POLYTECH LYON

Université Claude Bernard Lyon 1

NOTES DE COURS DE l'UE MNBmater

Matériaux 3A

MÉTHODES NUMÉRIQUES DE BASE

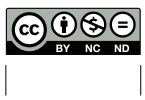
2024-2025, Automne

N. Débit & J. Bastien

Document compilé le 6 octobre 2024

Le lien original de ce document est le suivant :
http://utbmjb.chez-alice.fr/Polytech/MNBmater/coursMNBmater.pdf

Ce document est mis à disposition selon les termes de la licence Creative Commons : Paternité - Pas d'Utilisation Commerciale - Pas de Modification ; 3.0



http://creativecommons.org/licenses/by-nc-nd/3.0/
ou en français
http://creativecommons.org/licenses/by-nc-nd/3.0/deed.fr

Table des matières

Avant-propos v

Références conseillées vi

Avis de Recherche vii

Chapitre 0. Erreurs 1

Chapitre 1. Principes de l'analyse numérique (ou méthodes numériques) et approximations polynômiales de $\ln(1+x)$ et d
- 1.1. Introduction 2
- 1.2. Les principes de l'analyse numérique (ou méthodes numériques ou calcul scientifique) 2
- 1.3. Approximation de e^x (sous forme d'exercice corrigé) 5
- 1.4. Approximation de e^x 5
- 1.5. Approximation de $\ln(1+x)$ (sous forme d'un exercice facultatif corrigé) 10
- 1.6. Approximation de $\ln(1+x)$ 10
- 1.7. Simulations numériques sur les majorations des erreur absolue et relative 10

Chapitre 2. Interpolation 11
- 2.1. Motivation 11
- 2.2. Interpolation polynômiale 12
- 2.3. Un exercice type à savoir traiter parfaitement (Interpolation de Lagrange) 33
- 2.4. Interpolation de Tchebycheff (ou Chebyshev) 34
- 2.5. Interpolation par intervalles ou par morceaux (dite aussi interpolation composée ou composite) 36
- 2.6. Splines cubiques et autres courbes d'ajustement 39
- 2.7. Interpolation d'Hermite 40
- 2.8. Approximation au sens des moindres carrés 41

Chapitre 3. Intégration 46
- 3.1. Motivation 46
- 3.2. Introduction informelle 47
- 3.3. Méthodes élémentaires et composites (composées) 50
- 3.4. Formules de quadrature 68
- 3.5. Un exercice type à savoir traiter parfaitement 71

Chapitre 4. Équations non-linéaires 75
- 4.1. Motivation 75
- 4.2. Généralités 79
- 4.3. Méthode de bissection ou dichotomie 79
- 4.4. Méthode de point fixe 80
- 4.5. Méthode de Newton 98
- 4.6. Méthode de la sécante (ou de Lagrange) 102
- 4.7. Méthode de la corde et de la fausse position 103

4.8. Deux exercices types à savoir traiter parfaitement 103

Chapitre 5. Équations différentielles (ordinaires) 106
5.1. Motivations 106
5.2. Introduction et formalisme 107
5.3. Un peu de théorie 109
5.4. Schémas d'Euler progressif et rétrograde 109
5.5. Schémas de Runge Kutta 2 et 4 113
5.6. Équations différentielles d'ordres plus élevés 116
5.7. Retours sur les exemples introductifs 121
5.8. Un exercice type à savoir traiter parfaitement 122

Chapitre 6. Équations aux dérivées partielles 124
6.1. Dérivation numérique 124
6.2. Applications : résolution numérique d'équation aux dérivées partielles 124

Chapitre 7. Paradoxes 125

Annexe A. Compléments sur les approximations polynômiales de $\ln(1+x)$ et de e^x 126
A.1. Approximation de e^x par les séries 126
A.2. Approximation de $\ln(1+x)$ 129

Annexe B. Majoration de l'erreur relative 145
B.1. Principe théorique 145
B.2. Simulations numériques 147

Annexe C. Différents calculs de l'exemple 2.21 page 29 du cours 149

Annexe D. Étude théorique d'un problème de moindres carrés 151
D.1. Rappels sur la régression linéaire 151
D.2. Rappels sur la norme Euclidienne et le produit scalaire associé 152
D.3. Théorie 153

Annexe E. Définition et utilisation de la fonction W de Lambert 164
E.1. Définition de la fonction W de Lambert 164
E.2. Utilisation de la fonction W de Lambert : résolution de l'équation $ae^x + bx + c = 0$ 165
E.3. Utilisation de la fonction W de Lambert : résolution de l'équation $x^x = z$ 167

Annexe F. La primitive est l'opération inverse de la dérivation (sous forme d'exercice) 168
Énoncé 168
Corrigé 168

Annexe G. Calcul d'une intégrale "impropre" par la méthode de newton (sous forme d'exercice corrigé) 170
Énoncé de l'exercice 170
Corrigé de l'exercice 171

Annexe H. Formules d'intégration élémentaires à 0, 1 et 2 points 177

Annexe I. Formule d'intégration élémentaires à 4 points 179

Annexe J. Formules d'intégration élémentaires à 3 et 4 points et erreur associées (sous forme de problèmes corrigés) 181
Premier énoncé 181
Premier corrigé 182

Second énoncé 184
Second corrigé 185

Annexe K. Formules d'intégration élémentaires de Newton-Cotes (sous forme d'exercice corrigé) 187

Annexe L. Méthode de dichotomie ou de bissection (sous la forme d'un exercice corrigé) 194
Énoncé 194
Corrigé 195

Annexe M. Dichotomie discontinue 196
M.1. Introduction 196
M.2. Principe 196
M.3. Applications 200
M.4. Calcul de l en base 2 202
M.5. Simulations numériques 204

Annexe N. Convergence globale de la méthode du point fixe 211
N.1. Cas particuliers 211
N.2. Cas général 211

Annexe O. Étude de la convergence globale de la suite définie par $u_0 \in \mathbb{R}_+^*$ et $u_{n+1} = 1/2(u_n + A/u_n)$, sous la forme d'un
Énoncé 212
Corrigé 212

Annexe P. Étude du zéro de la fonction $e^x + x - K$, sous la forme d'un problème corrigé 217
Énoncé 217
Corrigé 220

Annexe Q. Un théorème de point fixe 241

Annexe R. Majoration de l'erreur pour une méthode d'ordre p 243

Annexe S. Convergence des méthodes d'ordre p 245

Annexe T. Compléments sur la divergence de la méthode du point fixe (sous forme d'un exercice corrigé) 249
Énoncé 249
Corrigé 249
Quelques remarques 256

Annexe U. Dégénérescence de la méthode de Newton et méthode de Newton modifiée 261

Annexe V. Étude et calcul de l tel que $l = \cos l$ sous la forme d'un problème corrigé 267
Énoncé 267
Corrigé 268

Annexe W. Exemple d'une méthode de Newton divergente partout 276
Énoncé 276
Corrigé 277

Annexe X. Racines multiples 283
X.1. Racines multiples d'un polynôme 283
X.2. Racines multiples d'une fonction quelconque 285

Annexe Y. Convergence globale de la méthode de Newton 287

Annexe Z. Étude de $x^{x^{x^{x^{\cdots}}}}$ (sous la forme d'un problème corrigé) 290
Énoncé 290
Corrigé 292

Annexe AA. Approximations de π 305
AA.1. Introduction 305
AA.2. Méthode d'Archimède 305
AA.3. Méthode de Cues 320
AA.4. Et la méthode originale des isopérimètres de Descartes ! 326
AA.5. Approximation quadratique par une méthode arithmético-géométrique 328
AA.6. Approximations d'ordres plus élevés 330
AA.7. Simulations numériques 330

Bibliographie 338

Avant-propos

Ce polycopié constitue les notes de cours de Méthodes Numériques de Base du département Matériaux 3A (2024-2025, Automne).

Ce polycopié de cours est normalement disponible à la fois

- en ligne sur `http://utbmjb.chez-alice.fr/Polytech/index.html` à la rubrique habituelle ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  — 'Poste de travail',
  — puis sur le répertoire '`P:`' (appelé aussi '\\`teraetu`\`Enseignants`'),
  — puis '`jerome.bastien`',
  — puis '`Polytech`',
  — puis '`Matériaux 3A`'.
  — enfin sur '`MNBmater`'.

Des notes en petits caractères comme suit pourront être omises en première lecture :
Attention, passage difficile ! ♢

Références conseillées

De nombreux résultats proviennent de l'ouvrage suivant :
J. Bastien et J.-N. Martin. *Introduction à l'analyse numérique. Applications sous Matlab*. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 519.4 BAS, 4e étage). Voir `https://www.dunod.com/sciences-techniques/introduction-analyse-numerique-applications-sous-matlab`. Paris : Dunod, 2003. 392 pages

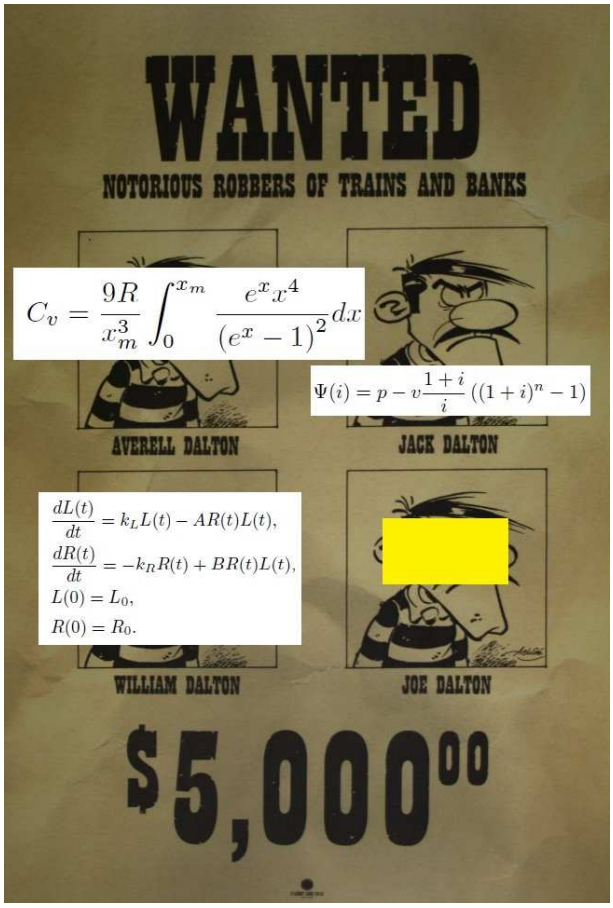
ils ne sont pas toujours rédigés dans ce polycopié, mais ils figurent dans cet ouvrage, disponible à la BU.

