Initiation à l'Analyse Numérique

Robert Rolland

C.N.R.S. Institut de Mathématiques de Luminy, case 930, F13288 Marseille cedex 9, France.
E-mail address: rolland@iml.univ-mrs.fr

Table des matières

Préface

Cet ouvrage est une initiation à l'analyse numérique. Les notions qui y sont exposées et développées concernent les objets et méthodes de base de ce vaste domaine des mathématiques. Le niveau est celui des deux ou trois premières années d'Université. Nous n'abordons pas les méthodes techniquement très élaborées. Nous n'abordons pas non plus l'immense domaine des équations aux dérivées partielles et équations fonctionnelles générales faisant intervenir les résultats généraux sur les espaces fonctionnels.

Nous essayons de montrer comment l'algèbre linéaire d'une part et quelques outils classiques de l'analyse d'autre part, permettent de poser clairement les problèmes d'approximation, d'interpolation, de calculs d'intégrales et de solutions d'équations différentielles. Nous donnons les méthodes numériques classiques concernant ces questions.

Robert Rolland

CHAPITRE 1

Approximations de solutions d'équations

1.1. Introduction

Nous voulons résoudre numériquement l'équation :

$$f(x) = 0 \tag{1.1}$$

où f est une fonction suffisamment régulière. Nous supposerons f au moins continue, parfois de classe C^1 ou plus. Il existe de nombreuses méthodes sophistiquées pour attaquer ce problème. Nous donnons ici deux idées importantes à la base de la plupart de ces méthodes : d'une part la dichotomie, d'autre part l'approximation de Newton. Ces deux idées seront illustrées par deux méthodes correspondantes.

1.1.1. Dichotomie. Nous supposons que nous avons isolé le zéro c que nous voulons calculer, c'est-à-dire que nous avons trouvé un intervalle [a,b] tel que :

$$c \in]a, b[,$$
$$\forall x \in [a, b] \setminus \{c\}, f(x) \neq 0$$

Nous supposerons en outre que f change de signe en c, donc que :

$$f(a)f(b) < 0.$$

La méthode de calcul de c par dichotomie est très simple : on coupe l'intervalle sur lequel on travaille en deux et on teste sur lequel des deux sous-intervalles obtenus se trouve c. On réitère le procédé à partir du sous-intervalle déterminé. Après n itérations on localise le nombre c sur un intervalle de longueur $(b-a)/2^n$. Voici les détails :

On fixe $\epsilon > 0$,

$A := a$;
$B := b$;
$E := \epsilon$;
tant que $B - A > E$ **faire**
  $C := (A + B)/2$;
  si $f(C) = 0$ **alors**
    retourner C ;
    sortir ;
  finsi ;
  si $f(A)f(C) < 0$ **alors**
    $B := C$;
  sinon
    $A := C$;

finsi ;
fintq ;
retourner C ;

1.1.2. Approximations successives, point fixe. Avant de présenter la méthode de Newton, disons quelques mots sur la méthode générale des approximations successives. Soit f une fonction de classe C^1 définie sur un intervalle $[a, b]$ à valeurs dans l'intervalle $[a, b]$. Soit $u_0 \in [a, b]$. On définie alors par récurrence la suite de terme général u_i en posant pour tout $i \geq 1$ $u_i = f(u_{i-1})$. On étudie la convergence de cette suite. Nous allons montrer que sous certaines conditions on peut affirmer qu'il existe un point fixe pour f, c'est-à-dire une solution de l'équation $f(x) = x$ (cf. figure 1).

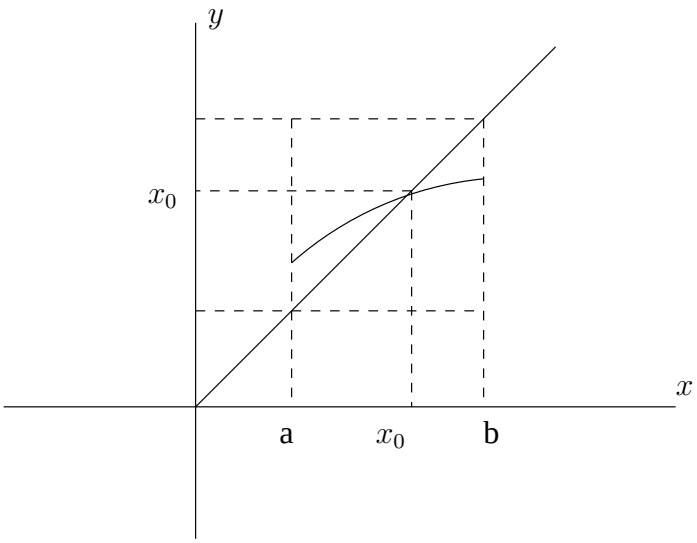

FIG. 1. Théorème du point fixe

THÉORÈME 1.1 (Théorème du point fixe). *Sous les hypothèses précédentes, si on suppose que :*

$$\sup_{x \in [a,b]} |f'(x)| = k < 1,$$

alors la fonction $f(x)$ admet un point fixe x_0 et un seul sur l'intervalle $[a, b]$. La suite $(u_i)_i$ définie précédemment converge vers le point fixe x_0.

DÉMONSTRATION. D'après le théorème des accroissements finis :

$$|u_2 - u_1| = |f(u_1) - f(u_0)| \leq k|u_1 - u_0|,$$

$$|u_3 - u_2| = |f(u_2) - f(u_1)| \leq k|u_2 - u_1| \leq k^2|u_1 - u_0|,$$

et en itérant ce calcul :

$$|u_n - u_{n-1}| \leq k^{n-1}|u_1 - u_0|.$$

On en déduit que si $n \geq p$:

$$|u_n - u_p| \leq \frac{k^p(1-k^{n-p})}{1-k}|u_1 - u_0|.$$

Donc la suite $(u_n)_n$ est une suite de Cauchy, elle converge vers un point x_0. Ce point est un point fixe, en effet, $u_{n+1} = f(u_n)$, comme la suite de terme général u_n converge vers x_0 et que la fonction f est continue, on conclut que $f(u_n)$ converge vers $f(x_0)$ et donc que $x_0 = f(x_0)$. On peut voir que le point fixe est unique en constatant que si s est un point fixe alors :

$$|u_n - s| = |f(u_{n-1}) - f(s)| \leq k|u_{n-1} - s|,$$

et en itérant ce calcul :

$$|u_n - s| \leq k^n|u_0 - s|.$$

Ceci prouve en effet que s est la limite de la suite de terme général u_n, et donc que $s = x_0$. □

Exemple 1.2 (Construction d'une table trigonométrique). Les propriétés géométriques des lignes trigonométriques permettent de calculer facilement $\sin(\pi/6)$, $\sin(\pi/4)$, $\sin(\pi/3)$, $\sin(\pi/5)$ ainsi évidemment que les cosinus des mêmes angles. En utilisant la formule qui donne le sinus d'une différence on trouve $\sin(\pi/30)$, en utilisant la formule qui donne le sinus de l'arc moitié on obtient $\sin(\pi/60)$. On a donc une table donnant les sinus (et les cosinus) de 3° en 3°. Il faudrait donc calculer $\sin(\pi/180)$ pour avoir une table trigonométrique donnant les lignes de tous les angles en degré entier. On utilise alors la formule :

$$\sin(3x) = 3\sin(x) - 4\sin^3(x).$$

On cherche alors à résoudre cette équation connaissant la valeur $a = \sin(3x)$. L'équation se ramène à :

$$\sin(x) = \frac{1}{3}(4\sin^3(x) + a).$$

Donc si on pose :

$$f(u) = \frac{1}{3}(4u^3 + a),$$

on est amené à chercher un point fixe de f. Ici u est proche de 0 ($\sin(\pi/180)$), donc la dérivée est dans un voisinage du point fixe bien plus petite que 1 en valeur absolue, et le théorème précédent s'applique. On peut prendre par exemple $u_0 = a/3$.

1.1.3. Méthode de Newton. Il s'agit ici de remplacer la fonction f dont on cherche un zéro par sa tangente en un point voisin du zéro cherché (cf. figure 2).

Ainsi à partir d'un point u_0 proche de la solution x_0 de l'équation $f(x) = 0$ (qu'on supposera unique, tout au moins dans un intervalle adapté), on construit la tangente à la courbe $y = f(x)$ au point $(u_0, f(u_0)$. Cette tangente coupe l'axe des abscisses au point u_1. On itère cette construction à partir de la valeur u_1 pour obtenir le point u_2. Le calcul de l'équation de la tangente au point d'abscisse x et de son intersection avec l'axe des x montre que si on introduit :

$$F(x) = x - \frac{f(x)}{f'(x)},$$

alors on peut écrire :

$$u_1 = F(u_0), u_2 = F(u_1), \cdots, u_n = F(u_{n-1}), \cdots$$

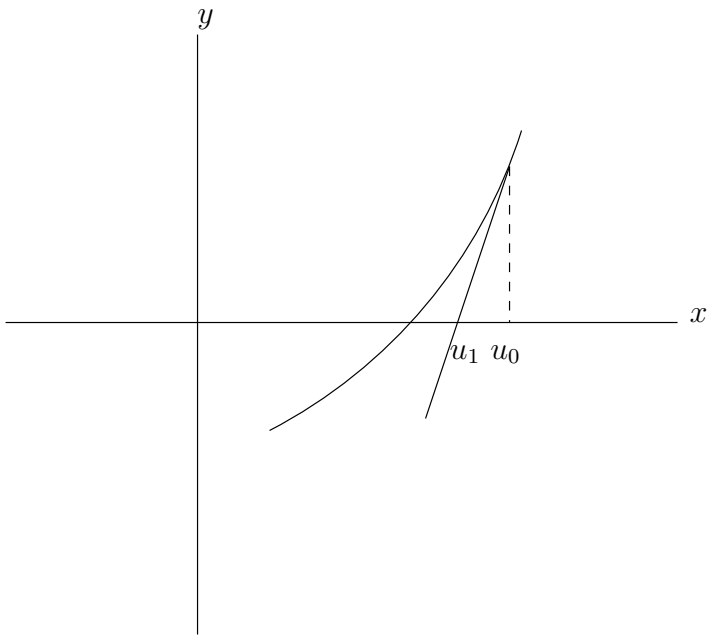

FIG. 2. Méthode de Newton

On vérifie imédiatement que x_0 est point fixe de la fonction $F(x)$. De plus si on suppose la fonction f de classe C^2 par exemple et si on suppose que la dérivée f' de f ne s'annule pas, alors $F(x)$ est dérivable et sa dérivée est nulle au point x_0. Il y aura donc un voisinage fermé de ce point fixe, pas nécessairement simple à déterminer effectivement, dans lequel la théorie du paragraphe précédent s'applique. De plus, comme la dérivée de F sera proche de zéro, on peut espérer une bonne convergence de la suite $(u_n)_n$, d'autant plus rapide qu'on va se rapprocher du point fixe. Nous n'en dirons pas plus dans le cas général, nous allons développer deux exemples.

1.1.3.1. *Exemple 1.* Considérons la fonction :

$$f(x) = x^2 - 2.$$

En nous restreignant à $x \geq 0$, nous allons déterminer la solution positive de l'équation $x^2 = 2$, c'est à dire $x = \sqrt{2}$. Introduisons donc, comme nous l'indique la méthode de Newton, la fonction :

$$F(x) = x - \frac{x^2 - 2}{2x} = x - \frac{x}{2} + \frac{1}{x} = \frac{1}{2}\left(x + \frac{2}{x}\right).$$

On voit que par exemple $F([1,2]) \subset [1,2]$ et sur cet intervalle la dérivée de $F(x)$ est en valeur absolue majorée par $1/2$. Donc le théorème du point fixe s'applique sur cet intervalle. Nous avons successivement :

$$\frac{1}{2}\left(u_n + \frac{2}{u_n}\right) - \sqrt{2} = F(u_n) - F(\sqrt{2}) = \frac{1}{2u_n}\left(u_n^2 + (\sqrt{2})^2 - 2u_n\sqrt{2}\right),$$

$$\frac{1}{2}\left(u_n + \frac{2}{u_n}\right) - \sqrt{2} = \frac{1}{2u_n}\left(u_n - \sqrt{2}\right)^2,$$

$$\left|\frac{1}{2}\left(u_n+\frac{2}{u_n}\right)-\sqrt{2}\right| \leq \frac{1}{2}\left(u_n-\sqrt{2}\right)^2,$$

c'est-à-dire :

$$|u_{n+1}-\sqrt{2}| \leq \frac{1}{2}\left(u_n-\sqrt{2}\right)^2.$$

La convergence est donc **quadratique**.

1.1.3.2. *Exemple 2.* Considérons la fonction :

$$f(x)=a-\frac{1}{x},$$

où $a>0$, que nous étudierons pour $x>0$. En appliquant la méthode de Newton à l'équation $f(x)=0$, nous obtiendrons une façon de calculer $1/a$. Introduisons la fonction :

$$F(x)=x(2-ax).$$

Soit I l'intervalle ouvert $]0,2/a[$. Si $u_0 \in I$ alors $0<F(u_0)\leq 1/a$. Posons ensuite $u_1=F(u_0)$ et par récurrence $u_n=F(u_{n-1})$. La suite $(u_n)_{n\geq 1}$ est croissante, majorée par $1/a$, donc converge vers le point fixe $1/a$ de F. En renumérotant la suite on peut supposer par exemple que $|u_1-1/a|\leq \frac{1}{10a}$. Nous pouvons écrire :

$$|au_{n+1}-1|=|2au_n-a^2u_n^2-1|=(au_n-1)^2.$$

Donc :

$$|au_{n+1}-1|=(au_1-1)^{2^n}\leq\left(\frac{1}{10}\right)^{2^n},$$

ou encore :

$$\left|u_{n+1}-\frac{1}{a}\right|\leq\frac{1}{a}\left(\frac{1}{10}\right)^{2^n}.$$

Là encore nous avons obtenu une convergence quadratique.

CHAPITRE 2

Les outils classiques d'approximation

2.1. Introduction

Lorsqu'on souhaite approcher une fonction, ses valeurs en des points particuliers, des sommes de séries, ou des valeurs d'intégrales etc. on est amené à utiliser les outils de l'analyse et en particulier le calcul intégral. Nous verrons par la suite diverses méthodes d'approximation d'objets divers de l'analyse, en particulier l'interpolation en des points bien choisis par des fonctions élémentaires tout aussi bien choisies. Ici nous commençons par un outil très simple mais particulièrement efficace : l'intégration par partie. Si l'on veut bien laisser de côté le volet anecdotique du calcul exact de certaines primitives, il faut comprendre l'intégration par partie comme l'écriture d'une expression sous forme d'une partie principale (la partie tout intégrée) et d'un reste (la deuxième intégrale)

$$\int_a^x f'(t)g(t)dt = [f(t)g(t)]_a^x - \int_a^x f(t)g'(t)dt.$$

Évidemment, si on espère que la partie

$$-\int_a^x f(t)g'(t)dt$$

puisse être considérée comme un reste, il faut qu'elle soit négligeable devant la partie tout intégrée

$$[f(t)g(t)]_a^x.$$

Cette remarque nous indique comment choisir, quand on fait une intégration par partie d'un produit de deux fonctions, celle qu'on intègre et celle qu'on dérive : celle qu'on dérive est en général celle qui varie peu, de manière à obtenir une dérivée petite. Attention ceci ne s'applique pas pour l'utilisation de l'intégration par partie pour des calculs de primitives, ni pour l'établissement de formules théoriques où l'on veut obtenir une forme particulière pour le terme tout intégré.

Dans la suite de ce chapitre, nous donnons deux applications très importantes de l'intégration par partie : la formule de Taylor et la formule d'Euler-Maclaurin. Ces deux formules constituent des outils de base des méthodes d'approximation. Nous exprimerons toutes ces formules asymptotiques en utilisant des restes intégraux qui donnent des formules exactes, et en évitant au maximum les formules faisant intervenir des restes écrits avec des points dont on ne connait pas la valeur mais juste une localisation. En général, dans un vrai problème on n'a jamais vraiment besoin de ces formules et les formules avec reste intégral ainsi que l'outil "intégration par partie" permettent d'obtenir les évaluations dont on a besoin.

2.1.1. Intégrons dans le bon sens.

Exemple 1. Il s'agit d'évaluer au voisinage de $+\infty$ l'intégrale

$$\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^2} du.$$

On choisit clairement de dériver la fonction $1/u^2$ et d'intégrer la fonction e^{-u}, on obtient

$$\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^2} du = \frac{1}{x(\ln(x))^2} - 2\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^3} du.$$

Cette manière de faire permet effectivement d'obtenir un reste négligeable devant la partie intégrée. En effet :

$$\frac{e^{-u}}{u^3} = o\left(\frac{e^{-u}}{u^2}\right),$$

donc :

$$\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^3} du = o\left(\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^2} du\right)$$

et en conséquence :

$$\int_{\ln(x)}^{+\infty} \frac{e^{-u}}{u^2} du \sim \frac{1}{x(\ln(x))^2}.$$

Exemple 2 (communiqué par Raymond Raynaud). Il s'agit de calculer

$$I_p = \int_1^e x^2 (\ln(x))^p dx.$$

On a $I_0 = (e^3 - 1)/3$.

2.1.1.1. *Le piège.* Le piège consiste ici à voir apparaître une formule très simple et à se laisser aller à intégrer par partie "à l'envers". C'est-à-dire qu'on va dériver $(\ln(x))^p$ et intégrer x^2.

$$I_p = \frac{e^3}{3} - \frac{p}{3} I_{p-1}.$$

Si nous itérons le calcul nous sommes amenés à écrire la formule exacte

$$I_p = \frac{e^3}{3}\left(1 - \frac{p}{3} + \frac{p(p-1)}{9} + \cdots + (-1)^{p-1}\frac{p!}{3^{p-1}}\right) + (-1)^p \frac{p!}{3^{p-1}}(e^3 - 1).$$

Hélas, comme à chaque étape on a pris un "reste plus grand que la partie qui aurait dû être principale", cette formule est très instable numériquement. En effet si on change un peu la valeur initiale I_0 en J_0, alors le terme J_p calculé vérifie

$$J_p - I_p = (-1)^p \frac{p!}{3^p}(J_0 - I_0),$$

ce qui fait que si $J_0 \neq I_0$, alors $|J_p - I_p| \to +\infty$. Donc une erreur d'arrondi va complètement modifier le calcul.

