$= \psi \circ \phi((\dots)_x) \triangleleft \psi \circ \phi((\dots)_y) \tag{44b}$$
$$= (\dots)_x \triangleleft (\dots)_y \tag{44c}$$

Théorème 4.14 : Les algèbres pré-Lie $(\mathbb{T}, \curvearrowleft)$ et $\mathcal{PL}[u]$ sont en bijection par l'application ϕ de réciproque ϕ.

Remarque : Nous avons construit une bijection entre l'algèbre pré-Lie $(\mathbb{T}, \curvearrowleft)$ et l'algèbre pré-Lie libre $\mathcal{PL}[u]$. Cela signifie que l'algèbre $(\mathbb{T}, \curvearrowleft)$ est libre. Un second résultat important est que l'ensemble $\mathbb{T}$ est engendré par $\bullet$ et l'opérateur $\curvearrowleft$. Il est aussi bon de noter que ce morphisme est unique.

12.2 Preuve de l'existence d'un unique morphisme

Démontrons le théorème 4.12. Avec le théorème 4.14 en main, cela devient une formalité.

Il existe une unique surjection de $\mathcal{PL}[u]$ sur toute algèbre pré-Lie $(V, \triangleleft)$ qui à u associe $v \in V$. En la composant par *phi*, on obtient le morphisme recherché.

Pour vérifier l'unicité, il suffit de vérifier que s'il existe deux morphisme ϕ et ϕ', ils ont alors les mêmes valeurs sur chaque expression en $\bullet$ et $\curvearrowleft$, c'est-à-dire pour chaque arbre, et qu'ils sont donc égaux.

12.3 Etude de cas paticuliers

Étudions quelques cas particuliers de morphismes sur les champs de vecteurs, formés de fonctions communes.

Soit $\Theta : (\mathbb{T}, \curvearrowleft) \to \mathrm{Vect}^{\mathcal{C}^{+\infty}}(\mathbb{R})$ un morphisme de l'espace vectoriel sur les arbres vers l'espace des champs de vecteurs infiniment différentiable. De plus, posons $y' = y^3$, c'est-à-dire $f(x)\partial_x = x^3\partial_x$. Alors, on remarque immédiatement que $e^x\partial_x$ ne figure pas dans l'image. On peut de plus déterminer facilement une famille génératrice : il s'agit de toute les combinaisons linéaires des expressions que l'on peut construire à partir de $x^3\partial_x$. En étudiant les expressions qui apparaissent, on remarque facilement que l'image est engendrée par les monômes $x^{2p+3}\partial_x$ où $p \in \mathbb{N}$.

On peut essayer "d'améliorer" la surjectivité en prenant le plus petit monôme "utilisable"[18], à savoir $x^2\partial_x$. Il manque toute fois $x\partial_x$ et $1\partial_x$ à l'image. Pour ce qui est de l'injectivité, c'est encore pire. En effet, $\Theta([\bullet, \bullet]) = K_1 x^4 \partial_x$ et $\Theta([[\bullet]]) = K_2 x^4 \partial_x$; elle est donc compromise.

13 Coefficients des différentielles élémentaires

Revenons sur les coefficients, fonction des arbres, que nous avions noté $\alpha(\tau)$ durant notre démonstration sur l'ordre d'une méthode de Runge-Kutta(cf: 4.1). Nous allons démontrer une bien belle façon de les exprimer. Pour la démontrer, nous aurons besoin de différencier les sommets

[18] Les cas $x\partial_x$ et $1\partial_x$ sont dégénérés, et n'engendrent que des polynômes de degré ≤ 1 pour le premier, et 0 pour le second.

d'un arbre, et nous leur associerons donc une étiquette, par exemple une lettre. Le choix des noms est parfaitement libre, mais doit rester fixé une fois posé. Par exemple, nous pouvons nommer la corolle à quatres sommets de la façon suivante : $[b, c, d]_a$ où la racine est a, et les trois autres sommets sont b, c et d.

De plus, nous considérerons que les arêtes entre les sommets sont orientées de la racine vers les feuilles. On notera $a < b$ pour dire qu'il existe une arête orientée de a vers b. Nous pouvons alors définir l'ensemble des automorphismes d'un arbre.

Définition 5.1 : On notera $Aut(\tau) = \{\sigma : \text{Som.}(\tau) \simeq \text{Som.}(\tau) | \sigma(\bullet) = \sigma(\bullet), \sigma(a) < \sigma(b) \Leftrightarrow a < b\}$ le sous-ensemble de $\{\sigma \in \text{Perm.}(\text{Sommets de } \tau)\}$ qui préserve "les voisins" de chaque sommets, c'est-à-dire que si $a < b$, alors $\sigma(a) < \sigma(b)$.

Remarque : Cet ensemble correspond aux différentes façon d'échanger les sommets d'un arbre sans que son dessin n'en soit affecté. Si l'on considère la corolle à quatre sommets, on a alors $\{\text{Perm.}(a, b, c)\}$. Si l'on considère plutôt $[[c]_b]_a$, la seule permutation est l'identité.