Une liste de référence de livres conseillés, en principe disponibles à la bibliothèque de Lyon I :

- M. Schatzmann. *Analyse numérique, une approche mathématique, Cours et exercices*. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 518.1 SCH, 4e étage). Dunod, 2001
- J. Nougier. *Méthodes de calculs numériques*. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : SCI 1351). Paris : Masson, 1993
- D. Conte et C. de Boor. *Elementary numerical analysis. An algorithmic approach*. Mc Graw-Hill, 1981
- S. D. Conte et C. de Boor. *Elementary numerical analysis*. Tome 78. Classics in Applied Mathematics. An algorithmic approach, Updated with MATLAB, Reprint of the third (1980) edition, For the 1965 edition see [MR0202267]. Society for Industrial et Applied Mathematics (SIAM), Philadelphia, PA, 2018, pages xxiv+456
- M. Crouzeix et A. L. Mignot. *Analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1984, pages viii+171
- M. Crouzeix et A. L. Mignot. *Exercices d'analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1986, page 183
- J. Baranger. *Introduction à l'analyse numérique*. Ouvrage disponible à la bibliothèque de Mathématiques de Lyon 1 (cote : 518.07 BAR, niveau Capes/Agreg). Hermann, Paris, 1977, pages vii+133
- F. Scheid. *Analyse numérique, Cours et problèmes*. Série Schaum. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 519.407 SCH, 4e étage). groupe Mc Graw-Hill, 1987

D'autres références sont aussi données à la page 338.

Avis de Recherche



Je recherche pour cette année des exemples d'illustration de ce cours issus du domaine de l'ingénierie des Matériaux :

— une ou plusieurs intégrales ;
— une équation non linéaires d'inconnue réelle (du type $\cos x = x$) ;
— une équation différentielle ordinaire.

Chapitre 0

Erreurs

Une fois n'est pas coutume ; nous essayerons de commencer ce cours par des présentations de petits problèmes simples de calculs. Nous verrons que les ordinateurs ne calculent pas si bien que ça ! La référence qui suit sera présentée de façon abrégée, lors du premier cours.

Voir [Bas14a] disponible sur `http://utbmjb.chez-alice.fr/INSA/zetetique/erreur_ordinateur.pdf`.

Les sources matlab de ces transparents, issus [BM03, chapitre 1] sont disponibles sur `http://utbmjb.chez-alice.fr/INSA/zetetique/zetetique_fichiersmatlab.zip` et pourront être par exemple, utilisées en TP.

Chapitre 1

Principes de l'analyse numérique (ou méthodes numériques) et approximations polynômiales de $\ln(1+x)$ et de e^x

1.1. Introduction

Dans ce chapitre, nous traitons un exemple simple d'analyse numérique sous forme d'exercice corrigé. Nous allons notamment démontrer

$$\forall x \in \mathbb{R}, \quad e^x = \sum_{n=0}^{+\infty} \frac{1}{n!} x^n, \tag{1.1a}$$

$$\forall x \in]-1, 1], \quad \ln(1+x) = \sum_{n=1}^{+\infty} (-1)^{n-1} \frac{x^n}{n}, \tag{1.1b}$$

et montrer que ces formules peuvent être simplement utilisées pour approcher numériquement les valeurs de e^x et de $\ln(1+x)$. Nous conclurons par quelques simulations numériques.

Nous utilisons dans ce chapitre, des notions simples sur les suites. Voir par exemple [Bas22a, chapitre 8] disponible sur `http://utbmjb.chez-alice.fr/Polytech/MFI/coursMFI.pdf`. Nous utilisons aussi des notions simples sur les série, notamment le fait qu'une série de terme général a_n pour $n \geq 0$ converge signifie qu'il existe un réel S tel que

$$\lim_{n \to +\infty} \sum_{k=0}^{n} a_k = S,$$

et nous poserons

$$S = \sum_{k=0}^{+\infty} a_k.$$

Plus de détail dans par exemple [Bas22a, chapitre 9].

Il existe sûrement des formules plus performantes pour approcher numériquement les valeurs de e^x et de $\ln(1+x)$, mais l'objet de ce chapitre est d'en proposer des versions simples.

1.2. Les principes de l'analyse numérique (ou méthodes numériques ou calcul scientifique)

(1) Les méthodes numériques ont pour objet de remplacer le calcul d'une quantité l, réelle, voire complexe, en principe non calculable ou difficilement calculable, par la détermination d'une suite (u_n) pour $n \geq n_0$ où $n_0 \in \mathbb{N}$, convergeant vers ce nombre l de sorte que :
   - Le calcul de u_n soit simple. Il sera souvent donné sous la forme d'un algorithme, lui-même implémentable informatiquement.
   - La suite (u_n) converge vers le réel l, soit

$$\lim_{n \to +\infty} u_n = l. \tag{1.2}$$

On pourra consulter par exemple [Bas22a, Chapitre "Suites"].

— On puisse "gérer" l'erreur, c'est-à-dire, déterminer, un majorant ε_n de l'erreur absolue η_n définie par

$$\forall n \geq n_0, \quad \eta_n = |u_n - l|, \tag{1.3}$$

qui vérifie donc

$$\forall n \geq n_0, \quad \eta_n \leq \varepsilon_n\,, \tag{1.4}$$

et

$$\lim_{n\to+\infty} \varepsilon_n = 0. \tag{1.5}$$

— Ainsi, le nombre l sera remplacé par son approximation u_n. L'erreur absolue commise, η_n sera donc, selon (1.4) majoré par le nombre connu ε_n. Selon (1.5), cette erreur pourra choisie aussi petite que l'on veut, à partir du moment où n est assez grand. Nous dirons donc pour résumer (1.3), (1.4) et (1.5) que

u_n constitue une approximation de l avec une erreur inférieure à ε_n,

qui tend vers zéro quand n tend vers l'infini. (1.6)

— De façon plus précise, on essayera de déterminer, si possible, pour tout $\varepsilon > 0$, un nombre $N \geq n_0$ tel que

$$\varepsilon_N \leq \varepsilon. \tag{1.7}$$

D'après (1.5), cet entier N existe. D'après (1.4), l'erreur absolue η_n sera majorée par ε.

Cette approximation d'un nombre l sera par exemple utilisée dans le cas de la détermination de la valeur approché l :

— d'une fonction en un point, cette fonction étant connue en d'autres valeurs (voir chapitre 2 page 11) ;
— d'une intégrale d'une fonction connue (voir chapitre 3 page 46) ;
— d'une solution d'une équation non linéaire, par exemple vérifiant $l = \cos l$ (voir chapitre 4 page 75) ;
— d'une fonction en un point, cette fonction étant la solution d'une équation différentielle ordinaire (voir chapitre 5 page 106) ou d'une équation aux dérivées partielles (voir chapitre 6 page 124).

On s'intéressera aussi dans le cas où $l \neq 0$ à déterminer, un majorant $\widetilde{\varepsilon}_n$ de l'erreur relative $\widetilde{\eta}_n$ définie par

$$\forall n \geq n_0, \quad \widetilde{\eta}_n = \left|\frac{u_n - l}{l}\right|, \tag{1.8}$$

qui vérifie donc les équivalents de (1.4) et (1.5), soit

$$\forall n \geq n_0, \quad \widetilde{\eta}_n \leq \widetilde{\varepsilon}_n, \tag{1.9}$$

$$\lim_{n\to+\infty} \widetilde{\varepsilon}_n = 0. \tag{1.10}$$

L'erreur relative commise, $\widetilde{\eta}_n$ sera donc, selon (1.9) majorée par le nombre connu $\widetilde{\varepsilon}_n$.

La quantité $\widetilde{\eta}_n$ n'est pas calculable en pratique car l n'est pas connue. Mais on peut utiliser une petite astuce pour déterminer le plus petit entier N vérifiant l'équivalent de (1.7) :

$$\widetilde{\varepsilon}_N \leq \varepsilon. \tag{1.11}$$

voir l'annexe B page 145.

Remarque 1.1. De façon plus générale, si une expression (réelle) X_{exacte} est connue de façon approchée avec la valeur $X_{\text{approchée}}$, l'erreur absolue est définie par une expression proche de l'expression de η_n définie par (1.3) :

$$\varepsilon_{\text{absolue}} = |X_{\text{exacte}} - X_{\text{approchée}}|\,, \tag{1.12a}$$

tandis que l'erreur relative (si bien sûr, X_{exacte} est non nulle !) est définie par une expression proche de l'expression de $\widetilde{\eta}_n$ définie par (1.8) :

$$\varepsilon_{\text{relative}} = \left| \frac{X_{\text{exacte}} - X_{\text{approchée}}}{X_{\text{exacte}}} \right|. \tag{1.12b}$$

Ces deux expressions (1.12) sont à rapprochées des notions d'incertitudes absolues et relatives vues par exemples dans [DB22, sections 2.5.2. et 2.5.3.].