2.1.1.2. *L'intégration dans le bon sens.* Bien sûr, si on intègre dans le bon sens en exhibant une partie principale et un reste plus petit que cette partie principale, alors ceci ne se produit plus. Intégrons donc $\frac{(\ln(x))^p}{x}$ et dérivons x^3. Nous obtenons

$$I_p = \frac{e^3}{p+1} - \frac{3}{p+1} I_{p+1},$$

et cette fois-ci en itérant le calcul on tombe sur une formule stable qui nous permet d'écrire tout de suite

$$I_p \sim \frac{e^3}{p}.$$

Si on veut pousser le développement plus loin on obtient :

$$I_p = e^3\left(\frac{1}{p} - \frac{4}{p^2}\right) + o\left(\frac{1}{p^2}\right).$$

Bien sûr la formule de récurrence trouvée est la même que dans le premier calcul, écrite différemment. Mais justement ceci nous permet de voir que si on a une vision "à l'envers" de l'intégration par partie, alors la suite des calculs qu'on est amené naturellement à faire conduit à de mauvaises situations.

2.2. La formule de Taylor

THÉORÈME 2.1. *Soit f une fonction à valeurs réelles définie sur le segment $[a-\epsilon, a+\epsilon]$ et $n+1$ fois continument dérivable sur ce segment. Alors pour tout point x du segment $[a-\epsilon, a+\epsilon]$ on peut écrire*

$$f(x) = f(a) + \frac{(x-a)}{1!}f(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)dt.$$

Preuve. On utilise une démonstration par récurrence. La formule

$$\int_a^x f'(t)dt = f(x) - f(a),$$

assure que le théorème est vrai pour $n = 0$. Si on suppose vraie la formule à l'ordre $n \geq 0$ alors

$$\int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)dt = \left[\frac{-(x-t)^{n+1}}{(n+1)!}f^{(n+1)}(t)\right]_a^x + \int_a^x \frac{(x-t)^{n+1}}{(n+1)!}f^{(n+2)}(t)dt,$$

$$\int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)dt = \frac{(x-a)^{n+1}}{(n+1)!}f^{(n+1)}(a) + \int_a^x \frac{(x-t)^{n+1}}{(n+1)!}f^{(n+2)}(t)dt,$$

ce qui nous donne la formule à l'ordre $n+1$. □

2.3. La formule d'Euler-Maclaurin

2.3.1. Un exemple à la main. Soit f une fonction de classe C^3 sur $[0,1]$. Cherchons à exprimer

$$\int_0^1 f(t)dt.$$

Pour cela on peut commencer par dire en remplaçant f par sa valeur en 0 qu'une valeur approchée de l'intégrale est $f(0)$. Étudions alors

$$W = \int_0^1 f(t)dt - f(0).$$

on voit que

$$W = \int_0^1 (f(t) - f(0))\,dt$$

ce qui par intégration par partie (on dérive f(t) -f(0) et on intégre 1) donne

$$W = \left[P_1(t)\,(f(t) - f(0))\right]_0^1 - \int_0^1 f'(t)P_1(t)dt,$$

où $P_1(t)$ est une primitive de 1 (donc de la forme $t-a$) à bien choisir.

$$W = (1-a)\,(f(1)-f(0)) - \int_0^1 f'(t)P_1(t)dt.$$

L'intégrale qui reste dans le second membre peut à son tour être intégrée par partie en dérivant f' et en intégrant P_1. Soit $P_2(t)$ une primitive de $P_1(t)$. Choisissons $P_1(t)$ tel que $P_2(1) = P_2(0)$ ou encore

$$\int_0^1 P_1(t)dt = 0.$$

Ceci impose de prendre $a = 1/2$ ($P_1(t) = t - 1/2$). On a alors

$$W = 1/2\,(f(1)-f(0)) - P_2(0)\,(f_1'(t)-f'(0)) + \int_0^1 P_2(t)f^{(2)}(t)dt.$$

Intégrons de nouveau par partie l'intégrale qui subsiste au second membre. Notons $P_3(t)$ une primitive de $P_2(t)$ et choisissons $P_2(t)$ de telle sorte que $P_3(1) = P_3(0)$. Comme $P_1(t) = t - 1/2$ on a $P_2(t) = t^2/2 - t/2 + C$ et la condition imposée

$$\int_0^1 P_2(t)dt = 0$$

donne $C = 1/12$. Alors $P_3(t) = t^3/6 - t^2/4 + t/12 + C$, et si on impose aussi la condition

$$\int_0^1 P_3(t)dt = 0$$

alors $P_3(t) = t^3/6 - t^2/4 + t/12$. On obtient après une nouvelle intégration par partie :

$$W = 1/2\,(f(1)-f(0)) - 1/12\,(f_1'(t)-f'(0)) - \int_0^1 P_3(t)f^{(3)}(t)dt.$$

Arrêtons là le développement et écrivons en conclusion

$$\int_0^1 f(t)dt = 1/2\,(f(0)+f(1)) - 1/12\,(f'(1)-f'(0)) - \int_0^1 P_3(t)f^{(3)}(t)dt.$$

2.3.2. Polynômes de Bernoulli. Les calculs précédents mettent en évidence la suite des polynômes de Bernoulli.

THÉORÈME 2.2. *Il existe une suite $(Q_n)_{n\geq 0}$ et une seule de polynômes telle que*

(1) $Q_0 = 1$
(2) $Q_n' = Q_{n-1}$, *pour* $n \geq 1$
(3) $\int_0^1 Q_n(u)du = 0$, *pour* $n \geq 1$.

Ces polynômes sont appelés **polynômes de Bernoulli**.

Preuve. Par récurrence, Q_0 est bien défini, Q_{n-1} étant construit, la condition (2) fixe Q_n à une constante près. La condition (3) fixe cette constante. □

Compte tenu du mode de construction de ces polynômes, il est facile de voir que leurs coefficients sont rationnels.

Voici les premiers polynômes de Bernoulli :

$$Q_0 = 1 \quad Q_1 = x - \frac{1}{2} \quad Q_2 = \frac{x^2}{2} - \frac{x}{2} + \frac{1}{12}$$
$$Q_3 = \frac{x^3}{6} - \frac{x^2}{4} + \frac{x}{12}.$$

PROPOSITION 2.3. *Pour $n \geq 2$, $Q_n(1) = Q_n(0)$.*

Preuve. En effet on doit avoir $\int_0^1 Q_{n-1}(u)du = 0$. Mais comme Q_n est une primitive de Q_{n-1} on a $\int_0^1 Q_{n-1}(u)du = Q_n(1) - Q_n(0)$, d'où le résultat. □

PROPOSITION 2.4. *Si $n \geq 1$, $Q_n(x+1) - Q_n(x) = \frac{x^{n-1}}{(n-1)!}$.*

Preuve. On constate que l'égalité est vraie pour $n = 1$. Supposons la vraie pour n, par primitivation on obtient alors l'égalité à l'ordre $n+1$ à une constante près. Mais la proposition précédente permet de conclure à la nullité de cette constante d'intégration. □

Cette égalité permet de calculer des sommes du type $\sum_{k=1}^p k^n$. Par exemple si $n = 2$ on obtient

$$Q_3(p+1) - Q_3(1) = \frac{1}{2}\sum_{k=1}^{p} k^2,$$

ce qui donne

$$\sum_{k=1}^{p} k^2 = \frac{p(p+1)(2p+1)}{6}.$$

PROPOSITION 2.5. *Pour tout $n \geq 0$, $Q_n(1-x) = (-1)^n Q_n(x)$.*

Preuve. Le résultat est vrai pour $n = 0$ et $n = 1$. Supposons le résultat vrai pour $n \geq 2$. Alors par primitivation on obtient au rang $n+1$ la formule voulue à une constante d'intégration près

$$Q_{n+1}(1-x) = (-1)^{n+1} Q_{n+1}(x) + C.$$

Si $n+1$ est pair en donnant à x la valeur 0 on calcule $C = 0$.

Si $n+1$ est impair alors par primitivation

$$Q_{n+2}(1-x) = Q_{n+2}(x) + Cx + D,$$

en prenant $x = 0$ on obtient d'abord $D = 0$, puis en prenant $x = 1$ on obtient $C = 0$. □

PROPOSITION 2.6. *Pour $n \geq 1$,*

$$Q_{2n+1}(0) = Q_{2n+1}(1/2) = Q_{2n+1}(1) = 0.$$

Preuve. Il suffit d'appliquer la proposition précédente avec $x = 0$, ce qui donne $Q_{2n+1}(0) = Q_{2n+1}(1) = 0$, puis avec $x = 1/2$, ce qui donne $Q_{2n+1}(1/2) = 0$. □

En étudiant par récurrence les variations des fonctions Q_n, on pourra montrer que

PROPOSITION 2.7. *Pour $n \geq 1$, $Q_{2n}(0) \neq 0$ et Signe$(Q_{2n}(0)) = (-1)^{n+1}$.*

On notera

$$B_k = (-1)^{k+1}(2k)!Q_{2k}(0).$$

Ainsi les B_k (**nombres de Bernoulli**) sont des rationnels positifs. Les premiers nombres de Bernoulli sont

$$B_1 = 1/6 \quad B_2 = 1/30 \quad B_3 = 1/42 \quad B_4 = 1/30.$$

2.3.3. Formule d'Euler-Maclaurin. Soit f une fonction réelle de classe C^∞ sur $[0,1]$.

$$f(1) - f(0) = \int_0^1 f'(t)dt,$$

$$f(1) - f(0) = [Q_1(t)f'(t)]_0^1 - \int_0^1 Q_1(t)f^{(2)}(t)dt,$$

$$f(1) - f(0) = 1/2(f'(1) + f'(0)) - \int_0^1 Q_1(t)f^{(2)}(t)dt,$$

et par récurrence on montre que :

THÉORÈME 2.8. *Soit f une fonction réelle de classe C^∞ sur $[0,1]$. Alors pour tout $n \geq 1$:*

$$f(1) - f(0) = \frac{1}{2}\big(f'(1) + f'(0)\big) + \sum_{k=1}^{n} \frac{(-1)^k B_k}{(2k)!}\big(f^{(2k)}(1) - f^{(2k)}(0)\big)$$

$$- \int_0^1 Q_{2n+1}(x)f^{(2n+2)}(x)dx.$$

En particulier si on applique ce résultat à une primitive de f on obtient

$$\int_0^1 f(t)dt = \frac{1}{2}\big(f(1) + f(0)\big) + \sum_{k=1}^{n} \frac{(-1)^k B_k}{(2k)!}\big(f^{(2k-1)}(1) - f^{(2k-1)}(0)\big)$$

$$- \int_0^1 Q_{2n+1}(x)f^{(2n+1)}(x)dx.$$

On peut appliquer le théorème 2.8 sur un intervalle $[a,b]$ au lieu de $[0,1]$. Il suffit comme toujours de faire le changement de variable affine qui envoie a sur 0 et b sur 1 : $t = a + u(b-a)$. On obtient alors :

THÉORÈME 2.9. *Soit f une fonction réelle de classe C^∞ sur $[a,b]$. Alors pour tout $n \geq 1$:*

$$f(b) - f(a) = \frac{b-a}{2}\big(f'(b) + f'(a)\big) + \sum_{k=1}^{n} \frac{(-1)^k B_k (b-a)^{2k}}{(2k)!}\big(f^{(2k)}(b) - f^{(2k)}(a)\big)$$

$$- \int_a^b Q_{2n+1}(\frac{u-a}{b-a})(b-a)^{2n-1} f^{(2n+2)}(u)du.$$

Découpons maintenant l'intervalle $[a,b]$ en q morceaux de même longueur $h = (b-a)/q$ sous la forme :

$$a = a_0 < a_1 < a_2 < \cdots < a_{q-1} < a_q = b,$$

appliquons le théorème 2.9 sur chaque morceau et sommons les résultats. Nous obtenons :

THÉORÈME 2.10. *Soit f une fonction réelle de classe C^∞ sur $[a,b]$. Soit $a = a_0 < a_1 < \cdots < a_{q_1} < a_q = b$ le partage de $[a,b]$ en q intervalles de même longueur $h = (b-a)/q$. Alors pour tout $n \geq 1$:*

$$f(b) - f(a) = f(a_q) - f(a_0) = \frac{h}{2}(f'(b) + f'(a)) + h\sum_{s=1}^{q-1} f'(a_s)$$

$$+\sum_{k=1}^{n} \frac{(-1)^k B_k h^{2k}}{(2k)!}\big(f^{(2k)}(b) - f^{(2k)}(a)\big)$$

$$-h^{2n-1}\int_0^h Q_{2n+1}\left(\frac{v}{h}\right)\left(\sum_{s=0}^{q-1} f^{(2n+2)}(a_s+v)\right)dv.$$

Cette dernière formule est appelée la formule sommatoire d'Euler-Maclaurin car elle permet de calculer la somme :

$$\sum_{s=1}^{q-1} f'(a_s).$$

2.3.4. Application à l'évaluation de restes de séries. Considérons la série de terme général $1/n^2$. On sait que cette série converge et que :

$$S = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Écrivons la somme S sous la forme :

$$S = S_p + R_p$$

où :

$$S_p = \sum_{k=1}^{p} \frac{1}{k^2} \quad \text{et} \quad R_p = \sum_{k=p+1}^{\infty} \frac{1}{k^2}.$$

Nous cherchons à évaluer R_p. Pour cela introduisons la fonction $f(x) = -1/x$ dont la dérivée est $f'(x) = 1/x^2$. Appliquons le théorème 2.10 à la fonction f sur l'intervalle $[p, r+1]$ (où p et r sont des entiers) avec $h = 1$ et $n = 1$. Nous obtenons successivement :

$$R_p - R_r = \frac{1}{(p+1)^2} + \frac{1}{(p+2)^2} + \cdots + \frac{1}{r^2},$$

$$R_p - R_r = \frac{1}{p} - \frac{1}{r+1} - \frac{1}{2}\left(\frac{1}{p^2} + \frac{1}{(r+1)^2}\right) + \frac{1}{6}\left(\frac{1}{p^3} - \frac{1}{(r+1)^3}\right) + T_{p,r},$$

et en faisant tendre r vers $+\infty$:

$$R_p = \frac{1}{p} - \frac{1}{2p^2} + \frac{1}{6p^3} + T_p,$$

où T_p se calcule facilement en utilisant le théorème 2.10. Un calcul simple (comparaison d'une somme de série avec une intégrale) permet de voir que :

$$T_p = O\left(\frac{1}{p^4}\right),$$

ce qui donne pour R_p le développement asymptotique :

$$R_p = \frac{1}{p} - \frac{1}{2p^2} + \frac{1}{6p^3} + O\left(\frac{1}{p^4}\right).$$

CHAPITRE 3

Interpolation des fonctions

3.1. Introduction

L'**interpolation** est un sujet très vaste lié aux questions d'**approximation** des fonctions. Très grossièrement il s'agit de trouver dans une classe fixée de fonctions (par exemple les fonctions polynomiales) un élément réalisant un certain nombre de contraintes. Souvent ces contraintes sont liées à la donnée d'une fonction f qu'on cherche à approcher par un procédé d'interpolation (par exemple la fonction cherchée doit prendre la même valeur que f en des points donnés). On se trouve alors confronté à plusieurs problèmes de natures différentes. Tout d'abord un problème algébrique, celui de trouver le ou les éléments de la classe choisie qui réalise les contraintes. Ensuite un problème d'approximation qui consiste lorsqu'on est parti d'une fonction f à mesurer la qualité de l'approximation théorique obtenue. Enfin un problème algorithmique, celui de déterminer un algorithme performant qui permette de calculer facilement et de manière aussi exacte que possible la ou les solutions.

Dans un premier temps nous partirons d'un exemple important : l'interpolation de Lagrange.

3.2. Interpolation de Lagrange

Soient $x_0, x_1, ..., x_n$ des nombres complexes distincts et $y_0, y_1, ..., y_n$ des nombres complexes. Il s'agit de trouver un polynôme $P(X)$ vérifiant $P(x_k) = y_k$ pour toutes les valeurs de k comprises entre 0 et n.

3.2.1. L'aspect algébrique.

Existence de solutions. Notons $\mathbb{C}[X]$ l'espace des polynômes à coefficients complexes et $\mathbb{C}_n[X]$ le sous espace des polynômes à coefficients complexes de degré inférieur ou égal à n. Considérons alors l'application linéaire T de $\mathbb{C}[X]$ dans $\mathbb{C}^{n+1}$ qui à un polynôme $P(X)$ fait correspondre $(P(x_0), P(x_1), ..., P(x_n))$. On voit que le noyau $Ker(T)$ de l'application T est l'espace constitué des multiples du polynôme $N(X) = (X - x_0)(X - x_1)...(X - x_n)$ et qu'on peut écrire

$$\mathbb{C}[X] = \mathbb{C}_n[X] \bigoplus Ker(T).$$

Ceci nous montre que la restriction de T à $\mathbb{C}_n[X]$ est une bijection de $\mathbb{C}_n[X]$ sur $\mathbb{C}^{n+1}$.

Le résultat obtenu est donc le suivant

THÉORÈME 3.1. *Pour tout élément $(y_0, y_1, ..., y_n)$ de $\mathbb{C}^{n+1}$ il existe un polynôme $P(X)$ de degré $\leq n$ et un seul (polynôme d'interpolation de Lagrange) tel que*

$P(x_k) = y_k$ $(0 \leq k \leq n)$. Tout polynôme $Q(X)$ (de degré quelconque) vérifiant aussi $Q(x_k) = y_k$ $(0 \leq k \leq n)$ s'écrit sous la forme

$$Q(X) = P(X) + K(X)N(X).$$

Ecriture sous la forme naturelle (base des monômes). Si on cherche à trouver explicitement le polynôme de Lagrange écrit sous la forme habituelle $P(X) = a_0 + a_1X + ... + a_nX^n$ on est amené à résoudre le système de $n+1$ équations en les $n+1$ inconnues $a_0, ..., a_n$

$$\left\{ \begin{array}{lcl} a_0 + a_1x_0 + ... + a_nx_0^n & = & y_0 \\ a_0 + a_1x_1 + ... + a_nx_1^n & = & y_1 \\ \cdots\cdots\cdots\cdots & \cdots & \cdots \\ a_0 + a_1x_n + ... + a_nx_n^n & = & y_n. \end{array} \right.$$

Ce système est un système de **Vandermonde** dont on sait bien sûr qu'il admet une solution unique. Nous fournirons ultérieurement un algorithme pour résoudre ce système plus rapidement que ne le ferait une méthode classique comme la méthode du pivot par exemple.

Ecriture sous forme de Lagrange. Ainsi le calcul des coefficients a_i du polynôme cherché se ramène à la résolution d'un système de Vandermonde. Sans être très compliqué ceci n'est cependant pas immédiat. Or il peut se faire qu'on n'ait pas absolument besoin des coefficients a_i et qu'on puisse se satisfaire d'exprimer le polynôme d'interpolation dans une base mieux adaptée que la classique base des monômes. Dans cet ordre d'idée, introduisons les $n + 1$ polynômes $L_k(X)$ (où $0 \leq k \leq n$) qui vérifient

$$\left\{ \begin{array}{lcl} L_k(x_k) & = & 1 \\ L_k(x_j) & = & 0 \quad (j \neq k). \end{array} \right.$$

D'après l'étude qui précède le polynôme $L_k(X)$ existe et est unique. On vérifie aisément que $L_k(X)$ s'écrit explicitement sous la forme

$$L_k(X) = \frac{(X - x_0)...(X - x_{k-1})(X - x_{k+1})...(X - x_n)}{(x_k - x_0)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_n)}.$$

Ces polynômes forment une base de $\mathbb{C}_n[X]$ et le polynôme d'interpolation s'exprime dans cette base

$$P(X) = \sum_{k=0}^{n} y_k L_k(X).$$

Il est intéressant de remarquer que le polynôme $L_k(X)$ s'écrit aussi

$$L_k(X) = \frac{N(X)}{(X - x_k)N'(x_k)}$$

où $N(X) = (X - x_0)(X - x_1)...(X - x_n)$.