De plus, nous noterons $\tau!$ le coefficients $\gamma(\tau) = |\tau|\gamma(\tau_1)\dots\gamma(\tau_k)$ qui est "la factorielle d'un arbre".

Théorème 5.2 : Les coefficients $\alpha(\tau)$ respectent

$$\alpha(\tau)\#Aut(\tau) = \frac{(|\tau|)!}{\tau!}$$

Démonstration : Pour démontrer ce résultat, nous allons observer un ensemble dont nous pourrons démontrer l'égalité de son cardinal avec chacun des deux membres de cette équation.

On notera $D(\tau) = \{\phi : \text{Som.}(\tau) \simeq \{1, \dots, n := |\tau|\} | a < b \Leftrightarrow \phi(a) < \phi(b)\}$, l'ensemble des bijections "croissantes" entre les sommets de l'arbre et l'ensemble $\{1, \dots, n\}$. Cela représente le nombre de façons de numéroter un arbre (en différenciant chaque sommets) en partant de la racine, et en remontant dans les feuilles, de telle façon que tout sommet $b > a$ se voit attribuer une valeur supérieure à celle du sommet sur lequel il repose.

Exemple : l'arbre $[b, c]_a$ peut être numéroté de deux façons : $[2, 3]_1$ et $[3, 2]_1$.

L'ensembles $Aut(\tau)$ et $D(\tau)$ est un groupe pour la loi $\circ$, et l'on peut donc le faire agir sur $D(\tau)$, par une application

$$\begin{array}{ccc} Aut(\tau) \times D(\tau) & \longrightarrow & D(\tau) \\ (\sigma, \phi) & \longmapsto & \sigma\circ \end{array}$$

On vérifie bien que $\phi \circ \sigma(a) < \phi \circ \sigma(b) \Leftrightarrow a < b$. On rappelle que l'orbite de $\phi \in D(\tau)$ est le sous-ensemble de $D(\tau)$ défini par $O_\phi = \{\phi \circ \sigma, \sigma \in Aut(\tau)\}$, et que l'on peut définir la relation $\tilde{}$ d'équivalence "appartenir à la même orbite". Or, l'orbite d'une bijection ϕ est l'ensemble des différentes façon de numéroter un arbre qui sont liées par un échange de sommets respectant les voisins. Par exemple, $[[4]_2, 3]_1$ et $[[3]_2, 4]$. Le nombre d'orbites est donc le nombre de façon de compter un arbre $\tau \in \mathbb{T}$, en considérant comme identiques deux numérotations qui n'ont pour différence qu'un échange de sommets. De plus, le coefficient $\alpha(\tau)$ correspond au nombre de façon de construire un arbre en ajoutant les noeuds en partant de la racine et en se dirigeant vers les feuilles. Or, le nombre de façons de construire un arbre est exactement le nombre de façons de numéroter de manière croissante, sans considérer comme important l'ordre dans lequel on choisit d'ajouter des sommets (c'est à dire que pour construire $[[c]_b, c]_a$, on peut d'abord choisir de construire $[[c]_b]_a$ avant d'ajouter d, ou bien d'abord $[b, c]_a$. et l'on a donc $\#D/\tilde{}(\tau) = \alpha(\tau)$.

Nous admettrons([Rob03], [Rot99]) que si le seul point fixe de l'action du groupe est l'élément neutre, c'est-à-dire que $\phi \circ \sigma = \phi \Rightarrow \sigma = id_\tau$, alors nous avons l'égalité

$$D/(\tau) \times \#Aut(\tau) = \#D(\tau)$$

Il est ici évident que le seul point fixe est l'identité, et nous avons donc terminé la première partie de notre démontration.

Si l'on considère $\tau = [\tau_1, \ldots, \tau_k]$, on peut numéroter chacun de ces sous-arbres et construire les $D(\tau_i)$ qui leur correspondent. Une fois chacun de ces ensembles choisis, on pourra alors construire une nouvelle bijection croissante. Pour ce faire, nous pouvons prendre une bijection de chacun des $D(\tau_i)$, puis, par la suite, si l'on appelle $n_i := |\tau_i$, nous pouvons alors renuméroter chacune des précédentes à partir de l'ensemble $\{1, \ldots, n_1 + \cdots + n_k\}$. Combien y a-t-il de façon de renuméroter ? Autant de façon que nous pouvons extraire de groupes de n_1, …, n_k, éléments croissants. En d'autres termes, c'est le nombre de façon de colorier $n_1 + \ldots n_k$ éléments en k couleurs différentes, où la couleur i apparaît n_i fois. On reconnaît là la définition du coefficient multinomial :

$$\binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k} = \frac{(n_1 + \cdots + n_k)!}{n_1! \ldots n_k!}$$

C'est bien le nombre de permutations totales divisé par le nombre de permutations dans chacun des k groupes de n_k éléments.