Remarque 1.2. Les équations (1.3), (1.8) et (1.12) et les suivantes sont aussi valable pour des grandeurs ou des suites à valeurs dans $\mathbb{C}$, dans $\mathbb{R}^N$ où N est un entier ou même dans des espaces vectoriels de dimensions finies ou même infinies normés, à condition de remplacer les valeurs absolues $|.|$ par des modules ou des normes.

$\diamondsuit$

(2) Dans le cas où la suite est réelle, on peut aussi déterminer deux suites $(m_n)_{n\geq n_0}$ et $(M_n)_{n\geq n_0}$ vérifiant

$$\forall n \geq n_0, \quad m_n \leq l \leq M_n \tag{1.13}$$

et

$$\lim_{n\to+\infty} M_n - m_n = 0. \tag{1.14}$$

m_n et M_n sont respectivement les approximations par défaut et par excès du nombre l. On a aussi

$$\lim_{n\to+\infty} M_n = \lim_{n\to+\infty} m_n = l. \tag{1.15}$$

En effet, on a, d'après (1.13)

$$0 \leq l - m_n \leq M_n - m_n,$$

qui tend vers zéro. De même, d'après (1.13)

$$0 \leq M_n - l \leq M_n - m_n,$$

qui tend vers zéro. On en déduit aussi que

$$\max\left(|l - m_n|, |l - M_n|\right) \leq M_n - m_n, \tag{1.16}$$

ce qui assure que les écarts entre entre l et m_n et l et M_n sont majorés par $M_n - m_n$. Nous dirons donc pour résumer (1.13) et (1.14) que

m_n et M_n constituent respectivement deux approximations par défaut et par excès de u_n
avec une erreur inférieure à $M_n - m_n$ qui tend vers zéro quand n tend vers l'infini. (1.17)

Remarque 1.3. Notons que les propriétés (1.3), (1.4) et (1.5) vues dans le cas 1 sont équivalentes aux propriétés (1.13) et (1.14) vues dans le cas 2.

Démontrons-le.

(1) Si (1.3), (1.4) et (1.5) ont lieu, on remarque que (pour $n \geq n_0$)

$$-|l - u_n| \leq l - u_n \leq |l - u_n|$$

et donc

$$u_n - |l - u_n| \leq l \leq u_n + |l - u_n|. \tag{1.18}$$

On pose donc (pour $n \geq n_0$)

$$m_n = u_n - |l - u_n|,$$
$$M_n = u_n + |l - u_n|.$$

Ainsi (1.13) est valable d'après (1.18). et on a

$$M_n - m_n = 2|l - u_n|$$

qui tend vers zéro selon (1.3) (1.4) et (1.5) et donc (1.14) est vérifiée.

(2) Réciproquement supposons que (1.13) et (1.14) aient lieu. Posons (pour $n \geq n_0$)

$$u_n = \frac{1}{2}(m_n + M_n), \tag{1.19a}$$

$$\varepsilon_n = M_n - m_m. \tag{1.19b}$$

D'après (1.13) et (1.19a) u_n et l sont tous les deux dans l'intervalle $[m_n, M_n]$ et donc

$$|u_n - l| \leq M_n - m_m = \varepsilon_n,$$

ce qui implique (1.3), (1.4) et (1.5), d'après (1.14).

♢

1.3. Approximation de e^x (sous forme d'exercice corrigé)

Nous donnons un exercice dont le corrigé se trouvera en section 1.4.1.

(1) Soient $n \in \mathbb{N}$ et $x \in \mathbb{R}$. Appliquer la formule de Taylor-Lagrange à l'ordre n à la fonction e^x sur l'intervalle $[0, x]$ (ou $[x, 0]$) et en déduire que

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad e^x = p_n(x) + R_n(x).$$

où $p_n(x)$ est un polynôme de degré n en x et $R_n(x)$ une expression à déterminer.

(2) En distinguant les cas $x < 0$ et $x > 0$, obtenir deux expressions polynômiales d'un minorant et d'un majorant de $e^x - p_n(x)$.

(3) Conclure sur une expression

(a) d'une approximation de e^x et la majoration d'erreur commise.

(b) de deux approximations de e^x par défaut et par excès et la majoration d'erreur commise.

1.4. Approximation de e^x

1.4.1. Par les formules de Taylor-Lagrange

Montrons le résultat suivant :

Proposition 1.4. *On pose, pour tout $n \in \mathbb{N}^*$,*

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}. \tag{1.20}$$

On a

$$\forall x \in \mathbb{R}, \quad \exists N(x) \in \mathbb{N}^*, \quad \forall n \geq N(x), \quad 1 - \frac{x^{n+1}}{(n+1)!} > 0. \tag{1.21}$$

On pose

(1) Si $x \geq 0$

$$\forall n \in \mathbb{N}, \quad a_n = \frac{x^{n+1}}{(n+1)!}, \tag{1.22a}$$

$$\forall n \geq N(x), \quad b_n = \left(\left(1 - \frac{x^{n+1}}{(n+1)!} \right)^{-1} - 1 \right) p_n(x), \tag{1.22b}$$

(2) Si $x \leq 0$

(a) Si n impair

(i) Si $p_n(x) \leq 0$,

$$\forall n \in \mathbb{N}, \quad a_n = 0, \tag{1.22c}$$

$$\forall n \in \mathbb{N}, \quad b_n = \frac{x^{n+1}}{(n+1)!}, \tag{1.22d}$$

(ii) Si $p_n(x) > 0$,

$$\forall n \geq N(x), \quad a_n = \left(\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} - 1\right) p_n(x), \tag{1.22e}$$

$$\forall n \in \mathbb{N}, \quad b_n = \frac{x^{n+1}}{(n+1)!}, \tag{1.22f}$$

(b) Si n pair

$$\forall n \in \mathbb{N}, \quad a_n = \frac{x^{n+1}}{(n+1)!}, \tag{1.22g}$$

$$\forall n \in \mathbb{N}, \quad b_n = \left(\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} - 1\right) p_n(x), \tag{1.22h}$$

On a alors

$$\forall n \geq N(x), \quad e^x - p_n(x) \in [a_n, b_n], \tag{1.23a}$$

et en posant $\varepsilon_n = b_n - a_n \geq 0$, on a

$$\lim_{n \to +\infty} \varepsilon_n = 0. \tag{1.23b}$$

En d'autres termes, pour tout $n \in \mathbb{N}^*$, pour tout $x \in \mathbb{R}$,

(1) $p_n(x)$ constitue une approximation de e^x avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini,

(2) $p_n(x) + a_n$ et $p_n(x) + b_n$ constituent respectivement deux approximations par défaut et par excès de e^x avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini.

En d'autres termes encore, le cas 1 correspond à (1.6) *du cas 1 page 2 et le cas 2 correspond à* (1.17) *du cas 2 page 4.*

On retrouve donc d'une part le développement limité usuel de e^x par exemple de [Bas22b, Annexe "Quelques développements limités usuels"] et d'autre part le résultat habituel sur les séries (voir par exemple [Bas22a, Chapitre "Séries"]) .

Remarque 1.5. Notons que si x est rationnel, la proposition 1.4 propose une approximation rationnelle de e^x.

Démonstration de la proposition 1.4. On s'appuiera pour montrer ce résultat sur la formule de Taylor-Lagrange.

(1) Considérons la fonction f définie par

$$\forall x \in \mathbb{R}, \quad f(x) = e^x. \tag{1.24}$$

On a immédiatement

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad f^{(n)}(x) = e^x \tag{1.25}$$

On a alors

$$\forall n \in \mathbb{N}^*, \quad f^{(n)}(0) = 1. \tag{1.26}$$

La formule de Taylor-Lagrange (voir par exemple [Bas22b, Chapitre "Dérivée, différentiation"]) à l'ordre n appliquée à la fonction f sur l'intervalle $[0, x]$ (ou $[x, 0]$) fournit

$$\forall x \in \mathbb{R}, \quad e^x = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k + \frac{1}{(n+1)!} f^{(n+1)}(\xi) x^{n+1},$$

où

$$\xi \in \begin{cases}]0, x[, & \text{si } x \geq 0, \\ ]x, 0[, & \text{si } x \leq 0. \end{cases} \tag{1.27}$$

et donc, pour $n \in \mathbb{N}^*$, d'après (1.26),

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad e^x = \sum_{k=0}^{n} \frac{x^k}{k!} + \frac{x^{n+1}}{(n+1)!} e^{\xi}. \tag{1.28}$$

Pour toute la suite, on pose

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}, \tag{1.29a}$$

$$R_n(x) = \frac{x^{n+1}}{(n+1)!} e^{\xi}, \tag{1.29b}$$

de sorte que (1.28) s'écrit

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad e^x = p_n(x) + R_n(x). \tag{1.30}$$

Il ne reste plus qu'à étudier la suite $R_n(x)$, selon les différentes valeurs de x.

(2) (a) Premier cas : $x \in \mathbb{R}_+$.