Ecriture sous la forme de Newton. L'inconvénient des polynômes $L_k(X)$ est que chacun d'eux fait intervenir tous les points d'interpolation et donc si on rajoute un nouveau point tout calcul fait à partir des polynômes $L_k(X)$ doit être entièrement refait.

Prenons alors les polynômes $N_k(X) = (X-x_0)(X-x_1)...(X-x_{k-1})$ où $0 \leq k \leq n$ (avec $N_0 = 1$). Ces polynômes forment aussi une base de $\mathbb{C}_n[X]$ et le polynôme d'interpolation se décompose sur cette base sous la **forme de Newton**

$$P(X) = \sum_{k=0}^{n} b_k N_k(X).$$

Le problème est alors de calculer les coefficients b_k. Pour ce faire définissons les différences divisées successives des valeurs y_i par rapport aux points x_i

$$[y_0] = y_0$$
$$[y_0, y_1, ..., y_k] = \frac{[y_1, ..., y_k] - [y_0, ..., y_{k-1}]}{x_k - x_0}$$

THÉORÈME 3.2. *Les coefficients de la décomposition du polynôme d' interpolation de Lagrange de degré n dans la base de Newton sont donnés par*

$$b_k = [y_0, y_1, ..., y_k]$$

où

$$0 \leq k \leq n.$$

Preuve. La formule à démontrer est clairement vraie si on a $n = 0$. Supposons la formule vraie pour les polynômes de degré $n-1$ interpolant en n points. Soit alors $P_{n-1}(X)$ le polynôme d'interpolation de Lagrange associé aux points $x_0, x_1, ..., x_{n-1}$ et aux valeurs $y_0, y_1, ..., y_{n-1}$, $Q_{n-1}(X)$ le polynôme d'interpolation de Lagrange associé aux points $x_1, x_2, ..., x_n$ et aux valeurs $y_1, y_2, ..., y_n$ et

$$P(X) = \sum_{k=0}^{n} b_k N_k(X)$$

le polynôme d'interpolation de Lagrange associé aux points $x_0, x_1, ..., x_n$ et aux valeurs $y_0, y_1, ..., y_n$. Il est facile de voir que

$$P_{n-1}(X) = \sum_{k=0}^{n-1} b_k N_k(X)$$

si bien qu'il reste simplement en vertu de l'hypothèse de récurrence à établir la formule pour le coefficient b_n.

Pour cela définissons

$$\tilde{P}_n(X) = \frac{(X - x_0)Q_{n-1}(X) - (X - x_n)P_{n-1}(X)}{x_n - x_0}.$$

On vérifie que

$$\tilde{P}_n(x_i) = y_i$$

pour $0 \leq i \leq n$. Donc

$$\tilde{P}_n(X) = P(X).$$

En égalant les coefficients du terme de degré n dans l'expression des polynômes P_n et $\tilde{P}_n$ on obtient la relation cherchée.

En comparant l'expression de Newton du polynôme P_k d'interpolation en $k+1$ points avec celle de Lagrange on trouve en regardant les coefficients des termes de degré k

$$[y_0, y_1, ..., y_k] = \sum_{j=0}^{k} \frac{y_j}{N'_{k+1}(x_j)}.$$

Il est clair dans cette représentation que le rajout d'un nouveau point ne fait que rajouter un nouveau terme au polynôme, les autres termes restant identiques.

3.2.2. L'aspect approximation.

Le théorème de division des fonctions différentiables. Soit f une fonction réelle définie sur $\mathbb{R}$ de classe C^{p+1}, où p est un entier naturel. On suppose que f s'annule en un point a de $\mathbb{R}$. Posons

$$g(x) = \int_0^1 f'(a + (x-a)u)du$$

alors $g(x)$ est l'unique fonction continue telle que

$$f(x) = (x-a)g(x).$$

De plus d'après le théorème de dérivation sous le signe intégrale on voit que la fonction $g(x)$ est de classe C^p et que pour tout $0 \leq q \leq p$

$$g^{(q)}(x) = \int_0^1 u^q f^{(q+1)}(a + (x-a)u)du$$

et par suite

$$|g^{(q)}(x)| \leq \frac{1}{q+1} sup_{t\in[a,x]} |f^{(q+1)}(t)|.$$

Majoration de l'erreur. Soit f une fonction de classe C^{n+1}, P le polynôme d'interpolation de Lagrange qui prend les mêmes valeurs que f aux point $x_0, x_1, ..., x_n$ et I un intervalle compact contenant $x, x_0, x_1, ..., x_n$. Appliquons alors le théorème précédent à $f(x) - P(x)$. On obtient

$$f(x) - P(x) = (x - x_0)g_0(x)$$

avec

$$|g_0^{(n)}(x)| \leq \frac{1}{n+1} sup_{t\in I} |f^{(n+1)}(t)|$$

(ne pas oublier que $P^{(n+1)}(x) = 0$)

puis

$$g_0(x) = (x - x_1)g_1(x)$$

avec

$$|g_1^{(n-1)}(x)| \leq \frac{1}{n} sup_{t\in I} |g_0^{(n)}(t)|$$

et ainsi de suite. Si bien que

$$|f(x) - P(x)| \leq \frac{1}{(n+1)!} |(x-x_0)(x-x_1)...(x-x_n)| sup_{t\in I} |f^{(n+1)}(t)|$$

3.2.3. L'aspect algorithmique.

Le calcul des différences divisées. L'algorithme des différences divisées est très simple. Il utilise l'écriture du polynôme d'interpolation sous la forme de Newton. Les coefficients sont alors calculés par la formule de récurrence établie précédemment

$$[y_0] = y_0$$
$$[y_0, y_1, ..., y_k] = \frac{[y_1, ..., y_k] - [y_0, ..., y_{k-1}]}{x_k - x_0}.$$

si bien que le calcul se fait conformément à la figure 1.
Le nombre d'opérations à effectuer est en $O(n^2)$.

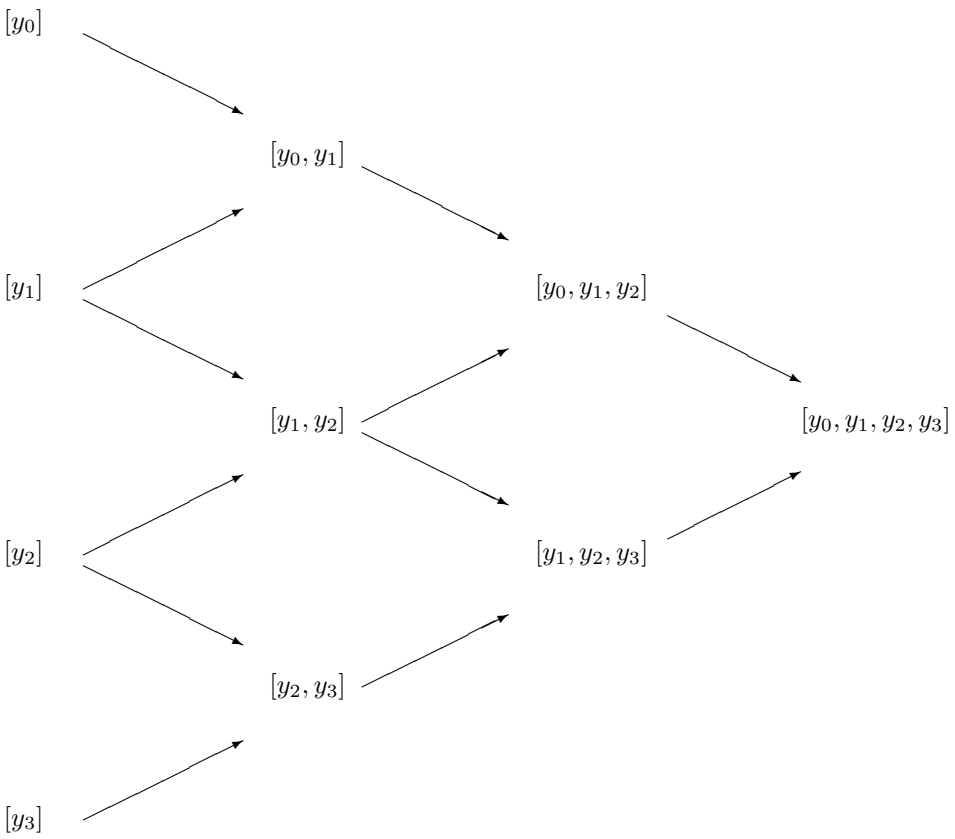

FIG. 1. Le calcul des différences divisées

Résolution d'un système de Vandermonde. Nous avons vu que le calcul effectif des coefficients du polynôme d'interpolation de Lagrange dans la base naturelle des monômes passe par la résolution d'un système de Vandermonde. Voici un algorithme qui permet de résoudre un tel système. Cet algorithme est basé en fait sur l'algorithme de Hörner pour l'évaluation de polynômes. Il est plus rapide que les algorithmes directs de résolution des systèmes linéaires généraux comme la méthode du pivot de Gauss ou la méthode de Householder qui sont en $O(n^3)$ alors que nous obtenons ici un algorithme en $O(n^2)$.

Soit N un entier ≥ 2. Etant donné $x = (x_1, x_2, ..., x_N)$ un N-uplet de réels deux à deux distincts on note B la matrice

$$B = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \\ x_1^2 & x_2^2 & \cdots & x_N^2 \\ \vdots & \vdots & & \vdots \\ x_1^{N-1} & x_2^{N-1} & \cdots & x_N^{N-1} \end{pmatrix}$$

On cherche à résoudre le système

$$BW = Q$$

où Q est la matrice colonne constituée des seconds membres $q_1, q_2, \cdots, q_N$ du système et où W est la matrice colonne constituée des inconnues $w_1, w_2, \cdots, w_N$ du système.

Pour tout entier j vérifiant $1 \leq j \leq N$ on pose

$$P_j(x) = \prod_{\substack{n=1 \\ n \neq j}}^{N} \frac{x - x_n}{x_j - x_n}$$

P_j est donc un polynôme en x de degré $N-1$ qui peut s'écrire

$$P_j(x) = \sum_{k=1}^{N} A_{j,k} x^{k-1}$$

En effectuant le produit de la matrice $A = (A_{j,k})_{j,k}$ par la matrice B on constate que

$$AB = (P_j(x_k))_{j,k}$$

ce qui prouve que A est l'inverse de B.

On peut alors écrire que W=AQ. On obtient ainsi les formules

$$w_j = \sum_{k=1}^{N} A_{j,k} q_k.$$

Nous allons dans la suite mettre en place une méthode de résolution qui calcule les coefficients des polynômes P_j, donc qui calcule l'inverse de la matrice B.Pour calculer les coefficients de P_j on sera amené à calculer les coefficients de

$$N_j(x) = \prod_{\substack{n=1 \\ n \neq j}}^{N} (x - x_n)$$

et aussi le dénominateur intervenant dans la formule qui définit P_j, c'est à dire le nombre $N_j(x_j)$.

Posons

$$P(x) = (x - x_1)(x - x_2)...(x - x_N)$$

$P(x)$ est donc un polynôme de degré N qui s'écrit sous la forme :

$$P(x) = x^N + c_N x^{N-1} + ... + c_2 x + c_1$$

Montrons tout d'abord comment si on connaît les coefficients c_j on peut calculer les coefficients du polynôme

$$N_j(x) = \prod_{\substack{n=1 \\ n \neq j}}^{N} (x - x_n)$$

Pour cela posons

$$N_j(x) = b_N x^{N-1} + ... + b_2 x + b_1$$

On vérifie immédiatement sur l'expression de $N_j(x)$ que le coefficient du terme de plus haut degré est 1. En remarquant que $P(x) = N_j(x)(x - x_j)$ on établit la formule

$$b_{k-1} = c_k + x_j b_k.$$

Si bien que

$$\begin{cases} b_N = 1 \\ b_{k-1} = c_k + x_j b_k \end{cases}$$

Connaissant les coefficients de N_j il est alors facile de calculer le dénominateur $N_j(x_j)$ intervenant dans la définition de P_j.

En effet posons $t_N = b_N = 1$ et définissons pour tout $k \leq N$

$$t_{k-1} = x_j t_k + b_{k-1}$$

On constate alors que $t_1 = N_j(x_j)$, le calcul proposé pour $N_j(x_j)$ n'étant rien d' autre que l'algorithme de Horner.

Il reste maintenant à calculer les coefficients c_j de P.

Pour tout entier k vérifiant $1 \leq k \leq N$ on définit

$$Q_k(x) = (x - x_1)(x - x_2)...(x - x_k)$$

et on écrit Q_k sous la forme

$$Q_k(x) = x^k + \alpha_{k,k} x^{k-1} + \alpha_{k,k-1} x^{k-2} + ... + \alpha_{k,1}$$

Il est facile de voir sur l'expression de $Q_1(x) = x - x_1$ que $\alpha_{1,1} = -x_1$.

De la formule

$$Q_k(x) = Q_{k-1}(x)(x - x_k).$$

découlent pour $k = 2, 3, ..., N$ les formules

$$\alpha_{k,k} = \alpha_{k-1,k-1} - x_k$$

$$\alpha_{k,j} = \alpha_{k-1,j-1} - x_k \alpha_{k-1,j} \quad j = k-1, ..., 2$$

ce qui achève l'algorithme.

On peut voir que le nombre d'opérations à faire dans cet algorithme est en $O(N^2)$, la partie la plus coûteuse étant le calcul des coefficients c_k.

3.2.4. Un cas particulier : les points d'interpolation sont les racines n^{e}de l'unité.

Interpolation de Lagrange et transformée de Fourier discrète. Rappelons que si $a = (a_0, a_1, ..., a_{n-1})$ est une suite finie de n nombres complexes on définit la transformée de Fourier $\widehat{a}$ de la suite a comme étant la suite $\widehat{a} = (\widehat{a}_0, \widehat{a}_1, ..., \widehat{a}_{n-1})$ où

$$\widehat{a}_v = \frac{1}{n}\sum_{u=0}^{n-1} a_u e^{-\frac{2i\pi uv}{n}}.$$

La transformation inverse se calcule facilement par

$$a_u = \sum_{v=0}^{n-1} \widehat{a}_v e^{\frac{2i\pi uv}{n}}.$$

On remarque que

$$\widehat{\widehat{a}}_u = \frac{1}{n} a_{n-u}.$$

A un coefficient près le même algorithme permettra de calculer la transformée de Fourier et son inverse.

Notons

$$P_{\widehat{a}}(X) = \left(\widehat{a}_0 + \widehat{a}_1 X + ... + \widehat{a}_{n-1} X^{n-1}\right)$$

On voit alors que

$$a_u = P_{\widehat{a}}(e^{\frac{2i\pi u}{n}}).$$

Cette dernière remarque met en évidence un aspect très important de la transformée de Fourier discrète : l'aspect interpolation . En effet il est facile de trouver grâce à ce que nous avons vu le polynôme d'interpolation de Lagrange qui prend les valeurs a_u aux points $x_u = e^{\frac{2i\pi u}{n}}$; C'est le polynôme $P_{\widehat{a}}(X)$ dont les coefficients sont donnés par la transformée de Fourier discrète de la suite $a = (a_0, a_1, ..., a_{n-1})$ des valeurs prises aux points d'interpolation .

Le calcul explicite : transformée de Fourier rapide. Il existe divers façons proches les une des autres de calculer une transformée de Fourier discrète. Toutes ces variantes sont des algorithmes de transformée de Fourier rapides (FFT). Nous nous placerons ici dans le cas où le nombre d'éléments de la suite à transformer est $n = 2^m$. Pour tout $r > 0$ et tout $0 \leq k \leq 2^r - 1$ posons

$$W_{2^r}^k = e^{-\frac{2ik\pi}{2^r}}.$$

Remarquons que

$$W_{2^r}^k = (W_{2^{r+1}}^k)^2 = (W_{2^{r+1}}^{k+2^r})^2$$
$$W_{2^{r+1}}^k = -W_{2^{r+1}}^{k+2^r}$$

par exemple

$$(W_8^3)^2 = (W_8^7)^2 = W_4^3$$
$$W_8^3 = -W_8^7.$$

On rappelle que si

$$a = (a_0, a_1, ..., a_{2^m-1})$$

et si

$$P_a(X) = \frac{1}{2^m}\left(a_0 + a_1 X + ... + a_{2^m-1} X^{2^m-1}\right)$$

alors

$$\widehat{a}_u = P_a(W_{2^m}^u).$$

Pour tout polynôme

$$P(X) = p_0 + p_1X + ... + p_{2^r-1}X^{2^r-1}$$

notons

$$P_0(X) = p_0 + p_2X + ... + p_{2^r-2}X^{2^{r-1}-1}$$

et

$$P_1(X) = p_1 + p_3X + ... + p_{2^r-1}X^{2^{r-1}-1}$$

alors

$$P(X) = P_0(X^2) + XP_1(X^2)$$

ce qui donne si $0 \leq k \leq 2^{r-1} - 1$

$$P(W_{2^r}^k) = P_0(W_{2^{r-1}}^k) + W_{2^r}^k P_1(W_{2^{r-1}}^k)$$

et

$$P(W_{2^r}^{k+2^{r-1}}) = P_0(W_{2^{r-1}}^k) - W_{2^r}^k P_1(W_{2^{r-1}}^k).$$

Ces dernières formules vont nous donner un algorithme pour calculer les valeurs de la transformée de Fourier.

Remarquons tout d'abord que si on a tabulé les valeurs de $W_{2^m}^k$ alors on dispose aussi des valeurs de $W_{2^r}^k$ pour tout $r \leq m$.

W_8^0	W_8^1	W_8^2	W_8^3	W_8^4 $-W_8^0$	W_8^5 $-W_8^1$	W_8^6 $-W_8^2$	W_8^7 $-W_8^3$
W_4^0		W_4^1		W_4^2 $-W_4^0$		W_4^3 $-W_4^1$	
W_2^0				W_2^1 $-W_2^0$			

Pratique du calcul. Le coefficient $\frac{1}{n}$ n'interviendra qu'à la fin. Pour cela au lieu de calculer avec le polynôme P_a nous calculerons avec $P = nP_a = a_0 + ... + a_{2^m-1}X^{2^m-1}$.