On a alors une bijection $D(\tau) \simeq \left(\prod_{i=1}^k D(\tau_i)\right) \times 1, \ldots, \binom{n_1+\cdots+n_k}{n_1,\ldots,n_k}$ et il vient

$$\#D(\tau) = \prod_{i=1}^k \#D(\tau_i) \binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k}$$

Par récurrence sur le nombre de sommets, on démontre que $\#D(\tau) = \frac{(|\tau|)!}{\tau!}$. En effet

$$\#D(T) = \prod_{i=1}^k D(\tau_i) \binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k} \tag{45a}$$

$$= \prod_{i=1}^k \frac{(|\tau_i|)!}{\tau_i!} \binom{n_1 + \cdots + n_k}{n_1, \ldots, n_k} \tag{45b}$$

$$= \prod_{i=1}^k \frac{(|\tau_i|)!}{\tau_i!} \frac{|\tau| - 1}{\prod_{i=1}^k (|\tau_i|)!} \tag{45c}$$

$$= \frac{(|\tau| - 1)!}{\prod_{i=1}^k \tau_i!} \tag{45d}$$

$$= \frac{(|\tau|)(|\tau| - 1)}{(|\tau|) \prod_{i=1}^k \tau_i!} \tag{45e}$$

Ce qui nous permet de conclure.

13.1 Comment jouer avec ces coefficients des différentielles élémentaires

Après avoir exprimer les coefficients des différentiels élémentaires, on est en droit de se questionner sur l'existence de formules plus simples les mettant en jeu. C'est effectivement le cas,

puisque nous allons établir la propriété suivante.

Lemme 5.4 :

$$\sum_{|\tau|=n} \alpha(\tau) = (n-1)!$$

Démonstration : Considérons l'équation différentielle $y' = e^y$. Une solution est $-\ln(1-t)$, puisque $-\ln(1-t)' = \frac{1}{1-t} = e^{(-\ln(1-t))}$.

Si l'on applique le développement en série entière de cette expression, et qu'on la compare à la série de Taylor contenant des $\alpha(\tau)$, on obtient

$$\sum_{n\geq 1} \frac{1}{n} t^n = \sum_{n\geq 1} \left(\sum_{|\tau|=n} \frac{\alpha_(\tau)}{n!} F(\tau)(y_0) \right) t^n$$

Puis, en appliquant la condition initiale $y_0 = 0$, les différentiels élémentaires sont de la forme $F(\tau)(0) = \prod_{i=1}^{|tau|} e^0 = 1$, et il vient

$$\sum_{n\geq 1} \frac{1}{n} t^n = \sum_{n\geq 1} \left(\sum_{|\tau|=n} \frac{\alpha_(\tau)}{n!} \right) t^n$$

d'où l'on tire immédiatement le résultat attendu.

14 Remerciements

Je remercie mes encadrants de TIPE, Frédéric Chapoton et Thierry Dumont, pour m'avoir prodigué leurs conseils, avoir su m'orienter, et avoir répondu à mes nombreuses questions. Je remercie aussi Antoine Pinochet Lobos pour avoir relu ce document, ma muse, Aurélie, et mes amis pour leur soutien.

Références

[AA01] Alexei A. Medovikov Assyr Abdulle. Second order chebyshev methods based on orthogonal polynomials. 2001.

[Cha00] Chapoton. Algèbres pré-lie et algèbres de hopf liées à la renormalisation. Novembre 2000.

[EH] Gerhard Wanner Ernst Hairer. Intégration numérique des équations différentielles raides.

[EH91] Gerhard Wanner Ernst Hairer. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. 1991.

[EH06] Gerhard Wanner Ernst Hairer, Christian Lubich. *Geometric numerical integration: structure-preserving algorithms for Ordinary Differential Equations*. Avril 2006.

[FC01] Muriel Livernet Frédéric Chapoton. Pre-lie algebras and the rooted trees operad. 2001.

[Fie07] Richard J. Field. Oregonator. 2007. `http://www.scholarpedia.org/article/Oregonator`.

[Hai] Ernst Hairer. Ch 3 : Equations différentielles ordinaires. `http://www.unige.ch/~hairer/poly/chap3.pdf`.

[IRE98] John Anthony Pojman Irving Robert Epstein. *An introduction to nonlinear chemical dynamics*. 1998.

[Rob03] Derek J.S. Robinson. *An introduction to abstract algebra*. 2003.

[Rot99] Joseph J. Rotman. *An introduction to the theory of groups*. 1999.

[Sze59] Szegö. *Orthogonal Polynomials*. 1959.

[W00] Stiffness and absolute stability. `http://sundials.wikidot.com/stiffness`.

[W01] Réaction oscillante de belousov-zhabotinsky. `http://scienceamusante.net/wiki/index.php?title=Belousov-Zhabotinsky`.

[W02] Belousov-zhabotinsky. `http://www.scholarpedia.org/article/Belousov-Zhabotinsky`.

[Win84] A. T. Winfree. The prehistory of the belousov-zhabotinsky oscillator. Aout 1984. `http://campus.usal.es/~licesio/Biofisica/Winfree_JCE1984.pdf`.

Index