D'après (1.27), on a

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}_+, \quad \frac{x^{n+1}}{(n+1)!} \leq R_n(x) \leq \frac{x^{n+1}}{(n+1)!} e^x. \tag{1.31}$$

On a

$$\lim_{n \to +\infty} \frac{x^{n+1}}{(n+1)!} = 0 \tag{1.32}$$

puisque $\frac{x^{n+1}}{(n+1)!}$ est le terme général d'une série convergente. (voir par exemple [Bas22a, Chapitre "Séries"] et en particulier la règle de d'Alembert [1]). Cela suffit pour montrer que l'on a

$$\lim_{n \to +\infty} R_n(x) = 0, \tag{1.33}$$

et donc

$$\forall x \in \mathbb{R}_+, \quad \lim_{n \to +\infty} p_n(x) = e^x, \tag{1.34}$$

et donc (1.1a). Cependant cela ne permet pas de majorer $R_n(x)$ indépendamment de x ! Pour cela, on écrit d'après (1.30) et (1.31),

$$\frac{x^{n+1}}{(n+1)!} \leq e^x - p_n(x) \leq \frac{x^{n+1}}{(n+1)!} e^x$$

et donc d'une part

$$\frac{x^{n+1}}{(n+1)!} \leq e^x - p_n(x) \tag{1.35}$$

et d'autre part

$$\left(1 - \frac{x^{n+1}}{(n+1)!}\right) e^x \leq p_n(x). \tag{1.36}$$

1. Le lecteur vérifiera qu'il n'y a pas de cercle vicieux dans ce raisonnement !

D'après (1.32), on a (1.21). Donc, d'après (1.21) et (1.36), on a

$$\forall x \in \mathbb{R}_+, \quad \forall n \geq N(x), \quad e^x \leq p_n(x)\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1}. \tag{1.37}$$

et donc

$$\forall x \in \mathbb{R}_+, \quad \forall n \geq N(x), \quad e^x - p_n(x) \leq \left(\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} - 1\right) p_n(x). \tag{1.38}$$

L'équation (1.23a) est une conséquence de (1.35) et de (1.38), en considérant a_n et b_n définis par (1.22a) et (1.22b). On vérifie facilement que $0 \leq a_n \leq b_n$ ce qui implique

$$R_n(x) \geq 0. \tag{1.39}$$

Enfin, on a d'après (1.32) et (1.34)

$$b_n - a_n = \left(\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} - 1\right) p_n(x) - \frac{x^{n+1}}{(n+1)!} \to 0, \tag{1.40}$$

(i) D'après (1.23a) et (1.40), on a

$$|e^x - p_n(x)| \leq b_n - a_n$$

et donc $p_n(x)$ constitue une approximation de e^x avec une erreur inférieure à $b_n - a_n$, ce qui permet de montrer le point 1 page 6.

(ii) Posons $m_n = p_n(x) + a_n$ et $M_n = p_n(x) + b_n$. On a donc, d'après (1.23a)

$$\forall n \geq N(x), \quad m_n \leq e^x \leq M_n, \tag{1.41}$$

et

$$\forall n \geq N(x), \quad M_n - m_n = b_n - a_n. \tag{1.42}$$

D'après (1.23a) et le point 2 page 4 appliqué à $l = e^x$, (1.40) (1.41) (1.42) impliquent que m_n et M_n tendent vers e^x et constituent des approximations par défaut et par excès de e^x, avec une erreur inférieur à $b_n - a_n$, ce qui permet de conclure en montrant le point 2 page 6.

(b) Second cas : $x \in \mathbb{R}_-$.

Il nous faut discuter cette fois-ci sur la parité de n.

(i) Si n est impair, on a

$$x^{n+1} \geq 0. \tag{1.43}$$

On a d'après (1.27) $e^x \leq e^\xi \leq 1$ et donc d'après (1.43)

$$\frac{x^{n+1}}{(n+1)!} e^x \leq \frac{x^{n+1}}{(n+1)!} \xi^{n+1} \leq \frac{x^{n+1}}{(n+1)!}$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}_-, \quad \frac{x^{n+1}}{(n+1)!} \geq R_n(x) \geq \frac{x^{n+1}}{(n+1)!} e^x, \tag{1.44}$$

ce qui est l'analogue de (1.31). Cela montre (1.34) ainsi que, d'une part,

$$\frac{x^{n+1}}{(n+1)!} \geq e^x - p_n(x) \tag{1.45}$$

et d'autre part

$$\left(1 - \frac{x^{n+1}}{(n+1)!}\right) e^x \geq p_n(x). \tag{1.46}$$

qui sont les analogues de (1.35) et (1.36).

- Si
$$p_n(x) \leq 0. \tag{1.47}$$
On écrit d'après (1.44)
$$\frac{x^{n+1}}{(n+1)!} \geq R_n(x) \geq 0,$$
et donc
$$0 \leq e^x - p_n(x). \tag{1.48}$$
- Si
$$p_n(x) > 0, \tag{1.49}$$
on a d'après (1.46), et d'après (1.21),
$$\forall n \geq N(x), \quad e^x \geq \left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} p_n(x). \tag{1.50}$$
qui est l'analogue de (1.37). Dans ce cas, (1.50) implique
$$\forall n \geq N(x), \quad e^x - p_n(x) \geq \left(\left(1 - \frac{x^{n+1}}{(n+1)!}\right)^{-1} - 1\right) p_n(x). \tag{1.51}$$
qui est l'analogue de (1.38). Comme pour les équations (1.35) et de (1.38), on dispose des minorations de $e^x - p_n(x)$ (1.48) ou (1.51) et de la majoration (1.45). On conclut comme dans le cas 2a page 7, en posant a_n et b_n définis par (1.22c), (1.22d) ou (1.22e), (1.22f). On vérifie facilement que $0 \leq a_n \leq b_n$ ce qui implique
$$R_n(x) \geq 0. \tag{1.52}$$
Puis on conclut de la même façon que dans le cas 2a page 7 en remarquant que dans ce cas encore a_n et b_n tendent aussi vers zéro.

(ii) Si n est pair, on a
$$x^{n+1} \leq 0, \tag{1.53}$$
et on a cette fois-ci, puisque $e^x \leq e^\xi \leq 1$
$$\frac{x^{n+1}}{(n+1)!} e^x \geq \frac{x^{n+1}}{(n+1)!} \xi^{n+1} \geq \frac{x^{n+1}}{(n+1)!}$$
et donc
$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}_-, \quad \frac{x^{n+1}}{(n+1)!} \leq R_n(x) \leq \frac{x^{n+1}}{(n+1)!} e^x,$$
ce qui est exactement (1.31). On finit donc exactement comme dans le cas 2a page 7, à la différence près que $\left(1 - \frac{x^{n+1}}{(n+1)!}\right)$ est toujours strictement positif et on peut donc considérer a_n et b_n définis par (1.22g) et (1.22h). On vérifie facilement que $b_n \leq a_n \leq 0$ ce qui implique
$$R_n(x) \leq 0. \tag{1.54}$$

□

1.4.2. Par les séries

Cette section facultative est donnée dans la section A.1.1 page 126 de l'annexe A.

1.4.3. Simulations numériques

Voir section A.1.2 page 128 de l'annexe A.

1.5. Approximation de $\ln(1+x)$ (sous forme d'un exercice facultatif corrigé)

Nous donnons un exercice dont le corrigé se trouvera en section 1.6.1.

(1) Soient $n \in \mathbb{N}$ et $x \in]-1,+\infty[$. Appliquer la formule de Taylor-Lagrange à l'ordre n à la fonction $\ln(1+x)$ sur l'intervalle et en déduire que

$$\forall n \in \mathbb{N}^*, \quad \forall x \in \mathbb{R}, \quad \ln(1+x) = p_n(x) + R_n(x).$$

où $p_n(x)$ est un polynôme de degré n en x et $R_n(x)$ une expression à déterminer.

(2) (a) En supposant d'abord que $x \in [0,1]$, obtenir deux expressions polynômiales d'un minorant et d'un majorant de $\ln(1+x) - p_n(x)$.

(b) Faire la même chose pour $x \in]-1/2, 0]$.

(c) Pour $x \in]-1,-1/2]$, on procèdera autrement : on intégrera la somme géométrique $1+t+...+t^{n-1}$.

(3) (a) Conclure, pour tout $x \in]-1,1]$, sur une expression

(i) d'une approximation de $\ln(1+x)$ et la majoration d'erreur commise.

(ii) de deux approximations de $\ln(1+x)$ par défaut et par excès et la majoration d'erreur commise.

(b) Que se passe-t-il quand $x = -1$ ou x appartient à l'intervalle $]1,+\infty[$?

1.6. Approximation de $\ln(1+x)$

1.6.1. Par les formules de Taylor-Lagrange

Cette section facultative est donnée dans la section A.2.1 page 129 de l'annexe A.

1.6.2. Par les séries

Cette section facultative est donnée dans la section A.2.2 page 137 de l'annexe A.

1.6.3. Simulations numériques

Voir section A.2.3 page 141 de l'annexe A.

1.7. Simulations numériques sur les majorations des erreur absolue et relative

Voir la section B.2 page 147 de l'annexe B.

Chapitre 2

Interpolation

1

2.1. Motivation

Latitude	$K = 0.67$	$K = 1.5$	$K = 2$
65	$-3,106$	$3,520$	$6,058$
55	$-3,228$	$3,621$	$6,055$
45	$-3,309$	$3,652$	$5,922$
35	$-3,327$	$3,522$	$5,707$
25	$-3,172$	$3,476$	$5,308$
15	$-3,074$	$3,253$	$5,020$
5	$-3,029$	$3,152$	$4,957$
-5	$-3,029$	$3,150$	$4,974$
-15	$-3,124$	$3,207$	$5,078$
-25	$-3,209$	$3,274$	$5,355$
-35	$-3,350$	$3,529$	$5,627$
-45	$-3,373$	$3,705$	$5,954$
-55	$-3,258$	$3,704$	$6,103$

TABLE 2.1. Température de l'air à proximité du sol en fonction de la latitude pour différentes valeurs de concentration K.

Il est connu que la température de l'air à proximité du sol varie selon la concentration de l'acide carbonique. Le tableau 2.1 (extrait de "Philosophical Magazine" 41,237) montre les variations annuelles de la température moyenne pour différentes latitudes en fonction de la concentration (relative) K de l'acide carbonique dans l'atmosphère.