L'exemple $m = 3$ est suffisamment instructif pour décrire l'algorithme. Remarquons que

$$P_{000}(X) = a_0, P_{001}(X) = a_4, P_{010}(X) = a_2, P_{011}(X) = a_6$$
$$P_{100}(X) = a_1, P_{101}(X) = a_5, P_{110}(X) = a_3, P_{111}(X) = a_7.$$

On commence donc à faire une permutation σ des éléments

$$a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7$$

pour les mettre dans l'ordre

$$a_0, a_4, a_2, a_6, a_5, a_3, a_7.$$

Ceci se fait facilement en remarquant qu'à chaque indice supposé écrit en binaire on fait correspondre l'indice obtenu en écrivant les bits dans l'ordre inverse. Ainsi l'indice $4 = 100$ est transformé en $1 = 001$. la suite du calcul de la transformée de Fourier se fait en trois étapes indiquées par la figure 2 et à la fin on divise les coefficient obtenus par 8.

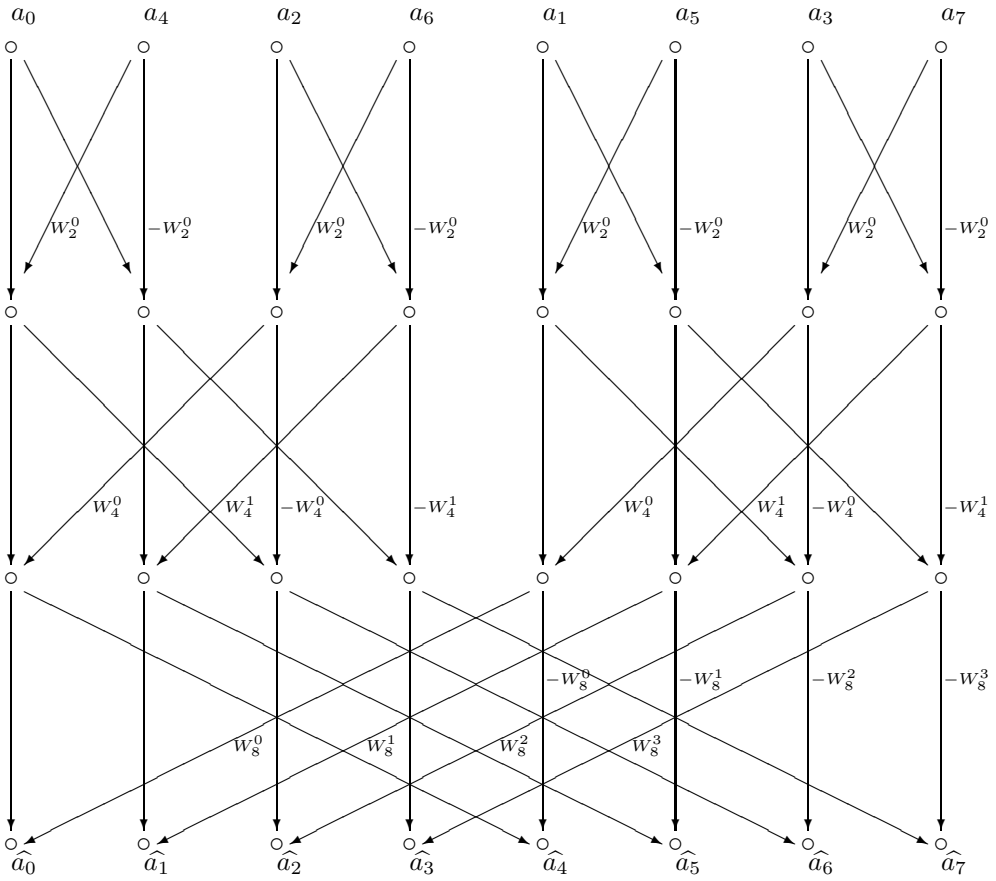

FIG. 2. La FFT sur 8 points

Appelons M_8^1, M_8^2, M_8^3 les matrices

$$M_8^1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$M_8^2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}$$

$$M_8^3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

S_1, S_2, S_3 les matrices diagonales définies par

$$S_1 = Diag(1, W_2^0, 1, W_2^0, 1, W_2^0, 1, W_2^0)$$
$$S_2 = Diag(1, 1, W_4^0, W_4^1, 1, 1, W_4^0, W_4^1)$$
$$S_3 = Diag(1, 1, 1, 1, W_8^0, W_8^1, W_8^2, W_8^3)$$

et enfin Σ la matrice de la permutation "reverse bit" σ.

Dans ces conditions la matrice F_8 de la transformation de Fourier sur 8 points s'écrit

$$F_8 = \frac{1}{8} M_8^3 S_3 M_8^2 S_2 M_8^1 S_1 \Sigma.$$

Ceci se généralise facilement pour $n = 2^m$. Le nombre d'opérations à effectuer pour calculer cette transformation est de l'ordre de $nlog(n)$.

3.3. Le problème général de l'interpolation

Interprétons de manière un peu plus algébrique le problème d'interpolation de Lagrange. En particulier notons g_i la forme linéaire sur $\mathbb{C}_n[X]$ qui à tout polynôme $Q(X)$ fait correspondre $Q(x_i)$. On cherche alors un élément $P(X)$ de $\mathbb{C}_n[X]$ qui réalise $g_i(P) = y_i$ pour tout i. Remarquons que les g_i forment une base du dual de $\mathbb{C}_n[X]$ et que cette base est la base duale de la base constituée par les L_i.

Nous poserons le problème général de l'interpolation en ces termes :

Soit E un espace vectoriel de dimension n, E^* le dual de E, $(g_i)_i$ une base de E^* et $(y_i)_i$ des nombres. Trouver un élément P de E tel que $g_i(P) = y_i$ pour tout $1 \leq i \leq n$.

Il est clair que si $(e_i)_i$ est la base de E dont $(g_i)_i$ est la base duale, alors

$$P = \sum_{i=1}^{n} y_i e_i.$$

Nous allons voir par la suite quelques exemples qui entrent dans ce cadre général.

3.4. Quelques exemples importants

3.4.1. Interpolation d'Hermite. Soient $x_0 < x_1$ et y_0, y_1, y_0', y_1' des nombres réels. On cherche un polynôme P de degré ≤ 3 tel que

$$\begin{cases} P(x_0) &= y_0 \\ P(x_1) &= y_1 \\ P'(x_0) &= y_0' \\ P'(x_1) &= y_1' \end{cases}$$

Notons E l'espace vectoriel de dimension 4 des polynômes de degré ≤ 3. Définissons les 4 formes linéaires sur E

$$\left\{\begin{array}{lcl} \delta_{0,x_0}(P) & = & P(x_0) \\ \delta_{0,x_1}(P) & = & P(x_1) \\ \delta_{1,x_0}(P) & = & P'(x_0) \\ \delta_{1,x_1}(P) & = & P'(x_1) \end{array}\right.$$

La question posée est donc de résoudre le problème linéaire suivant : trouver P tel que

$$\left\{\begin{array}{lcl} \delta_{0,x_0}(P) & = & y_0 \\ \delta_{0,x_1}(P) & = & y_1 \\ \delta_{1,x_0}(P) & = & y'_0 \\ \delta_{1,x_1}(P) & = & y'_1 \end{array}\right.$$

Pour cela cherchons si on peut trouver 4 polynômes $H_{0,x_0}, H_{0,x_1}, H_{1,x_0}, H_{1,x_1}$ tels que

$$\delta_{i,x_j}(H_{k,x_l}) = \left\{\begin{array}{l} 1 \ si \ i = k \ et \ j = l \\ 0 \ sinon \end{array}\right. .$$

Si on arrive à trouver ces polynômes cela prouvera à la fois que ces polynômes forment une base de E, que les formes linéaires introduites forment une base de E^*, qui est la base duale de la base polynômiale trouvée.

Un simple calcul nous permet de trouver effectivement ces polynômes (ce sont en fait des solutions dans des cas particuliers bien choisis du problème posé) ;

$$H_{0,x_0}(x) = \frac{(x-x_1)^2(3x_0 - x_1 - 2x)}{(x_0 - x_1)^3}$$

$$H_{0,x_1}(x) = \frac{(x-x_0)^2(3x_1 - x_0 - 2x)}{(x_1 - x_0)^3}$$

$$H_{1,x_0}(x) = \frac{(x-x_1)^2(x-x_0)}{(x_0 - x_1)^2}$$

$$H_{1,x_1}(x) = \frac{(x-x_0)^2(x-x_1)}{(x_0 - x_1)^2}.$$

Il ressort de toutes ces considérations que le problème a une solution unique (polynôme d'interpolation de Hermite) donnée par

$$P(x) = y_0 H_{0,x_0}(x) + y_1 H_{0,x_1}(x) + y'_0 H_{1,x_0}(x) + y'_1 H_{1,x_1}(x).$$

Si nous sommes partis d'une fonction f de classe $\mathcal{C}^4$ sur un intervalle compact I contenant les points x_0, x_1, et si nous appelons P le polynôme d'interpolation de Hermite associé aux points x_0, x_1 et aux valeurs $f(x_0), f(x_1), f'(x_0), f'(x_1)$, alors comme dans l'exemple de l'interpolation de Lagrange, l'application répétée du théorème de division des fonctions différentiables nous donne l'approximation

$$|f(x) - P(x)| \leq \frac{1}{4!}(x-x_0)^2(x-x_1)^2 \sup_{t \in I} |f^{(4)}(t)|.$$

3.4.2. Interpolation par les splines cubiques. Soit Δ un partage d'un segment $[a, b]$,

$$a = x_1 < x_2 < \cdots < x_n = b.$$

DÉFINITION 3.3. Une fonction f définie sur $[a, b]$ est une fonction spline cubique relativement au partage Δ si les conditions suivantes sont satisfaites :

1) f est de classe $\mathcal{C}^2[a, b]$;

2) f coïncide avec un polynôme de degré 3 sur chaque intervalle $[x_j, x_{j+1}]$;

Nous noterons $\mathcal{S}_\Delta$ l'ensemble de ces fonctions.

Ainsi, $\mathcal{S}_\Delta$ est un sous-espace vectoriel de l'espace des fonctions définies sur $[a, b]$.

Soit $Y = (y_1, y_2, \cdots, y_n)$ une suite de n nombres réels. Nous noterons $\mathcal{S}_{\Delta,Y}$ l'ensemble des fonctions splines cubiques qui vérifient la condition :

3) $f(x_j) = y_j$ pour $1 \leq j \leq n$.

REMARQUE 3.4. Soit Y_0 le n-uple particulier :

$$Y_0 = (0, 0, \cdots, 0).$$

On voit tout de suite que $\mathcal{S}_{\Delta,Y_0}$ est un sous-espace vectoriel de $\mathcal{S}_\Delta$.

Nous nous proposons de déterminer des fonctions splines cubiques répondant à certaines conditions supplémentaires.

Soit donc $f \in \mathcal{S}_{\Delta,Y}$ et posons $M_j = f''(x_j)$. Sur l'intervalle $[x_j, x_{j+1}]$ la dérivée seconde de f est de degré 1 et on a sur cet intervalle

$$f''(x) = M_j \frac{x_{j+1} - x}{x_{j+1} - x_j} + M_{j+1} \frac{x - x_j}{x_{j+1} - x_j}.$$

En intégrant deux fois,

$$f(x) = \frac{M_j}{6} \frac{(x_{j+1} - x)^3}{(x_{j+1} - x_j)} + \frac{M_{j+1}}{6} \frac{(x - x_j)^3}{(x_{j+1} - x_j)} + Q(x)$$

où $Q(x)$ est un polynôme de degré 1 que l'on peut présenter sous la forme

$$Q(x) = \alpha(x_{j+1} - x) + \beta(x - x_j).$$

Evaluons α et β en utilisant $f(x_j) = y_j$ et $f(x_{j+1}) = y_{j+1}$. On trouve

$$\alpha = \left(y_j - M_j \frac{(x_{j+1} - x_j)^2}{6} \right) \frac{1}{(x_{j+1} - x_j)}$$

$$\beta = \left(y_{j+1} - M_{j+1} \frac{(x_{j+1} - x_j)^2}{6} \right) \frac{1}{(x_{j+1} - x_j)},$$

si bien que si on pose

$$h_{j+1} = x_{j+1} - x_j$$

on obtient pour $x \in [x_j, x_{j+1}]$,

$$f(x) = \frac{M_j}{6} \frac{(x_{j+1} - x)^3}{h_{j+1}} + \frac{M_{j+1}}{6} \frac{(x - x_j)^3}{h_{j+1}} +$$

$$\left(y_j - M_j \frac{h_{j+1}^2}{6} \right) \frac{(x_{j+1} - x)}{h_{j+1}} + \left(y_{j+1} - M_{j+1} \frac{h_{j+1}^2}{6} \right) \frac{(x - xj)}{h_{j+1}}.$$

Nous allons exploiter maintenant les conditions sur la dérivée première. Pour $2 \leq j \leq n-1$ on a

$$f'(x_j^+) = -M_j \frac{h_{j+1}}{3} - M_{j+1}\frac{h_{j+1}}{6} + \frac{y_{j+1}-y_j}{h_{j+1}}$$

$$f'(x_j^-) = M_{j-1}\frac{h_j}{6} + M_j \frac{h_j}{3} + \frac{y_j - y_{j-1}}{h_j}$$

si bien que

$$\frac{h_j}{6}M_{j-1} + \frac{h_j + h_{j+1}}{3}M_j + \frac{h_{j+1}}{6}M_{j+1} = \frac{y_{j+1}-y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}.$$

On dispose donc de $n-2$ équations linéaires pour déterminer les n inconnues $M_1, M_2, \cdots, M_n$. Il convient donc si on espère avoir une solution unique de donner deux conditions supplémentaires. Pour la suite du calcul nous imposerons donc les valeurs de la dérivée de f en x_1 et en x_n, c'est-à-dire

$$f'(x_1) = y'_1$$
$$f'(x_n) = y'_n.$$

On a alors

$$2M_1 + M_2 = \frac{6}{h_2}\left(\frac{y_2 - y_1}{h_2} - y'_1\right)$$

et

$$M_{n-1} + 2M_n = \frac{6}{h_n}\left(y'_n - \frac{y_n - y_{n-1}}{h_n}\right).$$

Si on pose

$$\begin{cases} \lambda_1 &= 1 \\ \lambda_n &= 0 \\ \lambda_j &= \frac{h_{j+1}}{h_j + h_{j+1}} & 2 \leq j \leq n-1 \\ \mu_j &= 1 - \lambda_j & 1 \leq j \leq n \end{cases}$$

et

$$\begin{cases} b_1 &= \frac{6}{h_2}\left(\frac{y_2 - y_1}{h_2} - y'_1\right) \\ b_n &= \frac{6}{h_n}\left(y'_n - \frac{y_n - y_{n-1}}{h_n}\right) \\ b_j &= \frac{6}{h_j + h_{j+1}}\left(\frac{y_{j+1}-y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}\right) & 2 \leq j \leq n-1 \end{cases}$$

alors le système à résoudre est

$$\begin{bmatrix} 2 & \lambda_1 & 0 & 0 & \cdots & \cdots & 0 \\ \mu_2 & 2 & \lambda_2 & 0 & \cdots & \cdots & 0 \\ 0 & \mu_3 & 2 & \lambda_3 & & & \vdots \\ 0 & 0 & \mu_4 & 2 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \lambda_{n-2} & 0 \\ \vdots & \vdots & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ 0 & 0 & \cdots & \cdots & 0 & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ M_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{bmatrix}.$$

Ce système est tridiagonal, à diagonale strictement dominante ; il possède une solution et une seule. On peut employer pour un tel système une simplification de la méthode du pivot, qui dans ce cas particulier s'exécute en temps linéaire. Ainsi on peut énoncer le théorème suivant :

THÉORÈME 3.5. *Soient a' et b' deux nombres réels. Il existe une unique spline cubique $f \in \mathcal{S}_{\Delta,Y}$ telle que $f'(a) = a'$ et $f'(b) = b'$.*

REMARQUE 3.6. En particulier la seule fonction spline cubique qui s'annule sur tous les nœuds du partage Δ et dont la dérivée est nulle en a et en b est la fonction nulle.

Les fonctions splines cubiques minimisent approximativement l'énergie de flexion d'une tige. En effet si la fonction $g(x)$ représente l'équation d'une tige parfaitement élastique son énergie de flexion est

$$\int_a^b \frac{\Big(g''(t)\Big)^2 dt}{\Big(1 + \Big(g'(t)\Big)^2\Big)^{5/2}}$$

et donc si la dérivée de $g(x)$ est petite on peut approcher cette énergie de flexion par

$$\int_a^b \Big(g''(t)\Big)^2 dt.$$

Nous allons voir que les fonctions splines cubiques minimisent cette dernière expression. Sur $\mathcal{C}^2[a,b]$ on définit un semi-produit scalaire par

$$< f, g >= \int_a^b f''(t)g''(t)dt$$

qui donne la semi-norme

$$|f| = \left(\int_a^b |f''(t)|^2 dt\right)^{1/2}.$$

Soit $\mathcal{N}_\Delta$ le sous-espace de $\mathcal{C}^2[a,b]$ constitué des fonctions nulles aux points x_j et telles que $f'(a) = f'(b) = 0$.

LEMME 3.7. *Les espaces $\mathcal{N}_\Delta$ et $\mathcal{S}_\Delta$ sont orthogonaux et d'intersection réduite à $\{0\}$.*

Preuve. Soit $f \in \mathcal{N}_\Delta$ et $g \in \mathcal{S}_\Delta$. Calculons :

$$< f, g >= \int_a^b f''(t)g''(t)dt.$$

Pour cela calculons pour chaque intervalle $[x_j, x_{j+1}]$ du partage :

$$I_j = \int_{x_j}^{x_{j+1}} f''(t)g''(t)dt.$$

Par intégration par partie on obtient :

$$I_j = [f'(t)g''(t)]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} f'(t)g^{(3)}(t)dt,$$

ce qui donne en intégrant de nouveau par partie le reste et en tenant compte du fait que la dérivée quatrième de g est nulle et que f vaut 0 aux points x_j et x_{j+1} :

$$I_j = [f'(t)g''(t)]_{x_j}^{x_{j+1}} .$$

On en conclut que :

$$\int_a^b f''(t)g''(t)dt = \sum_{j=1}^{n-1} I_j = [f'(t)g''(t)]_a^b ,$$

et puisque $f'(t)$ s'annule en a et b on en conclut que $< f, g >= 0$. Le fait que l'intersection est réduite à la fonction nulle est une conséquence de la remarque 3.6. □

LEMME 3.8. *Toute fonction $f \in \mathcal{C}^2[a,b]$ s'écrit d'une façon unique sous la forme $f = s + h$ où $s \in \mathcal{S}_\Delta$ et $h \in \mathcal{N}_\Delta$.*

Preuve. Posons $Y = (f(x_1), f(x_2), \cdots, f(x_n))$. Comme h doit appartenir à $\mathcal{N}_\Delta$, s'il existe une spline cubique répondant à la question alors nécessairement $s \in \mathcal{S}_{\Delta,Y}$, $s'(a) = f'(a)$ et $s'(b) = f'(b)$. On sait qu'il existe une unique spline cubique répondant à la question. La fonction h est alors uniquement définie par $h = f - s$, et elle appartient bien à $\mathcal{N}_\Delta$. □

THÉORÈME 3.9. *Parmi toutes les fonctions $f \in \mathcal{C}^2[a,b]$ qui vérifient $f(x_j) = y_j$ et les conditions $f'(a) = y'_1$ et $f'(b) = y'_n$ la fonction qui minimise l'énergie de flexion est la spline cubique (l'élément de $\mathcal{S}_{\Delta,Y}$) qui vérifie en outre $f'(a) = y'_1$ et $f'(b) = y'_n$.*

Preuve. Soit $f \in \mathcal{C}^2[a,b]$. Écrivons d'après le lemme précédent $f = s+h$, où $s \in \mathcal{S}_\Delta$ et $h \in \mathcal{N}_\Delta$. On a en vertu de l'orthogonalité de s et h :

$$< f, f >=< s, s > + < h, h > .$$

Donc :

$$< s, s >=< f, f > - < h, h >\leq < f, f > .$$

□

On pourrait aussi montrer le résultat suivant

THÉORÈME 3.10. *Parmi toutes les fonctions $f \in \mathcal{C}^2[a,b]$ qui vérifient $f(x_j) = y_j$ la fonction qui minimise l'énergie de flexion est la spline cubique (l'élément de $\mathcal{S}_{\Delta,Y}$) qui vérifie en outre $f''(a) = f''(b) = 0$.*

3.4.3. Interpolation par des polynômes trigonométriques. Etant donnés $2n+1$ nombres réels distincts

$$-\pi \leq x_1 < x_2 < \cdots < x_{2n+1} < \pi$$

et $2n+1$ nombres complexes $y_1, \cdots, y_{2n+1}$ on cherche un polynôme trigonométrique

$$P(x) = \sum_{p=0}^{n} \Big(a_p cos(px) + b_p sin(px) \Big)$$

tel que $P(x_i) = y_i$.

On pourrait étudier directement ce problème, mais ici nous allons plutôt le relier au problème de l'interpolation de Lagrange que nous avons déjà étudié.

Pour cela posons pour tout $-n \leq k \leq n$

$$c_k = \frac{1}{2}\Big(a_{|k|} - i\text{Signe}(k)b_{|k|}\Big)$$

alors

$$P(x) = \sum_{k=-n}^{n} c_k e^{ikx}$$

et aussi

$$e^{inx}P(x) = \sum_{k=0}^{2n} c_{k-n}e^{ikx}.$$

En posant $X = e^{ix}$ on obtient

$$e^{inx}P(x) = \sum_{k=0}^{2n} c_{k-n}X^k,$$

ce qui ramène notre problème à un problème d'interpolation de Lagrange. Remarquons que dès que les c_k sont connus, on calcule facilement les a_p et les b_p grâce aux formules

$$a_p = c_p + c_{-p}$$
$$b_p = i(c_p - c_{-p}).$$

CHAPITRE 4

Calcul numérique des intégrales définies

4.1. Introduction

Nous voulons calculer numériquement les intégrales définies

$$\int_a^b f(x)dx$$

où f est une fonction suffisamment régulière pour assurer l'existence des dérivées et des bornes supérieures dont nous aurons besoin.

Les méthodes que nous décrivons **dans un premier temps** s'appuient sur la démarche suivante :

- On découpe l'intervalle d'intégration en N morceaux de même longueur
$$a = x_0 < x_1 < \cdots < x_N = b$$
où $x_{i+1} - x_i = \frac{b-a}{N}$.
- Sur chaque morceau on calcule la valeur approchée de l'intégrale
$$\int_{x_i}^{x_{i+1}} f(x)dx$$
en remplaçant sur l'intervalle $[x_i, x_{i+1}]$ la fonction f par une fonction polynômiale qui interpole f.

Les diverses méthodes qui entrent dans ce cadre diffèrent par le degré des polynômes d'interpolation utilisés et par le choix des points des segments $[x_i, x_{i+1}]$ en lesquels on interpole f. En particulier pour ce dernier point, nous verrons comment choisir au mieux les points d'interpolation. Il faut éviter de confondre les points x_i du partage en N morceaux du segment initial, avec les points des partages des intervalles $[x_i, x_{i+1}]$, que nous serons amenés à introduire pour appliquer sur ces intervalles des méthodes interpolatoires.

Comme nous le verrons ces méthodes peuvent être éventuellement complétées par une méthode d'accélération de convergence. Enfin nous donnerons un abord purement analytique d'une classe de formules de calcul approchée d'intégrales.

4.2. Mise en œuvre de méthodes interpolatoires

Dans toute cette section pour la clarté des calculs et des méthodes, nous travaillerons sur l'intervalle $[-1, 1]$ au lieu de $[x_i, x_{i+1}]$, puis donnerons les formules qu'on obtient de la même façon sur les intervalles $[x_i, x_{i+1}]$, et enfin nous sommerons ces diverses formules pour obtenir l'approximation de l'intégrale sur $[a, b]$. Dans chaque cas nous noterons e l'erreur de la méthode employée sur $[-1, 1]$, e_i l'erreur de la méthode employée sur $[x_i, x_{i+1}]$ et enfin E l'erreur globale introduite sur le segment $[a, b]$.

Signalons enfin que **nous supposerons la fonction f à intégrer sur le segment $[a,b]$ continument dérivable autant de fois qu'il le faut** pour pouvoir appliquer les théorèmes permettant l'établissement des majorations des erreurs. En particulier on aura à appliquer souvant les théorèmes de majoration de la différence d'une fonction et d'un de ses polynômes interpolateurs.

Nous noterons

$$m^{(p)}(f) = \sup_{t\in[-1,1]} \left|f^{(p)}(t)\right|,$$
$$m_i^{(p)}(f) = \sup_{t\in[x_i,x_{i+1}]} \left|f^{(p)}(t)\right|,$$
$$M^{(p)}(f) = \sup_{t\in[a,b]} \left|f^{(p)}(t)\right|.$$

4.2.1. Méthode des rectangles. La méthode que nous présentons mérite à peine le nom de méthode numérique, car elle ne donne pas un calcul efficace et n'est pas employée à cet usage. Cependant elle est importante pour la clarté de l'exposé, car elle initialise en quelque sorte un processus qui va nous conduire à des méthodes plus efficaces. Grâce à elle nous pourrons présenter et illustrer un certain nombre de problèmes de fond.

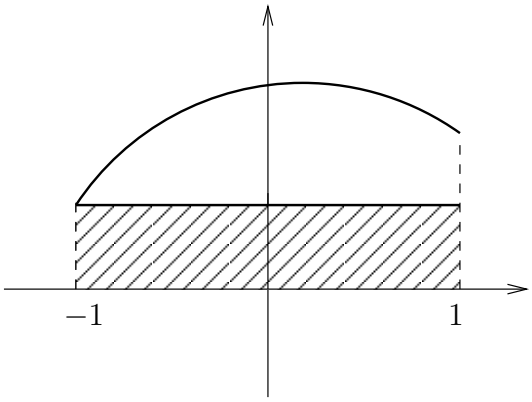

FIG. 1. Méthode des rectangles à gauche

La méthode des rectangles, inspirée de la définition même de l'intégrale de Riemann consiste à supposer la fonction constante sur tout l'intervalle $[-1,1]$, la valeur de la constante étant celle prise en un point déterminé par la fonction. Nous prendrons ici la valeur de la fonction au point -1 (méthode des rectangles à gauche). Remarquons qu'il s'agit bien d'une méthode d'interpolation de Lagrange en un point par un polynôme de degré zéro. Ainsi

$$\int_{-1}^{1} f(t)dt = \int_{-1}^{1} f(-1)dt + e = 2f(-1) + e.$$

En écrivant

$$f(u) = f(-1) + \int_{-1}^{u} f'(t)dt,$$

on obtient

$$\left|\int_{-1}^{1} f(u)du - 2f(-1)\right| \leq m^{(1)}(f)\int_{-1}^{1}(u+1)du,$$

ou encore

$$|e| \leq 2m^{(1)}(f).$$

Un calcul analogue sur l'intervalle $[x_i, x_{i+1}]$ donne

$$\int_{x_i}^{x_{i+1}} f(t)dt = (x_{i+1} - x_i)f(x_i) + e_i,$$

avec

$$|e_i| \leq \frac{(x_{i+1} - x_i)^2}{2} m_i^{(1)}(f).$$

Enfin sur l'intervalle $[a, b]$ nous obtenons la formule globale

$$\int_a^b f(u)du = \frac{(b-a)}{N}[f(x_0) + \cdots + f(x_{N-1})] + E,$$

avec

$$|E| \leq \frac{(b-a)^2}{N} M^{(1)}(f).$$

4.2.2. Méthode du milieu. La méthode des rectangles utilise la valeur de la fonction f en un point déterminé de l'intervalle. Il faudrait voir si un bon choix de ce point ne conduirait pas à une optimisation de la "qualité de la méthode". Encore faudrait il donner un sens à cette expression, ce que nous ferons par la suite. En attendant exposons la méthode du milieu qui consiste à utiliser comme valeur approchée de la fonction sur l'intervalle $[-1, 1]$, sa valeur au point milieu (c'est-à-dire en 0).

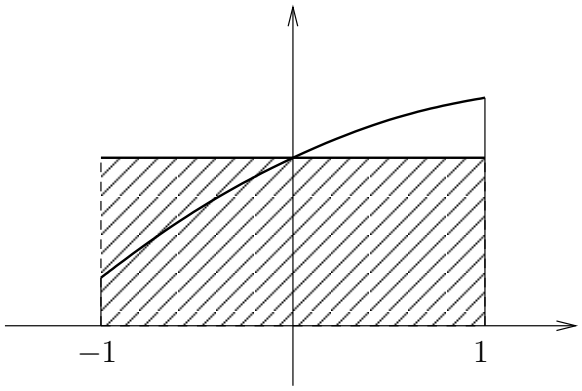

FIG. 2. Méthode du milieu

Ainsi

$$\int_{-1}^{1} f(t)dt = \int_{-1}^{1} f(0)dt + e = 2f(0) + e$$

En écrivant

$$f(u) = f(0) + \int_0^u f'(t)dt,$$

on obtient

$$\int_{-1}^{1} f(u)du = 2f(0) + \int_{-1}^{1} \left(\int_{0}^{u} f'(t)dt \right) du.$$

Or la méthode d'intégration par partie appliquée aux fonctions 1 et $f'(t)$ en dérivant $f'(t)$ et en intégrant 1 (on choisit $t-u$ comme primitive de la fonction 1) nous donne

$$\int_{0}^{u} f'(t)dt = uf'(0) - \int_{0}^{u} (t-u)f^{(2)}(t)dt.$$

Donc

$$\int_{-1}^{1} \left(\int_{0}^{u} f'(t)dt \right) du = - \int_{-1}^{1} \left(\int_{0}^{u} (t-u)f^{(2)}(t)dt \right) du.$$

On en conclut que

$$|e| = \left| \int_{-1}^{1} f(u)du - 2f(0) \right| \leq 1/3m^{(2)}(f).$$

Un calcul analogue sur l'intervalle $[x_i, x_{i+1}]$ donne

$$\int_{x_i}^{x_{i+1}} f(t)dt = (x_{i+1} - x_i)f(\frac{x_i + x_{i+1}}{2}) + e_i$$

avec

$$|e_i| \leq \frac{(x_{i+1} - x_i)^3}{24} m_i^{(2)}(f).$$

Sur l'intervalle $[a, b]$ on obtient par sommation

$$\int_{a}^{b} f(t)dt = \frac{(b-a)}{N}[f(x'_0) + \cdots + f(x'_{N-1})] + E$$

où

$$x'_i = \frac{x_i + x_{i+1}}{2}$$

et où

$$|E| \leq \frac{(b-a)^3}{24N^2} M^{(2)}(f).$$

4.2.3. Remarques concernant les deux méthodes précédentes. Ces deux méthodes sont toutes les deux basées sur une interpolation de degré 0 (c'est à dire par une constante). Cependant la deuxième a été en quelque sorte optimisée en choisissant un bon point d'interpolation. Pourquoi dire que cette deuxième méthode est meilleure que la première ? Ceci pour au moins deux raisons : d'une part l'erreur E est en $1/N$ dans la première méthode et en $1/N^2$ dans la deuxième, d'autre part la première méthode donne un calcul exact pour la classe des polynômes constants alors que la deuxième donne un calcul exact pour une classe plus vaste, celle des polynômes de degré 1 (ce qui ne veut pas dire qu'au hasard des choix, on ne puisse pas tomber sur des fonctions particulières autres, où l'une ou l'autre des formules donne une valeur exacte). Nous serons amenés plus tard à préciser ces critères de comparaison.

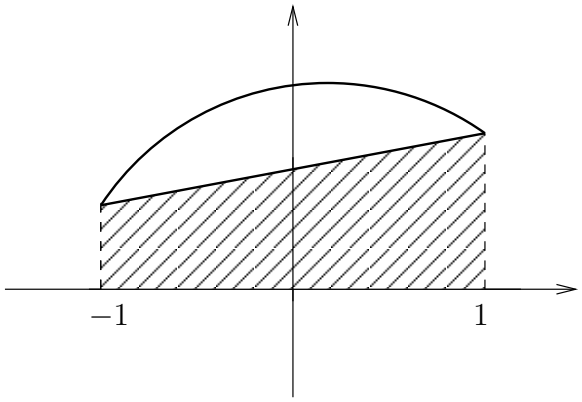

FIG. 3. Méthode des trapèzes

4.2.4. Méthode des trapèzes. Passons maintenant à une interpolation de Lagrange de degré 1 (c'est-à-dire par une fonction affine) aux bornes de l'intervalle.

Ainsi sur l'intervalle $[-1,1]$ on approchera $f(u)$ par

$$\frac{(u+1)f(1)+(u-1)f(-1)}{2}.$$

Alors on aura

$$\int_{-1}^{1} f(u)du = f(1)+f(-1)+e.$$

Le calcul de l'erreur e s'effectue en introduisant l'erreur dûe à l'interpolation de Lagrange, erreur calculée dans le chapitre portant sur l'interpolation :

$$\left| f(u) - \frac{(u+1)f(1)+(u-1)f(-1)}{2} \right| \leq \frac{1}{2}|(u-1)(u+1)|m^{(2)}(f).$$

En intégrant cette inégalité sur $[-1,1]$ on obtient la majoration suivante de l'erreur

$$|e| \leq \frac{2}{3}m^{(2)}(f).$$

Des calculs analogues sur l'intervalle $[x_i, x_{i+1}]$ nous donnent

$$\int_{x_i}^{x_{i+1}} f(u)du = \frac{x_{i+1}-x_i}{2}[f(x_i)+f(x_{i+1})]+e_i$$

avec

$$|e_i| \leq \frac{(x_{i+1}-x_i)^3}{12}m_i^{(2)}(f).$$

Enfin sur l'intervalle $[a,b]$ on peut écrire

$$\int_a^b f(u)du = \frac{b-a}{2}\left[\frac{f(x_0)+f(x_N)}{2}+f(x_1)+\cdots+f(x_{N-1})\right]+E$$

avec

$$|E| \leq \frac{(b-a)^3}{12N^2}M^{(2)}(f).$$

4.2.5. Méthode des tangentes. Cette méthode consiste à approcher la fonction par sa tangente au point milieu de l'intervalle. C'est donc une interpolation d'Hermite par un polynôme de degré 1.

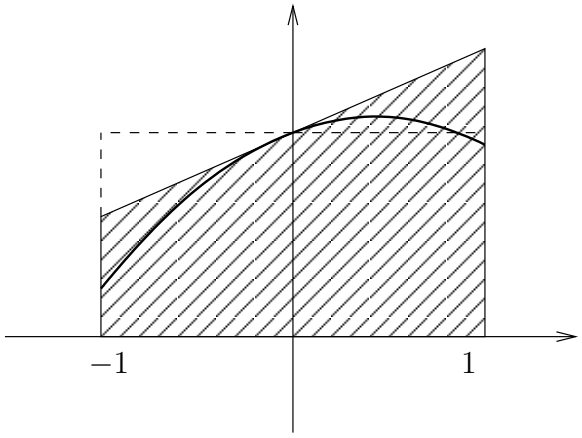

FIG. 4. Méthode des tangentes

Des considérations géométriques très simples montrent que cette méthode se ramène exactement à la méthode du milieu. Ainsi on dispose d'une piste pour expliquer que la méthode du milieu soit meilleure que celle des rectangles : en bien choisissant le point d'interpolation de degré 0, tout se passe comme si on disposait d'une interpolation de degré 1.

4.2.6. Méthode de Gauss. Comme dans le cas de l'interpolation de Lagrange de degré 0, nous allons essayer dans le cas de l'interpolation de Lagrange de degré 1, de choisir au mieux les deux points d'interpolation de manière à améliorer la méthode. Ainsi pour la méthode des trapèzes les point d'interpolation étaient aux bornes de l'intervalle, pour la méthode de Gauss les points d'interpolation seront des points bien choisis de l'intervalle.