On cherche une estimation de la variation de la température à la latitude $42,961\,111$ de Centuri (Haute-Corse) pour une concentration K de l'acide carbonique dans l'atmosphère égale à $K = 0.67$. Les techniques d'interpolation permettent de reconstruire, à partir des données disponibles, les valeurs de température pour des latitudes ou des concentrations non contenues dans le tableau.

Sur la figure 2.1 page suivante, ont été tracées les différentes courbes déterminées grâce aux techniques de ce chapitre, ainsi que plusieurs estimations de la variation de la température à la latitude $42,961\,111$ qui se tiennent toutes dans un mouchoir de poche. On obtient en effet

- $\tau = -3.333$, en utilisant un polynôme de degré 12 (voir section 2.2) ;
- $\tau = -3.321$, en utilisant un polynôme de degré 3 (voir section 2.2) ;
- $\tau = -3.323$, en utilisant une spline cubique (voir section 2.6.1) ;

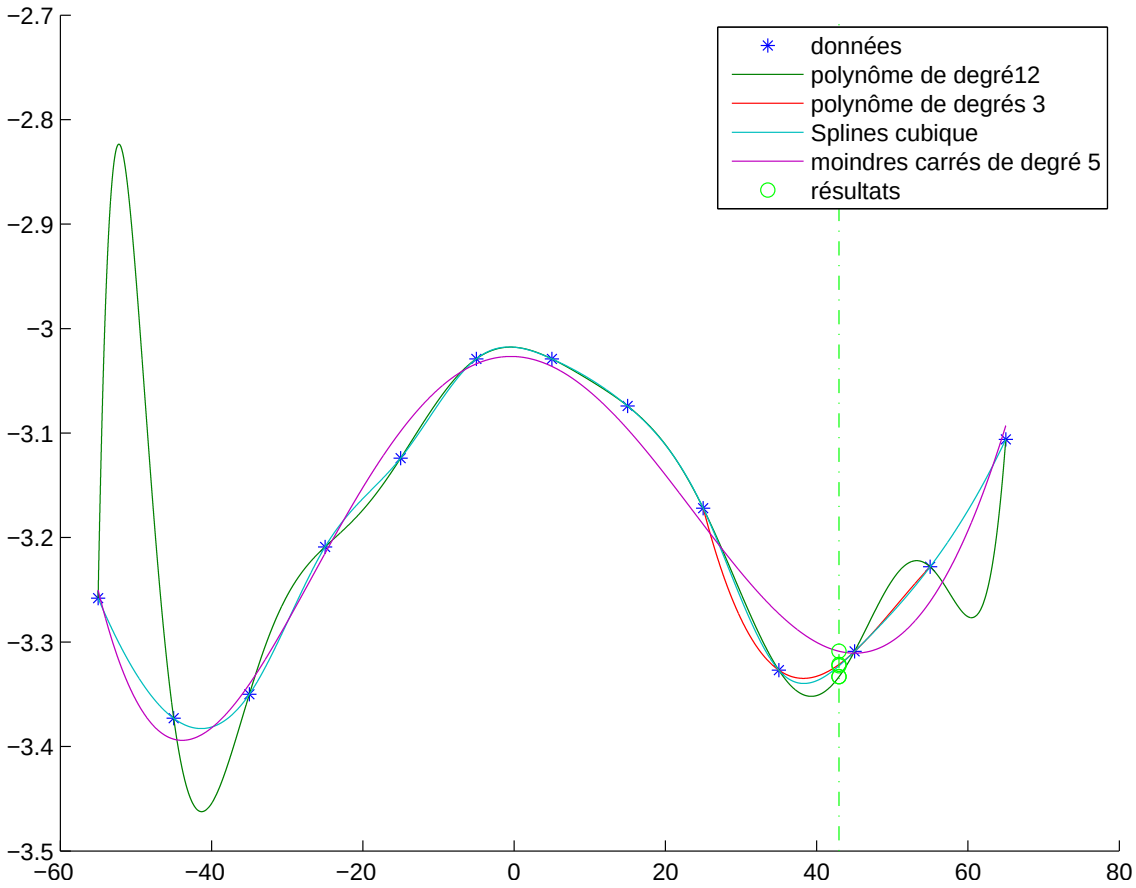


FIGURE 2.1. Les données et les différentes courbes déterminées grâce aux techniques de ce chapitre (pour une concentration K de l'acide carbonique dans l'atmosphère égale à $K = 0.67$).

- $\tau = -3.309$, en utilisant une approximation au sens des moindres carrés (voir section 2.8).

2.2. Interpolation polynômiale

2.2.1. Exemple et position du problème

EXEMPLE 2.1. Pour tout ce chapitre, on considérera les données suivantes : on connaît les valeurs d'une fonction f aux points $x_0 = 1$, $x_1 = 4$, $x_2 = 2$ et $x_3 = -1$:

$$f(x_0) = -2, \quad f(x_1) = 3, \quad f(x_2) = 1, \quad f(x_3) = 5. \tag{2.1}$$

On posera

$$\forall i \in \{0, 1, 2, 3\}, \quad y_i = f(x_i). \tag{2.2}$$

De façon plus générale, on se donne $n \geq 0$ un nombre entier. Étant donnés $n + 1$ points, deux à deux distincts, $x_0, x_1, \ldots, x_n$:

$$\forall i, j \in \{0, ..., n\}, \quad i \neq j \Longrightarrow x_i \neq x_j, \tag{2.3}$$

et $n + 1$ valeurs quelconques $y_0, y_1, \ldots, y_n$, on cherche un polynôme p de degré au plus n, tel que

$$\forall i \in \{0, ..., n\}, \quad p(x_i) = y_i. \tag{2.4}$$

On note $p = \Pi_n$, le polynôme d'interpolation aux points $(x_i)_{0\leq i\leq n}$ qui vérifie donc

$$\forall i \in \{0, ..., n\}, \quad \Pi_n(x_i) = y_i. \tag{2.5}$$

Les points (x_i, y_i) sont dits points d'interpolation. Les points x_i sont aussi appelés les nœuds du support $\{x_0, ..., x_n\}$. Si, comme dans l'exemple 2.1, les y_i correspondent aux valeurs d'une la fonction f, on dira que le polynôme Π_n interpole f sur le support $\{x_0, ..., x_n\}$. Dans ce cas, on notera parfois le polynôme sous la forme $\Pi_n(f)$.

Exemple 2.2. Pour $n = 1$, l'interpolation polynômiale est un problème déjà bien connu :

On cherche à déterminer l'équation de la droite passant par les points distincts A et B, de coordonnées respectives (x_A, y_A) et (x_B, y_B).

- Si $x_A = x_B$, la droite est "verticale" d'équation

$$x = x_A. \tag{2.6}$$

- Si $x_A \neq x_B$, on est exactement dans le cadre ci-dessus en posant

$$x_0 = x_A \text{ et } x_1 = x_B, \tag{2.7a}$$

$$y_0 = y_A \text{ et } y_1 = y_A. \tag{2.7b}$$

La droite est d'équation $Y = \alpha X + \beta$ et les égalités traduisant que la droite passe par A et B sont

$$\alpha x_A + \beta = y_A, \tag{2.8a}$$

$$\alpha x_B + \beta = y_B, \tag{2.8b}$$

système linéaire qui se résoud aisément et fournit, par différence des deux équations,

$$\alpha(x_A - x_B) = y_A - y_B,$$

et donc

$$\alpha = \frac{y_B - y_A}{x_B - x_A},$$

et si on réutilise l'équation (2.8a) on obtient

$$\beta = y_A - x_A \frac{y_B - y_A}{x_B - x_A}.$$

L'équation de la droite est donc

$$y = \frac{y_B - y_A}{x_B - x_A} x + y_A - x_A \frac{y_B - y_A}{x_B - x_A}. \tag{2.9}$$

soit

$$y = \frac{y_B - y_A}{x_B - x_A}(x - x_A) + y_A, \tag{2.10}$$

que l'on pourra écrire sous la forme suivante :

$$\frac{y - y_A}{x - x_A} = \frac{y_B - y_A}{x_B - x_A} \tag{2.11}$$

qui traduit que le taux d'accroissement entre x et x_A est égal au d'accroissement entre x_B et x_A.

Nous reviendrons sur cet exemple dans l'exemple 2.21 page 29.

♢

2.2.2. Construction de Π_n

2.2.2.1. Calcul direct (Matrice de Vandermonde).

Bien que la plus directe, cette technique est rarement utilisée en pratique à cause des instabilités numériques.

Donnons néanmoins le principe de ce calcul. Cherchons le polynôme Π_n sous la forme (canonique)

$$\Pi_n(x) = \alpha_0 + \alpha_1 x + ... + \alpha_n x^n = \sum_{k=0}^{n} \alpha_k x^k. \tag{2.12}$$

Si on écrit chacune des équations (2.5), on a donc

$$\forall i \in \{0, ..., n\}, \quad \alpha_0 + \alpha_1 x_i + ... + \alpha_n x_i^n = \sum_{k=0}^{n} \alpha_k x_i^k = y_i, \tag{2.13}$$

ce qui est équivalent au système linéaire

$$AX = B, \tag{2.14}$$

où la matrice $A \in \mathcal{M}_{n+1}(\mathbb{R})$ est définie par

$$\forall i, j, \in \{1, ..., n+1\}, \quad A_{ij} = x_{i-1}^{j-1}, \tag{2.15}$$

soit encore

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n & x_0^n, \\ 1 & x_1 & x_1^2 & \dots & x_1^{n-1} & x_1^n, \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} & x_2^n, \\ & \vdots & & \dots & \vdots & , \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} & x_n^n, \end{pmatrix} \tag{2.16}$$

et les vecteurs X et B sont les vecteurs colonnes de $\mathbb{R}^{n+1}$ donnés par

$$X = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}, \quad B = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{2.17}$$

La matrice A correspond à la matrice de Vandermonde, associée aux nœuds x_0, ..., x_n.