On se place dans l'espace $\mathbb{R}_1[X]$ des polynômes de degré ≤ 1 sur $\mathbb{R}$. Soient α et β tels que $-1 \leq \alpha < \beta \leq 1$. On note δ_α et δ_β les formes linéaires sur $\mathbb{R}_1[X]$ définies par $\delta_u(P) = P(u)$. Si on note

$$P_\alpha(X) = \frac{X-\beta}{\alpha-\beta}$$

et

$$P_\beta(X) = \frac{X-\alpha}{\beta-\alpha}$$

alors $\delta_u(P_v) = \delta_{u,v}$ (0 si $u \neq v$, 1 si $u = v$). On en conclut que $(\delta_\alpha, \delta_\beta)$ est une base du dual de $\mathbb{R}_1[X]$. Donc la forme linéaire I su $\mathbb{R}_1[X]$ définie par

$$I(P) = \int_{-1}^{1} f(u)du$$

se décompose sous la forme

$$I = \lambda_\alpha \delta_\alpha + \lambda_\beta \delta_\beta.$$

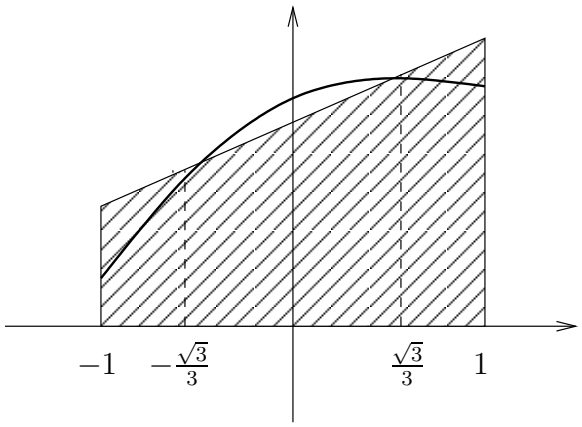

FIG. 5. Méthode de Gauss

Cette formule appliquée aux polynômes P_α et P_β donne les valeurs explicites des constantes λ_α et λ_β, ce qui permet d'écrire

$$I = \frac{2\beta}{\beta-\alpha}\delta_\alpha + \frac{2\alpha}{\alpha-\beta}\delta_\beta.$$

Ceci veut dire que pour tout polynôme $P \in \mathbb{R}_1[X]$ on a

$$\int_{-1}^{1} P(u)du = \frac{2\beta}{\beta-\alpha}P(\alpha) + \frac{2\alpha}{\alpha-\beta}P(\beta).$$

Pour une fonction f on va donc avec cette méthode écrire

$$\int_{-1}^{1} f(u)du = \frac{2\beta}{\beta-\alpha}f(\alpha) + \frac{2\alpha}{\alpha-\beta}f(\beta) + e,$$

et l'erreur e est nulle pour les polynômes de degré ≤ 1. Peut-on choisir les points α et β pour que la formule reste exacte ($e = 0$) pour un degré supérieur ? La réponse est donnée par :

PROPOSITION 4.1. *Il existe un couple et un seul de points α et β tels que pour tout polynôme P de degré ≤ 3 on ait*

$$\int_{-1}^{1} P(u)du = \frac{2\beta}{\beta-\alpha}P(\alpha) + \frac{2\alpha}{\alpha-\beta}P(\beta);$$

ces points sont

$$\alpha = \frac{-1}{\sqrt{3}} \quad \beta = \frac{1}{\sqrt{3}}.$$

De plus, quels que soient les points α et β il existe un polynôme de degré 4 qui ne vérifie pas la formule précédente.

Preuve. le calcul se fait à partir de la formule exacte pour les polynômes de degré ≤ 1 en imposant que cette formule reste exacte pour le polynôme X^2 et le polynôme X^3 (ce qui est nécessaire et suffisant pour que la formule reste exacte pour les polynômes de degré ≤ 3). Un calcul simple nous donne les points α et β attendus ainsi que les coefficients explicites de la formule. □

PROPOSITION 4.2. *Pour tout polynôme de degré ≤ 3 on a*

$$\int_{-1}^{1} P(u)du = P(-\frac{1}{\sqrt{3}}) + P(\frac{1}{\sqrt{3}}).$$

La **méthode de Gauss** consiste donc à utiliser cette formule pour approcher l'intégrale :

$$\int_{-1}^{1} f(u)du = f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}) + e.$$

La majoration de l'erreur e se fait maintenant en utilisant un polynôme de degré ≤ 3 qui interpole f. Plus précisément on considère le polynôme d'Hermite H de degré ≤ 3 tel que

$$H(-\frac{1}{\sqrt{3}}) = f(-\frac{1}{\sqrt{3}}) \quad H(\frac{1}{\sqrt{3}}) = f(\frac{1}{\sqrt{3}})$$
$$H'(-\frac{1}{\sqrt{3}}) = f'(-\frac{1}{\sqrt{3}}) \quad H'(\frac{1}{\sqrt{3}}) = f'(\frac{1}{\sqrt{3}}).$$

Dans ces conditions

$$\int_{-1}^{1} H(u)du = f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}),$$

donc

$$|e| = \left| \int_{-1}^{1} f(u)du - \int_{-1}^{1} H(u)du \right| \leq \int_{-1}^{1} |f(u) - H(u)|du.$$

On sait alors (cf. chapitre sur l'interpolation) que sur $[-1, 1]$

$$|f(u) - H(u)| \leq \frac{1}{24} \left| u^2 - \frac{1}{3} \right|^2 m^{(4)}(f),$$

d'où on tire par intégration

$$|e| \leq \frac{m^{(4)}(f)}{135}.$$

Sur chaque intervalle $[x_i, x_{i+1}]$ on obtient de manière analogue

$$\int_{x_i}^{x_{i+1}} f(t)dt = \frac{x_{i+1} - x_i}{2} \left[f\left(\frac{x_{i+1} + x_i}{2} + \frac{x_{i+1} - x_i}{2\sqrt{3}} \right) + f\left(\frac{x_{i+1} + x_i}{2} - \frac{x_{i+1} - x_i}{2\sqrt{3}} \right) \right] + e_i$$

avec

$$|e_i| \leq \left(\frac{x_{i+1} - x_i}{2} \right)^5 \frac{m_i^{(4)}(f)}{135}.$$

Par sommation on obtient sur $[a, b]$

$$\int_a^b f(u)du = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(u)du + E$$

avec

$$|E| \leq \frac{(b-a)^5}{N^4} \frac{M^{(4)}(f)}{4320}.$$

REMARQUE 4.3. En fait bien qu'au départ on ait travaillé sur une interpolation de Lagrange de degré 1, tout se passe ensuite de la même façon que pour une interpolation d'Hermite de degré 3. C'est ce qui fait la performance de la méthode.

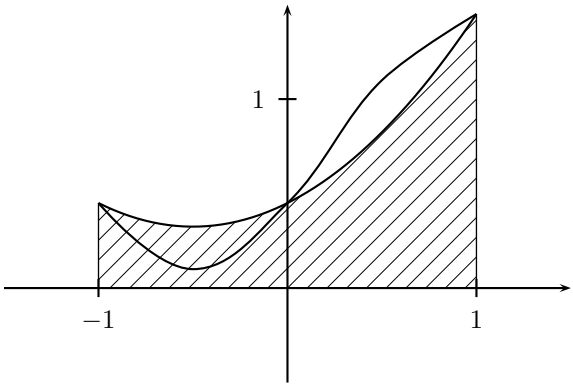

FIG. 6. Métode de Simpson

4.2.7. Méthode de Simpson. Nous partons maintenant de l'interpolation de Lagrange de degré 2 (par une fonction parabolique) aux bornes de l'intervalle et en un point γ tel que $-1 < \gamma < 1$. Nous employons une démarche analogue à celle du paragraphe précédent. Notons $\delta_{-1}, \delta_\gamma, \delta_1$ les formes linéaires sur l'espace $\mathbb{R}_2[X]$ des polynômes de degré ≤ 2 définies par

$$\delta_u(P) = P(u).$$

Les polynômes

$$P_{-1}(X) = \frac{(X-1)(X-\gamma)}{2(1+\gamma)},$$
$$P_\gamma(X) = \frac{(X-1)(X+1)}{(\gamma-1)(\gamma+1)},$$
$$P_1(X) = \frac{(X+1)(X-\gamma)}{2(1-\gamma)},$$

vérifient $\delta_u(P_v) = \delta_{u,v}$. On en déduit que $(\delta_{-1}, \delta_\gamma, \delta_1)$ est une base du dual de $\mathbb{R}_2[X]$ et donc que la forme linéaire I définie par

$$I(P) = \int_{-1}^{1} P(t)dt$$

se décompose sur cette base

$$I = \lambda_{-1}\delta_{-1} + \lambda_\gamma\delta_\gamma + \lambda_1\delta_1.$$

Ceci veut dire que pour tout polynôme P de degré ≤ 2 on a

$$\int_{-1}^{1} P(t)dt = \lambda_{-1}P(-1) + \lambda_\gamma P(\gamma) + \lambda_1 P(1).$$

Par un bon choix du point γ on va voir que cette formule persiste pour les polynômes de degré ≤ 3. On laisse au lecteur d'établir la proposition suivante

PROPOSITION 4.4. *Il existe un point γ et un seul tel que pour tout polynôme de degré ≤ 3 on ait*

$$\int_{-1}^{1} P(t)dt = \lambda_{-1}P(-1) + \lambda_\gamma P(\gamma) + \lambda_1 P(1).$$

Cette valeur de γ est 0. De plus il existe un polynôme de degré 4 pour lequel cette formule n'a pas lieu.

Ce point $\gamma = 0$ étant choisi on obtient la formule dite des 3 niveaux.

PROPOSITION 4.5. *Pour tout polynôme de degré ≤ 3*

$$\int_{-1}^{1} P(t)dt = \frac{1}{3}[P(-1) + P(1) + 4P(0)].$$

La **méthode de Simpson** consiste donc à utiliser cette formule pour approcher l'intégrale :

$$\int_{-1}^{1} f(u)du = \frac{1}{3}[f(-1) + f(1) + 4f(0)] + e.$$

Soit H le polynôme de degré ≤ 3 tel que

$$H(-1) = f(-1) \quad H(0) = f(0) \quad H(1) = f(1) \quad H'(0) = f'(0).$$

On a alors

$$|e| = \left| \int_{-1}^{1} f(u)du - \int_{-1}^{1} H(u)du \right| \leq \int_{-1}^{1} |f(u) - H(u)|\, du.$$

On sait (cf. chapitre sur l'interpolation) que sur $[-1, 1]$

$$|f(u) - H(u)| \leq \frac{1}{4!}u^2|(u-1)(u+1)|m^{(4)}(f)$$

ce qui donne par intégration

$$|e| \leq \frac{m^{(4)}(f)}{90}.$$

Sur les intervalles $[x_i, x_{i+1}]$ on obtient

$$\int_{x_i}^{x_{i+1}} f(u)du = \frac{x_{i+1} - x_i}{6}\left[f(x_i) + f(x_{i+1}) + 4f\left(\frac{x_i + x_{i+1}}{2}\right)\right] + e_i$$

avec

$$|e_i| \leq \left(\frac{x_{i+1} - x_i}{2}\right)^5 \frac{m_i^{(4)}(f)}{90}.$$

Sur l'intervalle $[a, b]$ tout entier on peut écrire

$$\int_a^b f(u)du = \frac{b-a}{6N}\left[f(x_0) + f(x_N) + 2\sum_{i=1}^{N-1} f(x_i) + 4\sum_{i=0}^{N-1} f\left(\frac{x_i + x_{i+1}}{2}\right)\right] + E$$

avec

$$|E| \leq \frac{(b-a)^5}{N^4}\frac{M^{(4)}(f)}{2880}.$$

L'erreur est là aussi comme pour la méthode de Gauss en $1/N^4$, mais avec des coefficients un peu moins bons. Cependant la formule ne fait pas intervenir comme dans celle de Gauss des nombres "compliqués" ($\sqrt{3}$) ce qui explique qu'elle ait été plus prisée avant l'apparition des ordinateurs.

Peut on là aussi optimiser mieux la méthode en choisissant de manière astucieuse les **trois points** d'interpolation ? Ceci est en partie l'objet de l'étude plus générale du paragraphe suivant.

4.2.8. Méthodes interpolatoires et polynômes orthogonaux. Soient $P_0, P_1, \cdots, P_n$ les $n+1$ premiers polynômes de Legendre. Rappelons que les polynômes de Legendre forment l'unique suite de polynômes telle que

a) $deg(P_i) = i$
b) $\int_{-1}^{1} P_i(u)P_j(u)du = 0 \quad (i \neq j)$
c) $P_i(1) = 1$.

Appelons $a_1, \cdots, a_n$ les racines de P_n ; elles sont distinctes et appartiennent à l'intervalle $[-1, 1]$. Les formes linéaires $(\delta_{a_i}, \cdots, \delta_{a_n})$ définies par $\delta_{a_i}(P) = P(a_i)$ forment une base du dual de l'espace $\mathbb{R}_{n-1}[X]$ des polynômes de degré $\leq n-1$ (espace de dimension n). On en conclut que la forme linéaire I définie sur $\mathbb{R}_{n-1}[X]$ par

$$I(P) = \int_{-1}^{1} P(u)du$$

se décompose sur cette base sous la forme

$$I = \lambda_{a_1}\delta_{a_1} + \cdots + \lambda_{a_n}\delta_{a_n},$$

ce qui veut dire que pour tout polynôme P de degré $\leq n-1$ on a

$$\int_{-1}^{1} P(u)du = \lambda_{a_1}P(a_1) + \cdots + \lambda_{a_n}P(a_n).$$

Le choix des racines du polynôme de Legendre P_n pour les points a_i est un très bon choix dans le mesure où

PROPOSITION 4.6. *Pour tout polynôme de degré $\leq 2n-1$ on a*

$$\int_{-1}^{1} P(u)du = \lambda_{a_1}P(a_1) + \cdots + \lambda_{a_n}P(a_n).$$

Preuve. Il suffit de montrer que la formule est vraie pour les polynômes d'une base de l'espace $\mathbb{R}_{2n-1}[X]$ des polynômes de degré $\leq 2n-1$. La famille $(P_0, \cdots, P_{n-1}, P_0P_n, \cdots, P_{n-1}P_n)$ est une base de $\mathbb{R}_{2n-1}[X]$ (tous les degrés sont représentés une fois et une seule). Il est clair que $P_0, \cdots, P_{n-1}$ vérifient la formule puisque ces polynômes sont de degré $\leq n-1$. Quand aux P_iP_n $(0 \leq i \leq n-1)$ il est clair en utilisant l'orthogonalité de la suite des polynômes de Legendre que $\int_{-1}^{1} P_i(u)P_n(u)du = 0$ et que sachant que les a_i sont les racines de P_n, $\delta_i(P_jP_n) = P_j(a_i)P_n(a_i) = 0$. Ils vérifient donc eux ausi la formule.

Toute autre famille de points est moins bonne que la famille (a_i). □

PROPOSITION 4.7. *Soient $\alpha_1, \cdots, \alpha_n$ des points distincts de l'intervalle $[-1, 1]$ dont l'un au moins n'est pas racine du polynôme de Legendre de degré n. Alors il existe un polynôme Q de degré $\leq 2n-1$ tel que*

$$\int_{-1}^{1} Q(u)du \neq \lambda_{\alpha_1}Q(\alpha_1) + \cdots + \lambda_{\alpha_n}Q(\alpha_n).$$

Preuve. Posons $Q_n(X) = (X-\alpha_1)\cdots(X-\alpha_n)$.Par hypothèse, Q_n n'est pas proportionnel au polynôme de Legendre P_n, il n'est donc pas orthogonal au sous espace des polynômes de degré $\leq n-1$. Il existe un polynôme R de ce sous espace tel que $\int_{-1}^{1} R(u)Q_n(u)du \neq 0$. Il suffit alors de prendre $Q = RQ_n$. □

Cette propriété optimale des racines des polynômes de Legendre peut être utilisée pour avoir une meilleure majoration de l'erreur quand on remplace l'intégrale à calculer par l'intégrale du polynôme d'interpolation de Lagrange de degré $\leq n-1$ en n points de l'intervalle. C'est ainsi que nous avons procédé pour la méthode de Gauss, avec $n = 2$. Les point $-1/\sqrt{3}$ et $1/\sqrt{3}$ sont les racines du polynôme de Legendre de degré 2 ($P_2(X) = \frac{3X^2-1}{2}$).

Dans le cas de l'interpolation de degré 2 (avec les notations précédentes $n = 3$), nous pouvons obtenir mieux que par la méthode de Simpson et par la méthode de Gauss. La méthode de Simpson en effet est une méthode utilisant l'interpolation de degré 2, partiellement optimisée par le choix du point milieu de l'intervalle, mais qui n'est pas complètement optimisée puisque les deux autres points d'interpolation sont les bords de l'intervalle et non pas des racines du polynôme de Legendre de degré 3. Le polynôme de Legendre de degré 3 est

$$P_3(X) = \frac{5X^2 - 3X}{2},$$

ce qui donne pour racines

$$a_1 = -\frac{\sqrt{3}}{\sqrt{5}} \quad a_2 = 0 \quad a_3 = \frac{\sqrt{3}}{\sqrt{5}}.$$

Un calcul simple donne alors pour tout polynôme P de degré ≤ 5

$$\int_{-1}^{1} P(u)du = \frac{1}{9}\left[5P\left(-\frac{\sqrt{3}}{\sqrt{5}}\right) + 8P(0) + 5P\left(\frac{\sqrt{3}}{\sqrt{5}}\right)\right].$$

Ainsi pour toute fonction f de classe C^5 on a

$$\int_{-1}^{1} f(u)du = \frac{1}{9}\left[5P\left(-\frac{\sqrt{3}}{\sqrt{5}}\right) + 8P(0) + 5P\left(\frac{\sqrt{3}}{\sqrt{5}}\right)\right] + e$$

où $|e|$ se majore en utilisant l'inégalité

$$|f(u) - P(u)| \leq \frac{1}{6!}u^2(u^2 - 3/5)^2 m^{(6)}(f),$$

ce qui par intégration fournit

$$|e| \leq \frac{1}{15750} m^{(6)}(f).$$

Sur chaque intervalle $[x_i, x_{i+1}]$ on a une erreur

$$e_i \leq \left(\frac{x_{i+1} - x_i}{2}\right)^7 \frac{m_i^{(6)}(f)}{15750},$$

et sur $[a, b]$ une erreur

$$|E| \leq \frac{(b-a)^7}{N^6} \frac{M^{(6)}(f)}{15750 \times 128}.$$

4.3. Présentation générale des quadratures élémentaires - Ordre d'une méthode

Toutes les méthodes vues jusqu'à présent entrent dans le cadre général suivant (**méthodes de quadrature élémentaires**). On a sur l'intervalle $[x_i, x_{i+1}]$ une formule du type

$$\int_{x_i}^{x_{i+1}} f(u)du = (x_{i+1} - x_i) \sum_{j=0}^{l_i} \omega_{i,j} f(\eta_{i,j}) + e_i,$$

où

$$\eta_{i,j} \in [x_i, x_{i+1}] \quad \text{et} \quad \sum_{j=0}^{l_i} \omega_{i,j} = 1.$$

On essaie d'adapter les paramètres pour que la méthode soit la meilleure possible.

DÉFINITION 4.8. On dit qu'une méthode de quadrature est **d'ordre** N si la formule approchée est exacte pour les polynômes de degré $\leq N$ et inexacte pour au moins un polynôme de degré $N + 1$.

Remarquons que puisque $\sum_{j=0}^{l_i} \omega_{i,j} = 1$, une méthode de quadrature élémentaire est toujours au moins d'ordre 0.

Le lecteur est invité à préciser les paramètres qui correspondent aux diverses méthodes décrites précédemment.

4.4. Accélération de convergence - Méthode de Romberg

La méthode que nous allons décrire maintenant part d'une méthode d'accélération de convergence sur les suites.

4.4.1. Formule de Richardson. Soit u une suite de nombres réels convergeant vers un nombre réel a. On suppose que

$$u_n - a = \lambda k^n + O(k'^n)$$

où k et k' sont des réels tels que $0 < |k'| < |k| < 1$ et où λ est un réel non nul.

Alors

$$u_{n+1} - a = \lambda k^{n+1} + O(k'^n)$$
$$ku_n - ka = \lambda k^{n+1} + O(k'^n)$$

et par suite

$$\frac{u_{n+1} - ku_n}{1 - k} - a = O(k'^n).$$

On construit de cette manière une suite

$$v_n = \frac{u_{n+1} - ku_n}{1 - k}$$

qui converge vers a plus vite que la suite initiale u.

Plus généralement on peut supposer que

$$u_n - a = \lambda_1 k_1^n + \lambda_2 k_2^n + \cdots + \lambda_p k_p^n + O(k_{p+1}^n)$$

où

$$0 < |k_{p+1}| < |k_p| < \cdots < |k_1| < 1$$

et où $\lambda_1, \cdots, \lambda_n$ sont des réels non nuls. Pour toute suite w, notons $R_k w$ la suite définie par

$$R_k w_n = \frac{w_{n+1} - k w_n}{1-k}.$$

Alors

$$R_{k_1} u_n = \frac{u_{n+1} - k_1 u_n}{1-k_1}.$$

Puisque

$$u_{n+1} - a = \lambda_1 k_1^{n+1} + \lambda_2 k_2^{n+1} + \cdots + \lambda_p k_p^{n+1} + O(k_{p+1}^n)$$

et que

$$k_1 u_n - k_1 a = \lambda_1 k_1^{n+1} + \lambda_2 k_1 k_2^n + \cdots + \lambda_p k_1 k_p^n + O(k_{p+1}^n)$$

on obtient

$$u_{n+1} - k_1 u_n - (a - k_1 a) = (\lambda_2' k_2^n + \cdots + \lambda_p' k_p^n)(1-k_1) + O(k_{p+1}^n)$$

où

$$\lambda_2' = \lambda_2 \frac{(k_2 - k_1)}{1-k_1}, \cdots, \lambda_p' = \lambda_p \frac{(k_p - k_1)}{1-k_1}.$$

On en déduit que

$$R_{k_1} u_n - a = \lambda_2' k_2^n + \cdots + \lambda_p' k_p^n) + O(k_{p+1}^n).$$

Par conséquent on peut itérer le procédé et on obtient alors

$$R_{k_p} R_{k_{p-1}} \cdots R_{k_1} u_n - a = O(k_{p+1}^n).$$

REMARQUE 4.9. Reprenons la formule de Richardson à l'ordre 1. Il se peut que k ne soit pas connu. On pose alors

$$v_n = \frac{u_{n+1} - \mu_n u_n}{1 - \mu_n}$$

où

$$\mu_n = \frac{u_{n+1} - u_n}{u_n - u_{n-1}}.$$

Dans ces conditions

$$\mu_n - k = O\left((k'/k)^n\right),$$

donc

$$v_n - a = O(k'^n).$$

REMARQUE 4.10. On peut donner une version continue de la formule de Richardson.

Soit $f(y)$ telle que $\lim_{y \to 0} f(y) = \alpha_0$. On suppose que

$$f(y) = \alpha_0 + \alpha_1 y + \cdots + \alpha_p y^p + O(y^{p+1}).$$

Soit alors $0 < r < 1$ et $y_0 > 0$.

Formons la suite

$$A_{m,0} = f(r^m y_0),$$

et remarquons que

$$\lim_{m \to \infty} A_{m,0} = \alpha_0,$$

$$A_{m,0} = \alpha_0 + \alpha_1 y_0 r^m + \cdots + \alpha_p y_0^p (r^p)^m + O\left((r^{p+1})^m\right).$$

Donc en posant $k_1 = r, \cdots, k_p = r^p$ on se ramène à la formule de Richardson précédente. Plus précisément, pour $0 \leq n \leq p-1$ on définit

$$A_{m,n+1} = \frac{A_{m,n} - r^{n+1} A_{m-1,n}}{1 - r^{n+1}}.$$

Dans ces conditions

$$A_{m,p} - \alpha_0 = O\left((r^{p+1})^m\right).$$

4.4.2. Méthode de Romberg. Nous allons montrer comment marche cette méthode .

Découpons l'intervalle $[a, b]$ par un partage équidistant en $N = 2^n$ morceaux et posons $h = (b-a)/2^n$. Appelons

$$I_{2^n} = 1/2h[f(a) + f(b) + 2\sum_{i=1}^{2^n-1} f(x_i)]$$

la valeur approchée de l'intégrale donnée par la méthode des trapèzes.

La formule d'Euler-Maclaurin nous donne

$$\int_a^b f(u)du - I_{2^n} = \alpha(1/4)^n + O\left((1/8)^n\right).$$

On peut donc appliquer la méthode de Richardson à l'ordre 1 avec $k = 1/4$ et $k' = 1/8$. On pourra remarquer que dans ce cas on obtient la formule donnée par la méthode de Simpson sur 2^{n-1} intervalles. Il faut appliquer la méthode de Richardson à un ordre ≥ 3 pour trouver une formule qui n'est pas donnée par les considérations des paragraphes précédents.

CHAPITRE 5

Analyse numérique des équations différentielles

5.1. Introduction

5.1.1. Position du problème. Nous nous intéressons ici aux méthodes à un pas d'intégration des équations différentielles du premier ordre. Précisons le problème que nous nous posons. Nous cherchons à résoudre numériquement le problème de Cauchy

$$\left\{ \begin{array}{l} y' = f(x,y) \\ y(x_0) = y_0. \end{array} \right. \tag{5.1}$$

Dans toute la suite on supposera qu'il existe un voisinage compact V du point (x_0, y_0) tel que la fonction f soit continue sur V et lipschitzienne par rapport à y uniformément par rapport à x. C'est-à-dire qu'il existe une constante L telle que pour tout x, tout y_1, tout y_2 tels que $(x, y_1) \in V$ et $(x, y_2) \in V$ on ait

$$|f(x,y_2) - f(x,y_1)| \leq L|y_2 - y_1|.$$

Nous noterons aussi

$$M = \sup_{(x,y)\in V} |f(x,y)|.$$

Soit $T > 0$ tel que

$$[x_0 - T, x_0 + T] \times \{y \mid |y - y_0| \leq MT\} \subset V.$$

On sait alors qu'il existe une solution unique qu'on notera ϕ du problème de Cauchy (5.1) définie sur l'intervalle $[x_0 - T, x_0 + T]$.

Nous allons chercher à approcher la fonction ϕ sur l'intervalle $[x_0, x_0 + T]$ (ou sur l'intervalle $[x_0 - T, x_0]$).

Dans certains cas on sera amené à faire des hypothèses plus fortes sur la régularité de la fonction f, ce qui entrainera une plus forte régularité de la fonction solution ϕ.

5.1.2. Notations. On choisit un partage

$$x_0 < x_1 < \cdots < x_n = x_0 + T$$

de l'intervalle d'étude $[x_0, x_0 + T]$, et on pose

$$h_k = x_{k+1} - x_k.$$

On définit aussi

$$H = \max_{0\leq k\leq n-1} h_k.$$

La fonction ϕ étant la solution du problème (5.1) on pose pour $0 \leq k \leq n$

$$y_k = \phi(x_k).$$

Une **méthode numérique à un pas** permettra de calculer une approximation de y_k à partir d'une approximation de y_{k-1}.

5.2. Généralités sur les méthodes

Partons de la formule

$$\phi(x+h) = \phi(x) + \int_0^h \phi'(x+t)dt.$$

Une première idée pour calculer $y_1 = \phi(x_1) = \phi(x_0+h_0)$ est d'utiliser cette formule, donc d'écrire

$$y_1 = y_0 + \int_0^{h_0} \phi'(x_0+t)dt,$$

puis d'évaluer l'intégrale

$$\int_0^{h_0} \phi'(x_0+t)dt$$

par une méthode de calcul numérique approché d'intégrales.

5.2.1. Méthode de la tangente d'Euler.

5.2.1.1. *Description.* Utilisons comme méthode de calcul la méthode de Riemann. Nous obtenons alors

$$\int_0^{h_0} \phi'(x_0+t)dt = h_0\phi'(x_0) + O(h_0^2),$$

ce qui donne

$$y_1 = y_0 + h_0\phi'(x_0) + O(h_0^2).$$

Ceci nous conduit à prendre pour approximation du point y_1 le point u_1 défini par

$$u_1 = u_0 + h_0 f(x_0, u_0),$$

où u_0 est une approximation de y_0. Cette méthode nous donne au rang k

$$u_k = u_{k-1} + h_{k-1} f(x_{k-1}, u_{k-1}).$$

Cette méthode est appelée méthode de la tangente d'Euler en raison de l'interprétation géométrique suivante.

5.2.1.2. *Interprétation géométrique.* On peut voir une équation différentielle comme la donnée d'un champ de vecteurs : à tout point (x, y) on fait correspondre le vecteur $(1, f(x, y))$ de coefficient directeur $f(x, y)$. Le problème de Cauchy consiste à trouver la ligne de champ qui passe par le point (x_0, y_0). On la construit de manière approchée en traçant une ligne polygonale partant de (x_0, y_0) et suivant tout d'abord la direction de la tangente de coefficient directeur $f(x_0, y_0)$ à la trajectoire en ce point. On arrive à un point (x_1, y_1) et en ce point on prend la direction indiquée par le champ de vecteurs, c'est-à-dire la droite de coefficient directeur $f(x_1, y_1)$. Évidemment en général dès le point (x_1, y_1) on a quitté la bonne trajectoire. Mais on peut espérer que si on fait de tous petits pas, on ne décolle pas trop (cf. figure 1).

5.2.2. Métode d'Euler modifiée.

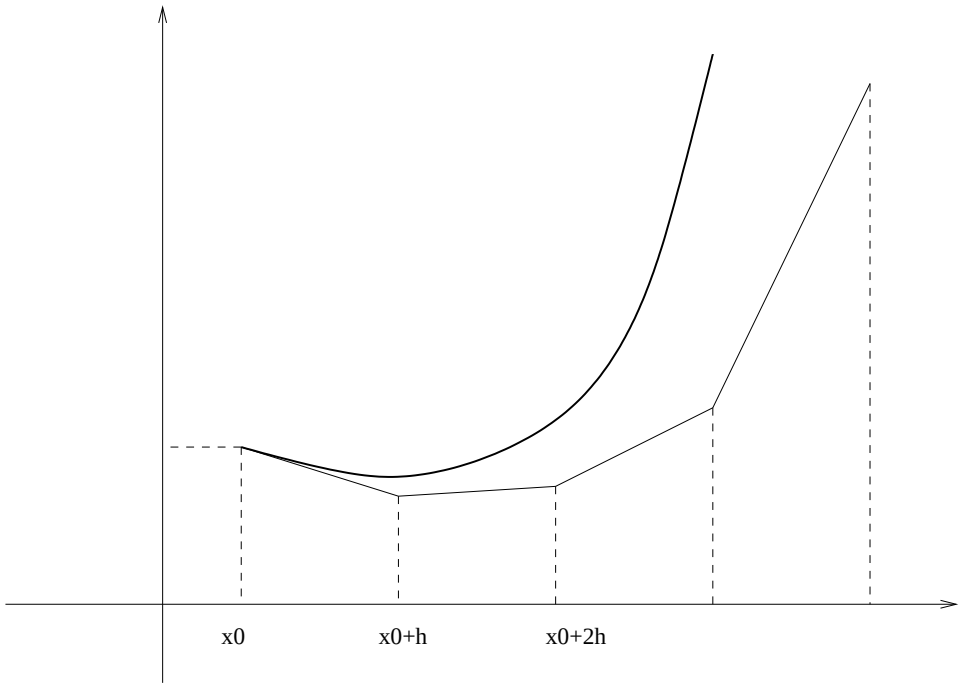

FIG. 1. Méthode de la tangente d'Euler

5.2.2.1. *Description.* En améliorant la méthode d'intégration de

$$\int_0^{h_0} \phi'(x_0+t)dt$$

par la méthode des trapèzes par exemple

$$\int_0^{h_0} \phi'(x_0+t)dt = \frac{h_0}{2}\left(\phi'(x_0)+\phi'(x_0+h_0)\right)+O(h_0^3),$$

alors

$$\phi(x_0+h) = \phi(x_0)+\frac{h_0}{2}\left(f(x_0,y_0)+f(x_0+h_0,\phi(x_0+h_0)\right)+O(h_0^3),$$

soit encore

$$y_1 = y_0+\frac{h_0}{2}\left(f(x_0,y_0)+f(x_1,y_1)\right)+O(h_0^3).$$

Malheureusement au second membre intervient la valeur y_1 que l'on veut justement calculer. On peut penser à utiliser une méthode d'approximations successives pour calculer y_1 : on injecte une valeur approchée de y_1 dans le second membre de la formule précédente et on obtient une nouvelle valeur approchée de y_1 qu'on espère meilleure dans le premier membre de la formule. Pour cela on part de la valeur approchée v_1 de y_1 donnée par la méthode précédente de la tangente d'Euler

$$v_1 = u_0+h_0 f(x_0,y_0)$$

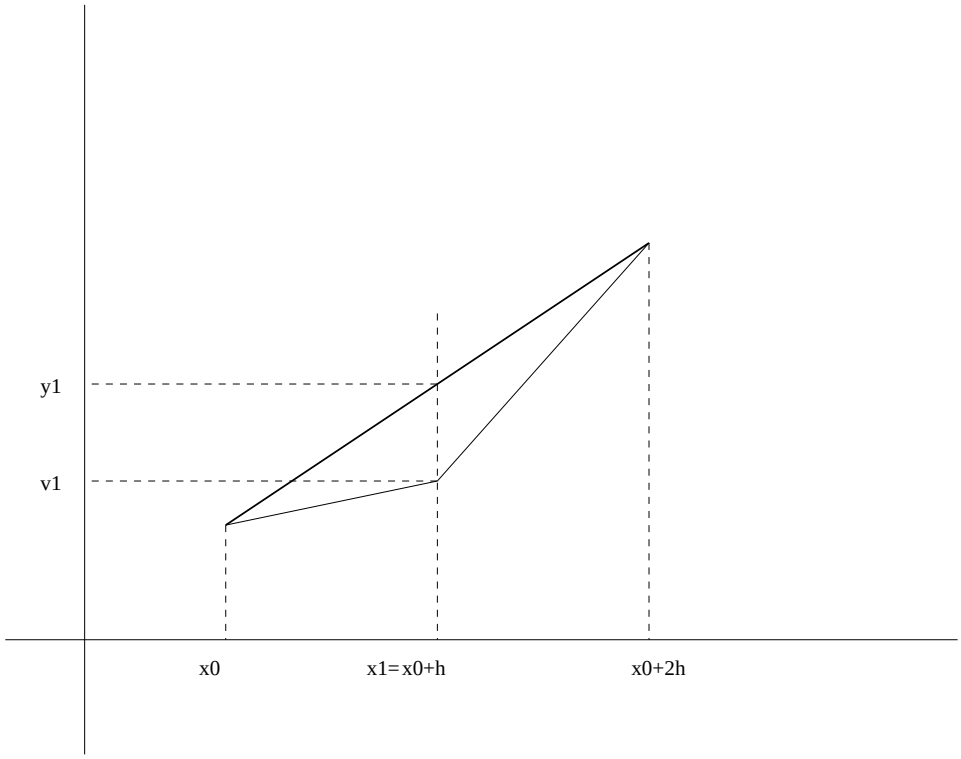

FIG. 2. Méthode d'Euler modifiée

(dans ce calcul on considèrera que $v_0 = y_0$). Est-il la peine d'aller plus loin dans la méthode des approximations successives ? Une évaluation de l'erreur nous donne

$$f(x_1, \phi(x_0 + h_0)) - f(x_1, v_1) = (y_1 - v_1)\frac{\partial f}{\partial y}(x_1, y_1),$$

ce qui donne en utilisant l'évaluation de l'erreur de la méthode d'Euler

$$|f(x_1, \phi(x_0 + h_0)) - f(x_1, v_1)| = O(h_0^2).$$

En conséquence, il n'est pas utile d'aller plus loin pour rester dans les limites d'une erreur en $0(h_0^3)$ et on obtient en définitive

$$y_1 = y_0 + \frac{h_0}{2}\left(f(x_0, y_0) + f(x_1, v_1)\right) + 0(h_0^3),$$

où

$$v_1 = y_0 + h_0 f(x_0, y_0).$$

En conclusion nous sommes amenés à prendre le schéma suivant

$$\left\{\begin{array}{l} u_0 \text{ est une valeur initiale proche de } y_0 \\ v_k = u_{k-1} + h_{k-1} f(x_{k-1}, u_{k-1}) \\ u_k = u_{k-1} + \dfrac{h_{k-1}}{2}\left(f(x_{k-1}, u_{k-1}) + f(x_k, v_k)\right) \end{array}\right. \tag{5.2}$$

Cette méthode est appelée méthode d'Euler modifiée et admet l'interprétation géométrique suivante.

5.2.2.2. *Interprétation géométrique.* Si partant du point (x_0, y_0) la méthode de la tangente d'Euler faisant un bout de chemin sur la droite de coefficient directeur $f(x_0, y_0)$ on arrive au point (x_1, v_1) alors on fait la moyenne entre le coefficient directeur en (x_0, y_0) et le coefficient directeur en (x_1, v_1) et finalement on emprunte à partir de (x_0, y_0) la droite ayant pour coefficient directeur cette moyenne (cf. figure 2).

5.2.3. Généralisation. Le nombre u_0 étant une valeur approchée de y_0, on construit la suite $(u_k)_{k\geq 0}$ en partant de u_0 et en calculant u_k pour $k \geq 1$ par une formule récurrente du type :

$$u_k = u_{k-1} + h_{k-1}\Phi(x_{k-1}, u_{k-1}, h_{k-1}). \tag{5.3}$$

C'est ce qu'on appelle une **méthode à un pas**. Par exemple dans la méthode d'Euler on prend :

$$\Phi(x, y, h) = f(x, y)$$

et dans la méthode d'Euler modifiée :

$$\Phi(x, y, h) = \frac{1}{2}\left(f\left(x, y\right) + f\left(x + h, y + hf\left(x, y\right)\right)\right).$$

La question est alors la suivante : comment obtenir une fonction Φ intéressante.

5.2.3.1. *Première idée : la formule de Taylor.* On peut penser à la formule de Taylor, ce qui nous donne :

$$y_1 = y_0 + h\phi'(x_0) + \cdots + \frac{h_0^p}{p!}\phi^{(p)}(x_0) + O(h_0^{p+1}).$$

Posons :

$$f^{(0)}(x, y) = f(x, y),$$

$$f^{(1)}(x, y) = \frac{\partial(f^{(0)})}{\partial x}(x, y) + \frac{\partial(f^{(0)})}{\partial y}(x, y)f(x, y),$$

et plus généralement :

$$f^{(k)}(x, y) = \frac{\partial(f^{(k-1)})}{\partial x}(x, y) + \frac{\partial(f^{(k-1)})}{\partial y}(x, y)f(x, y).$$

Il est clair que :

$$f^{(k)}(x, \phi(x)) = \phi^{(k+1)}(x),$$

si bien que :

$$y_1 = y_0 + h\left(f^{(0)}(x_0, y_0) + \frac{h}{2}f^{(1)}(x_0, y_0) + \cdots + \frac{h^{(p-1)}}{p!}f^{(p-1)}(x_0, y_0)\right) + O(h^{p+1}).$$

On est donc amené à poser :

$$\Phi(x, y, h) = f^{(0)}(x, y) + \frac{h}{2}f^{(1)}(x, y) + \cdots + \frac{h^{(p-1)}}{p!}f^{(p-1)}(x, y).$$

Il est visible que dans cette méthode, le calcul des dérivées successives peut s'avérer compliqué. De plus l'erreur est difficile à contrôler et peut être catastrophique si les dérivées de f ne sont pas "petites".