On peut remarquer que ce calcul est une généralisation de l'exemple 2.2 page précédente.

On sait qu'elle est inversible. Voir par exemple [BM03, Exercice 2.5 p. 55]. On peut en calculer à la main son inverse, en utilisant la théorie de l'interpolation ! L'inversibilité de cette matrice sera aussi une conséquence de la section 2.2.2.2. ♢

On a donc l'expression théorique de X d'après (2.14) :

$$X = A^{-1}B. \tag{2.18}$$

Exemple 2.3. Si on traite de cette façon l'exemple 2.1, on obtient successivement :

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 16 & 64 \\ 1 & 2 & 4 & 8 \\ 1 & -1 & 1 & -1 \end{pmatrix}, \tag{2.19a}$$

$$B = \begin{pmatrix} -2 \\ 3 \\ 1 \\ 5 \end{pmatrix}, \tag{2.19b}$$

$$X = \begin{pmatrix} -9/5 \\ -\frac{44}{15} \\ \frac{33}{10} \\ -\frac{17}{30} \end{pmatrix}, \tag{2.19c}$$

et donc

$$\Pi_3(x) = -\frac{17}{30}\,x^3 + \frac{33}{10}\,x^2 - \frac{44}{15}\,x - 9/5. \tag{2.19d}$$

Si on évalue ce polynômes aux points du support x_0,...,x_n, on obtient bien les y_i :

$$\begin{pmatrix} -2 & 3 & 1 & 5 \end{pmatrix}. \tag{2.19e}$$

2.2.2.2. Base de Lagrange.

Nous avons le lemme suivant :

LEMME 2.4. *Soit $n \in \mathbb{N}$. Pour tout $i \in \{0, ..., n\}$, il existe un unique polynôme l_i de degré n tel que*

$$\forall j \in \{0, ..., n\} \setminus \{i\}, \quad l_i(x_j) = 0, \tag{2.20a}$$

$$l_i(x_i) = 1, \tag{2.20b}$$

soit encore

$$\forall j \in \{0, ..., n\}, \quad l_i(x_j) = \delta_{ij}, \tag{2.21}$$

où δ_{ij} est le symbole de Kronecker[1]*. Pour tout $i \in \{0, ..., n\}$, l_i est donné par*

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{2.23}$$

Les polynômes $(l_i)_{0 \leq i \leq n}$ sont appelés les polynômes de Lagrange, relatifs aux support $\{x_0, ..., x_n\}$.

DÉMONSTRATION. On montre à la fois l'unicité et l'existence. On cherche Q un polynôme nul en x_0, x_1, x_2, ... x_{i-1}, x_{i+1}, ..., x_{n-1}, x_n et égal à 1 en x_i, de degré n. Voir figure 2.2. C'est donc équivalent à :

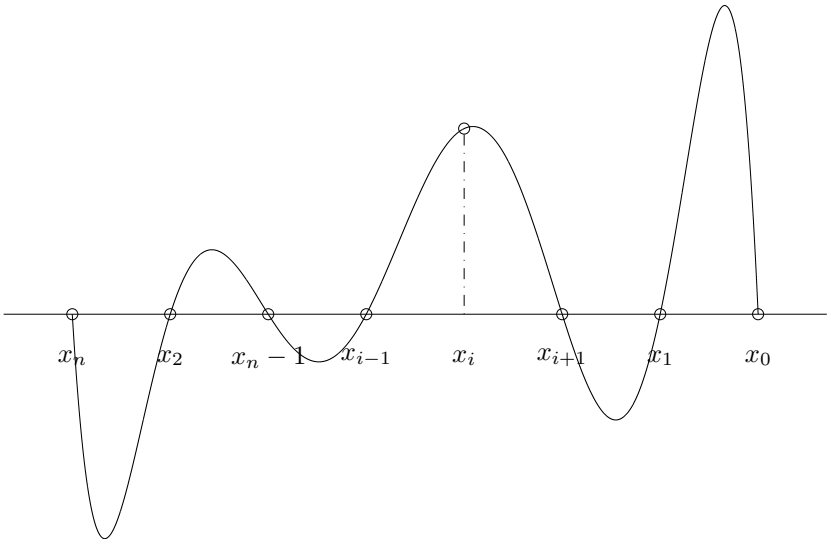


FIGURE 2.2. Le polynômes l_i, nul en x_0, x_1, x_2, ... x_{i-1}, x_{i+1}, ..., x_{n-1}, x_n et égal à 1 en x_i.

- Il admet pour racines les n nombres x_j, pour $j \in \{0, ..., n\} \setminus \{i\}$ et donc il existe un réel c tel que

$$Q = c\,(X - x_0)\,(X - x_1)\,...\,(X - x_{i-1})\,(X - x_{i+1})\,...\,(X - x_n) = c \prod_{\substack{j=0 \\ j \neq i}}^{n} (X - x_j)\,. \tag{2.24}$$

1. défini par

$$\forall i, j \in \{0, ..., n\}, \quad \delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases} \tag{2.22}$$

- De plus $l_i(x_i) = 1$ si et seulement si

$$c\,(x_i - x_0)\,(x_i - x_1)\,...\,(x_i - x_{i-1})\,(x_i - x_{i+1})\,...\,(x_i - x_n) = 1\ ;$$

d'après (2.3), les nombres x_i sont deux à deux distincts et le terme de gauche de cette égalité est non nul et on a donc

$$c = \frac{1}{(x_i - x_0)\,(x_i - x_1)\,...\,(x_i - x_{i-1})\,(x_i - x_{i+1})\,...\,(x_i - x_n)} = \frac{1}{\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j)},$$

et, d'après, (2.24), il vient

$$Q = \frac{\prod_{\substack{j=0\\ j\neq i}}^{n}(X - x_j)}{\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j)} = \prod_{\substack{j=0\\ j\neq i}}^{n}\frac{X - x_j}{x_i - x_j}.$$

□

Exemple 2.5. On vérifie que[2] pour $n = 0$:

$$l_0(x) = 1, \tag{2.25a}$$

puis, pour $n = 1$:

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}, \tag{2.25b}$$

$$l_1(x) = \frac{x - x_0}{x_1 - x_0}, \tag{2.25c}$$

puis, pour $n = 2$:

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \tag{2.25d}$$

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \tag{2.25e}$$

$$l_2(x) = \frac{(x - x_1)(x - x_0)}{(x_2 - x_0)(x_2 - x_1)}. \tag{2.25f}$$

Proposition 2.6. *Soit $n \in \mathbb{N}$. Il existe un unique polynôme Π_n de degré au plus n tel que*

$$\forall j \in \{0, ..., n\} \quad \Pi_n(x_j) = y_j. \tag{2.26}$$

Il est donné par

$$\Pi_n(x) = \sum_{i=0}^{n} y_i l_i(x). \tag{2.27}$$

Notons que le degré de Π_n est au plus n. Il peut être plus petit !

Démonstration.

2. En prenant pour convention qu'un produit vide vaut 1 ; voir exercice de TD 2.1.

(1) Démontrons d'abord l'unicité d'un tel polynôme. Supposons qu'il existe deux polynômes P_1 et P_2 deux degré au plus n qui vérifient

$$\forall i \in \{0, ..., n\}, \quad P_1(x_i) = P_2(x_i) = y_i.$$

Ainsi, $P_1 - P_2$ est un polynôme de degré au plus n qui vérifie

$$\forall i \in \{0, ..., n\}, \quad (P_1 - P_2)(x_i) = 0. \tag{2.28}$$

Remarquons que, d'après (2.3), on a

$$\text{Si } Q \text{ est un polynôme de degrés au plus } n, \text{ alors : } (\forall i \in \{0, ..., n\}, \quad Q(x_i) = 0) \Longrightarrow Q = 0. \tag{2.29}$$

En effet, dans ce cas, Q est de degré au plus n et possède au moins $n+1$ racines deux à deux distinctes ; il ne peut être que nul. Ainsi, d'après (2.28) et (2.29) appliqué à $Q = P_1 - P_2$, on a $Q = 0$ et donc $P_1 = P_2$.

(2) Enfin, si on choisit Π_n, défini par (2.27), alors d'après (2.21), pour tout j, on a

$$\sum_{i=0}^{n} y_i l_i(x_j) = \sum_{i=0}^{n} y_i \delta_{ij} = \sum_{\substack{0 \leq i \leq n \\ i \neq j}} y_i \delta_{ij} + y_j \delta_{jj} = 0 + y_j = y_j. \tag{2.30}$$

Cela assure l'existence du polynôme Π_n, défini par (2.27).

□

Démonstrations alternatives.

Nous donnons deux démonstrations alternatives.