5.2.3.2. *Deuxième idée : une formule d'intégration.* Cette idée, que nous avons évoquée au début de cette section, consiste à partir de $q > 0$ nombres réels $c_1, c_2, \cdots, c_q$ distincts ou non, d'écrire des formules de quadrature numérique du type :

$$\int_0^{c_i} \psi(t)dt \approx \sum_{j=1}^{q} a_{i,j}\psi(c_j) \tag{5.4}$$

et :

$$\int_0^{1} \psi(t)dt \approx \sum_{j=1}^{q} b_{j}\psi(c_j). \tag{5.5}$$

Rappelons que l'intervalle $[x_0, x_0 + T]$ a été découpé sous la forme :

$$x_0 < x_1 \cdots < x_n = x_0 + T.$$

Chaque sous-intervalle $[x_k, x_{k+1}]$ où $x_{k+1} = x_k + h_k$ contient lui-même un sous-découpage basé sur q points. Ces q points sont les points :

$$x_{k,i} = x_k + c_i h_k,$$

où $1 \leq i \leq q$. Alors les formules de quadrature numérique choisies nous permettent d'écrire :

$$\phi(x_{k,i}) = y_k + \int_{x_k}^{x_{k,i}} f(t, \phi(t))dt,$$

ce qui donne l'approximation :

$$\phi(x_{k,i}) = y_k + h_k \sum_{j=1}^{q} a_{i,j} f(x_{k,j}, \phi(x_{k,j})).$$

On peut aussi écrire :

$$\phi(x_k + h_k) = y_k + \int_{x_k}^{x_k+h_k} f(t, \phi(t))dt,$$

ce qui donne l'approximation :

$$y_{k+1} \approx y_k + h_k \sum_{j=1}^{q} b_{j} f(x_{k,j}, \phi(x_{k,j})).$$

Ces considérations permettent de penser au schéma suivant :

$$\left\{ \begin{array}{lcl} u_{k,i} & = & u_k + h_k \displaystyle\sum_{j=1}^{q} a_{i,j} f(x_{k,j}, u_{k,j}) \\ u_{k+1} & = & u_k + h_k \displaystyle\sum_{j=1}^{q} b_{j} f(x_{k,j}, u_{k,j}) \\ x_{k,j} & = & x_k + c_j h_k, \end{array} \right. \tag{5.6}$$

où

$$\left\{ \begin{array}{ccccc} 0 & \leq & k & \leq & n-1 \\ 1 & \leq & j & \leq & q \\ 1 & \leq & i & \leq & q. \end{array} \right.$$

En posant :

$$K_{k,i} = f(x_{k,i}, u_{k,i})$$

on obtient :

$$\text{(5.7)}\qquad \begin{cases} K_{k,i} &= f\left(x_{k,i}, u_k + h_k \displaystyle\sum_{j=1}^{q} a_{i,j} K_{k,j}\right) \\ u_{k+1} &= u_k + h_k \displaystyle\sum_{j=1}^{q} b_j K_{k,j}. \end{cases}$$

Le schéma ainsi décrit peut être **explicite** : le calcul de $u_{k,i}$ (ou de $K_{k,i}$) ne fait intervenir que des $u_{k,j}$ (des $K_{k,j}$) avec $j < i$, c'est-à-dire déjà calculés, ou **implicite** dans le cas contraire.

Il reste à savoir comment choisir les coefficients $a_{i,j}$, b_j, c_j, c'est-à-dire quelles formules de quadrature choisir. Toutes ces méthodes rentrent sous la dénomination commune de méthode de Runge-Kutta. Cependant une de ces méthodes porte le nom de méthode de Runge-Kutta classique.

5.2.4. Les méthodes de Runge-Kutta. Nous retrouvons bien entendu par ce principe général les cas déjà traitées.

- **Métode d'Euler.** Dans ce cas on prend $q = 1$, $c_1 = 0$, $a_{1,1} = 0$, $b_1 = 1$. Avec ces valeurs on retrouve effectivement la méthode d'Euler.
- **Méthode d'Euler modifiée.** Dans cette méthode aussi appelée **méthode de Heun** on prend $q = 2$, $c_1 = 0, c_2 = 1$, $a_{1,1} = 0, a_{1,2} = 0, a_{2,1} = 1, a_{2,2} = 0$, $b_1 = 1/2, b_2 = 1/2$. Les méthodes d'intégration utilisées sont d'une part la méthode des rectangles pour la méthode d'intégration numérique (5.4) où $i = 2$ (pour $i = 1$ il n'y a rien à faire) et celle des trapèzes pour la méthode d'intégration numérique (5.5).

  Toujours pour $q = 2$ on peut généraliser cette méthode en prenant un nombre $0 < \alpha \leq 1$ puis $c_1 = 0, c_2 = \alpha$, $a_{1,1} = 0, a_{1,2} = 0, a_{2,1} = 1, a_{2,2} = 0$, $b_1 = 1 - 1/(2\alpha)), b_2 = 1/(2\alpha)$. Pour $\alpha = 1$ on obtient la méthode de Heun, pour $\alpha = 1/2$ on obtient une méthode basée sur la méthode du milieu pour la méthode d'intégration numérique (5.5) et la métode des rectangles pour le calcul (5.4).
- **Méthode de Runge-Kutta classique.** On prend $q = 4$, $c_1 = 0, c_2 = 1/2, c_3 = 1/2, c_4 = 1$, $a_{2,1} = 1/2, a_{3,2} = 1/2, a_{4,3} = 1$, les autres $a_{i,j}$ sont nuls, $b_1 = 1/6, b_2 = 1/3, b_3 = 1/3, b_4 = 1/6$.

  On peut alors écrire le shéma de calcul de la façon suivante :

$$\text{(5.8)}\qquad \begin{cases} K_{k,1} &= f\left(x_k, u_k\right) \\ K_{k,2} &= f\left(x_k + \dfrac{h_k}{2}, u_k + \dfrac{h_k}{2} K_{k,1}\right) \\ K_{k,3} &= f\left(x_k + \dfrac{h_k}{2}, u_k + \dfrac{h_k}{2} K_{k,2}\right) \\ K_{k,4} &= f\left(x_{k+1}, u_k + h_k K_{k,3}\right) \\ u_{k+1} &= u_k + h_k \left[\dfrac{K_{k,1}}{6} + \dfrac{K_{k,2}}{3} + \dfrac{K_{k,3}}{3} + \dfrac{K_{k,4}}{6}\right]. \end{cases}$$

ANNEXE A

Représentation des nombres

A.1. Introduction

Nous nous intéressons à la façon de définir les objets que nous allons utiliser dans les calculs sur machine pour représenter les nombres. Ceci passe par au moins deux stades :

- La **définition de la nature mathématique** de ces objets
- Leur **représentation interne**

Nous regarderons les cas des **réels** et des **entiers**.

A.2. Les nombres réels

Les nombres réels sont approchés par les **nombres à virgule flottante** (en simple précision, double précision, précisions étendues).

Les normes **IEEE-754** et **IEEE-854** (**I**nstitute of **E**lectrical and **E**lectronics **E**ngineers) définissent ces nombres ainsi que certaines façons d'opérer dessus $(+, -, \div, \times, \sqrt{})$.

A.2.1. Première approche. On fixe une **base** de numération b, un nombre fixe p de digits, un signe $S = (-1)^s$ (avec $s = 0$ ou $s = 1$) un exposant e compris entre deux valeurs fixées $e_{min} \leq e \leq e_{max}$. On considère alors les nombres de la forme :

$$x = (-1)^s x_0.x_1x_2\cdots x_{p-1}b^e,$$

où $0 \leq x_i < b$.

la partie S est le **signe**, $x_0x_1\cdots x_{p-1}$ la **mantisse** et e l'**exposant**.

A.2.2. Les flottants normalisés. Un **flottant normalisé** est un réel non nul qui peut s'écrire sous la forme précédente avec $x_0 \neq 0$. Les nombres réels qui seront exactement représentables en machine seront 0 et les flottants normalisés. Un réel pourra ne pas être exactement représenté pour les raisons suivantes : mantisse trop longue (infinie pour les nombres non b-adiques), exposant trop grand (**overflow**) ou trop petit (**underflow**).

A.2.3. Cas concrets. Les bases utilisées sont 2 et 10. Chacune a ses avantages.

La base 2 sera choisie quand surtout des pas de calcul et peu d'affichages (langages de programmation généralistes par exemple).

La base 10 sera choisie si on a des affichages à chaque pas (calculette par exemple) et donc beaucoup de traductions en décimal à faire.

La norme **IEEE-754 régit le cas de la base** 2.

A.2.4. La base 2, simple précision. • **La définition :** $b = 2$, $p = 24$, $e_{max} = 127$, $e_{min} = -126 = -e_{max} + 1$. On introduit aussi l'exposant **biaisé** $E = e + e_{max}$.

• **Le stockage en machine :** 32 bits,

$$sE_7 \cdots E_0 x_1 \cdots x_{23}.$$

Remarquons que x_0 qui vaut 1 n'est pas écrit.

exemple : $(-1.00110001000111111110011)_2 \times 2^{(11)_2}$ se représente par les 32 bits suivants :

$$1 \mid 10000010 \mid 00110001000111111110011.$$

A.2.5. Bits non utilisés. Remarquons que $1 \leq E \leq 2e_{max} = 254$. Il reste donc la valeur $E = 0$ et $E = 255$ qui sont non encore utilisées.

On utilise $E = 0$ avec une mantisse nulle pour 0.

On utilise $E = 255$ avec une mantisse nulle pour $\pm\infty$ (overflow).

On utilise $E = 255$ avec une mantisse non nulle pour "Not a Number".

On utilise $E = 0$ avec une mantisse non nulle pour l'underflow.

A.2.6. La base 2, double précision. $b = 2$, $p = 53$, $e_{max} = 1023$, $e_{min} = -1022 = -e_{max} + 1$, $E = e + e_{max}$. Le stockage se fait sur 64 bits suivant le même principe que pour la simple précision, avec 1 bits de signe, puis 11 bits d'exposant, puis 52 bits de mantisse (x_0 n'est pas écrit).

Il existe aussi une notion de **double précision étendue** avec un stockage sur ≥ 80 bits.

A.2.7. La base 10. Les calculatrices utilisent en général la base 10 de manière à avoir un affichage facile. Les digits sont alors stockés en **D**écimal **C**odé **B**inaire. Chaque digit décimal est codé sur 4 bits par sa valeur en binaire. Ainsi 7 par exemple est codé 0111. On a de ce fait une petite perte (car sur 4 bits on pourrait stocker 16 symboles).

A.2.8. Taille du stockage. L'exposant e vérifie en général $-99 \leq e \leq 99$ et est stocké en DCB sur 1 octet. Les signes du nombre et de l'exposant sont stockés sur 1/2 octet.

Si on dispose de n octets la mantisse aura une longueur $p = 2n-3$ (par exemple si $n = 8$ on aurra 13 chiffres pour la mantisse). Attention il ne faut pas confondre p avec le nombre de chiffres affichés qui est souvent plus petit.

Exercice : Monter un calcul qui permette de déduire la véritable taille de la mantisse.

A.2.9. Arrondi. La norme IEEE définit quatre façons d'arrondir un résultat :

- Arrondi vers $+\infty$: x est arrondi au plus petit nombre représentable $\geq x$.
- Arrondi vers $-\infty$: x est arrondi au plus grand nombre représentable $\leq x$.
- Arrondi vers 0 : la valeur absolue de x est arrondie vers $-\infty$ (le signe est conservé).
- Arrondi au plus près : x est arrondi au nombre représentable le plus proche.

A.2.10. Arrondi correct. On choisit un mode d'arrondi. Soit f une fonction d'une ou de plusieurs variables réelles à valeurs réelles. On veut réaliser le calcul de f de telle manière que si les valeurs de $u = (u_1, \cdots, u_k)$ sont des nombres représentables alors la valeur approchée obtenue pour $f(u)$ soit l'arrondi de la vraie valeur de $f(u)$. (**Table Maker's dilemma**)

La norme IEEE impose l'arrondi correct pour les fonctions de base $+, -, \div, \times, \sqrt{}$ et les conversions. Ceci n'est pas imposé pour $\sin, \exp \cdots$.

A.2.11. Quelques difficultés. Supposons u, v, w des flottants. L'addition n'est pas associative : notons $F(x)$ le nombre flottant qui représente x.

$$F\left((u+v)+w\right) = F\left(F(u+v)+w\right),$$

$$F\left(u+(v+w)\right) = F\left(u+F(v+w)\right).$$

Or il se peut que $F(u+v) = u$ et $F(u+w) = u$ tandis que $F\left(u+F(v+w)\right) \neq u$. Par exemple si $b = 10$, $p = 11$ on prend $u = 1$, $v = w = 5 \times 10^{-11}$.

A.3. Les entiers

On veut représenter des nombres entiers en machine. Supposons par exemple que nous ayons **un octet** pour le faire. Avec un octet on dispose de $2^8 = 256$ écritures différentes, ces écritures pouvant être considérées comme les développements binaires des entiers de l'intervalle $\{0..255\}$.

Comme on veut répartir équitablement les entiers que l'on représente entre des entiers positifs et des entiers négatifs, on décide de s'intéresser aux 256 entiers de l'intervalle $\{-128..127\}$. Il convient donc d'établir une bijection qui permette des calculs commodes, entre l'intervalle $\{-128..127\}$ des entiers qu'on veut représenter, et l'intervalle $\{0..255\}$ des représentations.

En résumé, à tout entier x de l'intervalle $\{-128..127\}$ on va faire correspondre sa représentation $R(x)$ qui sera un entier de l'intervalle $\{0..255\}$. En outre comme $R(x)$ doit être stocké dans la mémoire d'une machine, on regardera plus spécialement les propriétés du développement binaire (sur un octet) de $R(x)$.

A.3.1. Représentation en "complément à 2". Rappelons que si n est un entier, $n \bmod 256$ est le reste de la division de n par 256 ou encore l'unique entier m tel que $0 \leq m \leq 255$ et m congru à n modulo 256.

Notons I l'intervalle $\{-128..127\}$ et J l'intervalle $\{0..255\}$. Soit R l'application de I dans J définie par

$$R(x) = x \bmod 256.$$

A.3.2. Premières propriétés. a) Montrer que R est une application bijective.

b) Calculer $R(0), R(100), R(127), R(-1), R(-100), R(-128)$. Donner les développements binaires des résultats obtenus.

c) Calculer $R(x)$ en fonction de x.

d) Déterminer l'image par R de l'ensemble des $x \geq 0$ de I ainsi que l'image par R de l'ensemble des $x < 0$ de I.

A.3.3. L'opposé. Comment reconnaître sur le développement binaire de $R(x)$ le signe de x ?

e) On suppose que $x \in I \setminus \{-128, 0\}$. Calculer $R(-x)$ en fonction de $R(x)$.

On constatera que $R(-x) = (255 - R(x)) + 1$. En déduire un algorithme simple permettant de calculer le développement binaire de $R(-x)$ à partir de celui de $R(x)$ (algorithme dit de complémentation à 2).

f) Pour écrire en binaire la représentation $R(x)$ d'un entier x de l'intervalle I on applique la stratégie suivante :

- Si $x \geq 0$, on développe x en binaire.
- Si $x < 0$, on développe $-x$ en binaire et on applique l'algorithme de complémentation à 2 (cf. e)).

Appliquer cette méthode pour calculer l'écriture binaire de $R(18)$, $R(-20)$.

A.3.4. Addition des entiers et représentation. Soit T l'application de $\mathbb{N}$ dans $\{0..255\}$ qui à tout $n = \sum_{j=0}^{\infty} a_j 2^j$ fait correspondre $T(n) = \sum_{j=0}^{7} a_j 2^j$ (troncature limitée aux 8 premiers bits).

a) Quel est le lien entre n mod 256 et $T(n)$?

b) Montrer que si $x_1, x_2, x_1 + x_2$ sont des éléments de I alors

$$R(x_1 + x_2) = \big(R(x_1) + R(x_2)\big) \bmod 256,$$

$$R(x_1 + x_2) = T\big(R(x_1) + R(x_2)\big).$$

c) Il est bien entendu que l'addition de deux éléments de I ne sera valide que si le résultat est aussi dans I. Comme la machine ne connaît que les représentants des nombres qu'on additionne, nous mettons en place ici une méthode (bien adaptée aux circuits électroniques) qui opère sur les représentations et permette à la fois de détecter si l'opération d'addition est valide et de calculer dans ce cas le résultat.

Soient x_1 et x_2 deux éléments de I. Soit C le carry, retenue du $8ieme$ bit vers l'exterieur, et α la retenue du $7ieme$ bit vers le $8ieme$, obtenus en faisant l'addition $R(x_1) + R(x_2)$. On pose $V = C \oplus \alpha$.

- Si $V = 0$ l'addition est valide et $R(x_1 + x_2)$ s'obtient en faisant en binaire l'addition de $R(x_1)$ avec $R(x_2)$ et en négligeant tout débordement au delà du huitième bit (cf. b)).
- Si $V = 1$ l'addition n'est pas valide.

En effet : c1) Supposons $0 \leq x_1 \leq 127$ et $0 \leq x_2 \leq 127$. Examiner les deux cas $x_1 + x_2 \leq 127$ (opération valide) et $x_1 + x_2 > 127$ (opération invalide), et dans chaque cas calculer C, α, V.

c2) Supposons $0 \leq x_1 \leq 127$ et $-128 \leq x_2 < 0$ (opération toujours valide). On examinera suivant les valeurs possibles de $R(x_1) + R(x_2)$ quelles sont les valeurs possibles de C, α, V.

c3) Supposons $-128 \leq x_1 < 0$ et $-128 \leq x_2 < 0$. Examiner les deux cas $x_1 + x_2 < -128$ (opération invalide) et $x_1 + x_2 \geq -128$ (opération valide), et dans chaque cas calculer C, α, V. (Indication : pour calculer α on pourra regarder si l'addition des deux nombres de 7 bits $(R(x_1) - 128)$ et $(R(x_2) - 128)$ a une retenue vers le huitième bit.)

c4) En conclure la validité de l'algorithme annoncé.