(1) Les polynômes $(l_i)_{0 \leq i \leq n}$ forment une base de l'espace vectoriel P_n, des polynômes de degrés au plus n et on a, pour tout $P \in P_n$:

$$P = \sum_{i=0}^{n} P(x_i) l_i. \tag{2.31}$$

Démontrons en effet l'existence et l'unicité de la décomposition d'un polynôme de degré n sur la famille $(l_i)_{0 \leq i \leq n}$, c'est-à-dire

$$\forall P \in P_n, \quad \exists! (a_i)_{0 \leq i \leq n} \in \mathbb{R}^{n+1}, \quad P = \sum_{i=0}^{n} a_i l_i. \tag{2.32}$$

Notons que

$$\forall j \in \{0, ..., n\}, \quad \left(\sum_{i=0}^{n} a_i l_i \right)(x_j) = \sum_{i=0}^{n} a_i l_i(x_j)$$

et donc, d'après (2.21),

$$\forall j \in \{0, ..., n\}, \quad \left(\sum_{i=0}^{n} a_i l_i \right)(x_j) = \sum_{i=0}^{n} a_i \delta_{ij},$$

et donc

$$\forall j \in \{0, ..., n\}, \quad \left(\sum_{i=0}^{n} a_i l_i \right)(x_j) = a_j. \tag{2.33}$$

Ainsi, (2.32) est immédiat. En effet, d'après (2.29), dire que les polynômes de degré au plus n P et $\sum_{i=0}^{n} a_i l_i$ sont égaux est équivalent à dire qu'ils ont les mêmes valeurs aux $n+1$ points $\{x_j\}_{0 \leq j \leq n}$, c'est-à-dire :

$$\forall j \in \{0, ..., n\}, \quad P(x_j) = \sum_{i=0}^{n} a_i l_i(x_j),$$

ce qui est équivalent, compte tenu de (2.33)

$$\forall j \in \{0, ..., n\}, \quad P(x_j) = a_j.$$

Cette égalité fournit donc à la fois l'existence et l'unicité des coefficients a_i et leur valeur, ce qui montre en même temps (2.31), et donc (2.27).

(2) (a) Une autre version alternative consiste à d'abord montrer que les polynômes de Lagrange forment une base de P_n. Là encore, deux méthodes sont possibles.

(i) Pusique les polynômes de lagrange l_i sont en nombre égal à $n+1$, dimension de l'espace vectoriel P_n, il suffit de démontrer qu'ils constituent une famille libre. Pour cela, on suppose qu'il existe des scalaires $(a_i)_{0\leq i\leq n+1}$ tel que

$$\sum_{i=0}^{n} a_i l_i = 0.$$

Si on évalue cela en x_j, on obtient d'après (2.33) pour tout j, $a_j = 0$ et donc la famille est libre.

(ii) Sinon, de façon plus algébrique encore, on considère l'application Φ définie de P_n dans $\mathbb{R}^{n+1}$ par

$$\forall \in P_n, \quad \Phi(P) = \begin{pmatrix} P(x_0) \\ P(x_1) \\ \vdots \\ P(x_n) \end{pmatrix}$$

qui est clairement linéaire. De plus, elle est injective, puisque $\Phi(P) = 0$ implique que pour tout $i \in \{0, ..., n\}$, $P(x_i) = 0$, ce qui implique, d'après (2.29), la nullité de P. Puisque les deux espaces vectoriels P_n et $\mathbb{R}^{n+1}$ sont tous les deux de dimensions $n+1$, Φ est une application linéaire injective et donc bijective. Ainsi, Φ^{-1} est définie et est une application bijective et l'image par Φ^{-1} d'une base de $\mathbb{R}^{n+1}$ est une base de P_n. Or, il est clair que pour tout $i \in \{0, ..., n\}$, $\Phi(l_i)$ vaut d'après (2.21)

$$\Phi(l_i) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

vecteur dont toutes les composantes sont nulles sauf la $i+1$-ième qui vaut 1. L'ensemble de ces vecteurs constitue la base canonique de $\mathbb{R}^{n+1}$ et donc les l_i forment une base, en tant qu'image de la base canonique par Φ^{-1}.

(b) Ensuite, on écrit que pour tout $P \in P_n$, il existe un unique $n+1$-uplets de réels $(a_i)_{0\leq i\leq n+1}$ tels que

$$P = \sum_{i=0}^{n} a_i l_i,$$

et en évaluant cela en x_j pour tout j, on obtient grâce à (2.33) encore $a_j = P(x_j)$.

□

♢

Exemple 2.7. Reprenons les données de l'exemple 2.1 page 12. Chacun des polynômes de Lagrange l_i (de degré 3) est donné par la formule (2.23). On a donc successivement

$$\begin{aligned} l_0(x) &= \frac{(x-4)(x-2)(x+1)}{(1-4)(1-2)(1+1)}, \\ l_1(x) &= \frac{(x-1)(x-2)(x+1)}{(4-1)(4-2)(4+1)}, \\ l_2(x) &= \frac{(x-1)(x-4)(x+1)}{(2-1)(2-4)(2+1)}, \\ l_3(x) &= \frac{(x-1)(x-4)(x-2)}{(-1-1)(-1-4)(-1-2)}. \end{aligned}$$

soit encore après calculs :

$$l_0(x) = 1/6\,x^3 - 5/6\,x^2 + 1/3\,x + 4/3, \tag{2.34a}$$

$$l_1(x) = 1/30\,x^3 - 1/15\,x^2 - 1/30\,x + 1/15, \tag{2.34b}$$

$$l_2(x) = -1/6\,x^3 + 2/3\,x^2 + 1/6\,x - 2/3, \tag{2.34c}$$

$$l_3(x) = -1/30\,x^3 + \frac{7}{30}\,x^2 - \frac{7}{15}\,x + \frac{4}{15}. \tag{2.34d}$$

Ensuite, le polynôme interpolateur de degré 3, Π_3, est donné par la formule (2.27). Ici, on a donc :

$$\Pi_3(x) = y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x) + y_3 l_3(x).$$

Après calculs, il vient :

$$\Pi_3(x) = -\frac{17}{30}\,x^3 + \frac{33}{10}\,x^2 - \frac{44}{15}\,x - 9/5. \tag{2.35}$$

On retrouve bien (2.19d).

Exemple 2.8. On reprend les données de l'exemple 2.1.

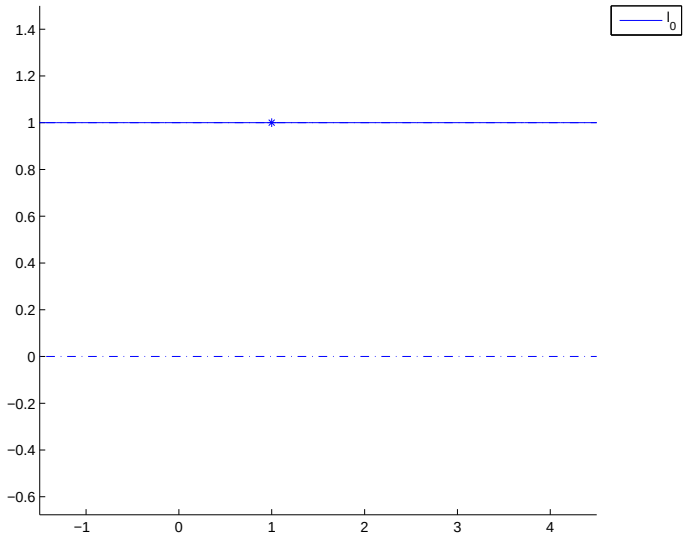


Figure 2.3. Le polynôme de Lagrange l_0

Sur les figures 2.3, 2.4, 2.5 et 2.6, ont été tracés : le polynôme l_0, interpolant la fonction f sur le support $\{x_0\}$, puis les polynômes l_0 et l_1, interpolant la fonction f sur le support $\{x_0, x_1\}$, puis les polynômes l_0, l_1 et l_2, interpolant la fonction f sur le support $\{x_0, x_1, x_2\}$, puis les polynômes l_0, l_1, l_2 et l_3, interpolant la fonction f sur le support $\{x_0, x_1, x_2, x_3\}$.

Voir aussi la figure 2.7.

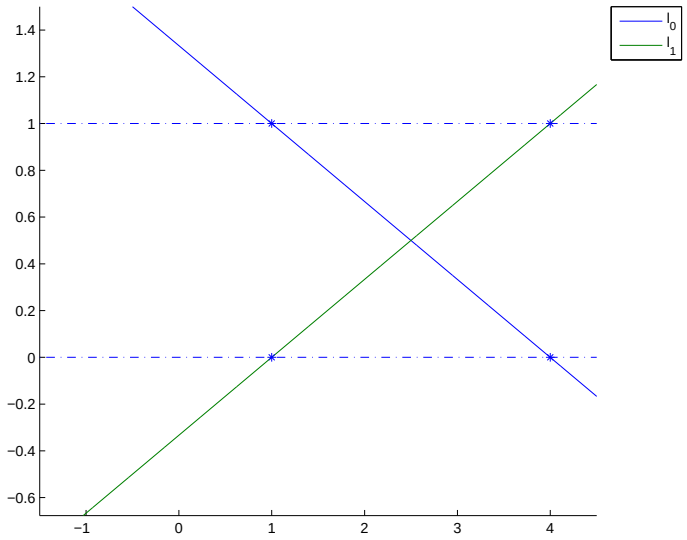


FIGURE 2.4. Les polynômes de Lagrange l_0 et l_1

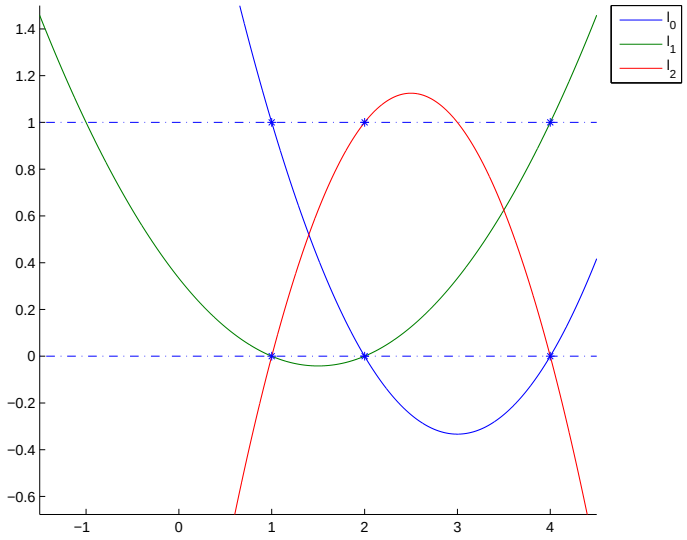


FIGURE 2.5. Les polynômes de Lagrange l_0, l_1 et l_2

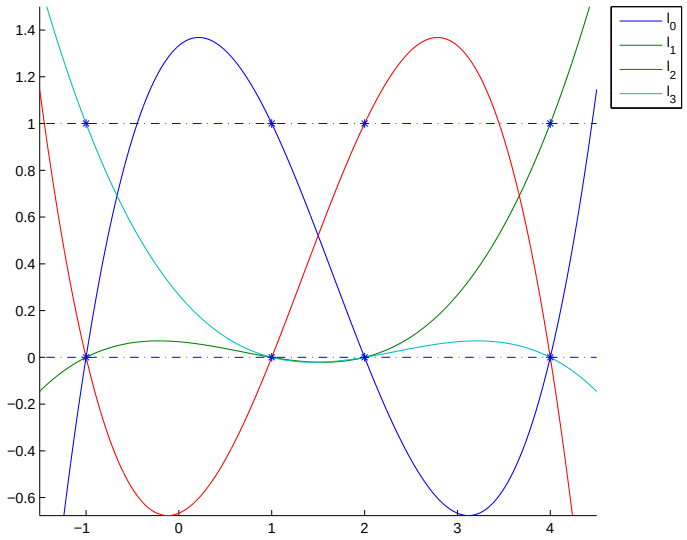


FIGURE 2.6. Les polynômes de Lagrange l_0, l_1, l_2 et l_3

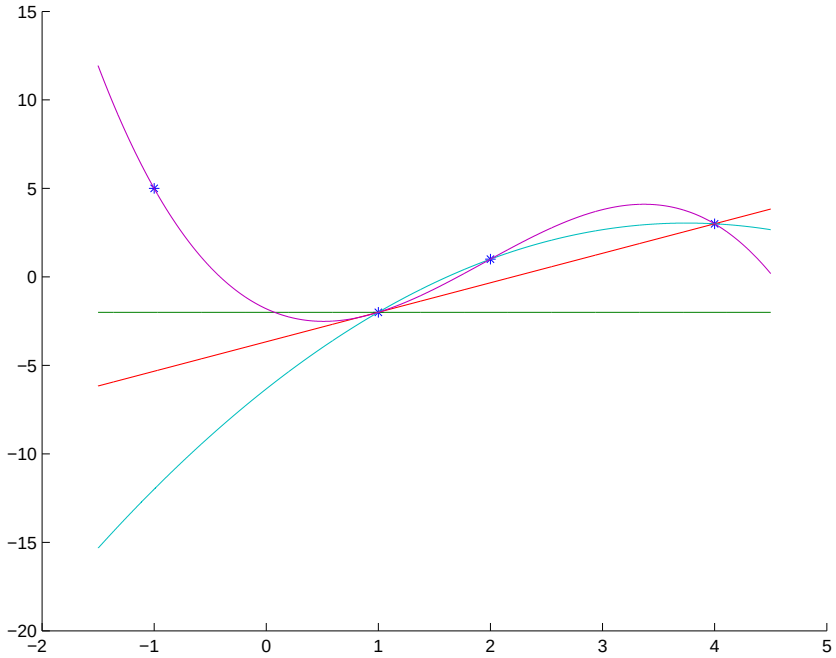


FIGURE 2.7. Les polynômes d'interpolation de f sur $\{x_0\}$, $\{x_0, x_1\}$, $\{x_0, x_1, x_2\}$ et $\{x_0, x_1, x_2, x_3\}$. .

2.2.2.3. Forme de Newton et différences divisées.

C'est la forme la plus appropriée à l'interpolation. On trouve aussi la terminologie "polynôme de Newton". Par convention, on pose [3]

$$\prod_{j=0}^{-1}(x-x_j)=1. \tag{2.36}$$

Nous allons chercher Π_n sous la forme

$$\Pi_n(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+...+a_n(x-x_0)(x-x_1)...(x-x_{n-1}), \tag{2.37a}$$

où les $(a_i)_{0\leq i\leq n}$ sont des réels à déterminer. En utilisant la convention (2.36), on écrit (2.37a) sous la forme

$$\Pi_n(x)=\sum_{i=0}^{n}a_i\prod_{j=0}^{i-1}(x-x_j)\,. \tag{2.37b}$$

Remarque 2.9. Dans l'espace vectoriel P_n des polynômes de degré au plus n, nous avons précédemment déterminé Π_n sur la base canonique 1, x, x^2 ...., x^n (section 2.2.2.1) ou sur la famille des l_i qui en forme aussi une base (section 2.2.2.2). Ici, nous utilisons la famille 1, $x-x_0$, $(x-x_0)(x-x_1)$, ..., $(x-x_0)...(x-x_{n-1})$ qui est aussi une base de P_n.

♢

Exemple 2.10. Pour $n=0$, on a, grâce à (2.5) pour $i=0$:

$$a_0=f(x_0).$$

Exemple 2.11. Pour $n=1$, on a, grâce à (2.5) pour $i=0$:

$$a_0=f(x_0),$$

puis pour $i=1$,

$$a_0+a_1(x_1-x_0)=f(x_1),$$

d'où

$$a_1=\frac{f(x_1)-f(x_0)}{x_1-x_0}.$$

Définition 2.12. Les quantités $(f[x_i])_{0\leq i\leq n}$ définies par

$$\forall i\in\{0,...,n\},\quad f[x_i]=f(x_i), \tag{2.38}$$

sont appelées les différences divisées d'ordre 0. Les quantités $(f[x_i,x_{i+1}])_{0\leq i\leq n-1}$ définies par

$$\forall i\in\{0,...,n-1\},\quad f[x_i,x_{i+1}]=\frac{f[x_{i+1}]-f[x_i]}{x_{i+1}-x_i}, \tag{2.39}$$

sont appelées les différences divisées d'ordre 1. Les quantités $(f[x_i,x_{i+1},x_{i+2}])_{0\leq i\leq n-2}$ définies par

$$\forall i\in\{0,...,n-2\},\quad f[x_i,x_{i+1},x_{i+2}]=\frac{f[x_{i+1},x_{i+2}]-f[x_i,x_{i+1}]}{x_{i+2}-x_i}, \tag{2.40}$$

sont appelées les différences divisées d'ordre 2. De manière plus générale, pour tout $k\in\{1,...,n\}$, les quantités $(f[x_i,...,x_{i+k}])_{0\leq i\leq n-k}$ définies par récurrence sur k (tandis que les différences divisées d'ordre 0 sont définies par (2.38)) par

$$\forall k\in\{1,...,n\},\quad \forall i\in\{0,...,n-k\},\quad f[x_i,...,x_{i+k}]=\frac{f[x_{i+1},...,x_{i+k}]-f[x_i,...,x_{i+k-1}]}{x_{i+k}-x_i} \tag{2.41}$$

sont appelées les différences divisées d'ordre k. Ainsi, l'unique différence divisée d'ordre n est définie par

$$f[x_0,...,x_n]=\frac{f[x_1,...,x_n]-f[x_0,...,x_{n-1}]}{x_n-x_0} \tag{2.42}$$

3. Voir note de bas de page numéro 2.

On peut alors montrer, avec les notations (2.37b), que

PROPOSITION 2.13. *On a*

$$\forall i \in \{0, ..., n\}, \quad a_i = f[x_0, x_1, ..., x_i], \tag{2.43}$$

soit encore

$$\Pi_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + ... + f[x_0, x_1, x_2, ..., x_n](x - x_0)(x - x_1)...(x - x_{n-1}), \tag{2.44a}$$

soit encore, avec la convention (2.36)

$$\Pi_n(x) = \sum_{i=0}^{n} f\left[x_0, ..., x_i\right] \prod_{j=0}^{i-1} (x - x_j)\,. \tag{2.44b}$$

DÉMONSTRATION. Voir [BM03, Définition 2.28 et Théorème 2.34], où l'on montre en particulier que $f\left[x_i, ..., x_{i+k}\right]$ est le coefficient dominant du polynôme interpolateur de f sur le support $\{x_i, ..., x_{i+k}\}$. □

♢

PROPOSITION 2.14 (Calcul complet des différences divisées). *Les égalités* (2.38) *permettent d'initialiser le calcul des différentes différences divisées. Les égalités* (2.41) *permettent de calculer successivement les différentes différences divisées.*

EXEMPLE 2.15. Pour $n = 0$, on a, grâce à (2.38) et (2.44a) :

$$f[x_0] = f(x_0), \tag{2.45a}$$

$$\Pi_0(x) = f(x_0). \tag{2.45b}$$

EXEMPLE 2.16. Pour $n = 1$, on a, grâce à (2.38), (2.44a) et (2.41) pour $k = 1$:

$$f[x_0] = f(x_0), \tag{2.46a}$$

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \tag{2.46b}$$

$$\Pi_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0). \tag{2.46c}$$

On retrouve les résultats obtenus dans l'exemple 2.2 page 13 (dans le cas où $x_A \neq x_B$).

REMARQUE 2.17.

(1) Le coefficient dominant de Π_n est égal à la différence divisé $f[x_0, ..., x_n]$.

(2) Ainsi, si $f[x_0, ..., x_n] = 0$, alors Π_n est de degré inférieur ou égal à $n - 1$.

2.2.2.4. Calcul pratique des différences divisées et du polynôme interpolateur.

DÉFINITION 2.18 (Tableau des différences divisée).
On définit le tableau des différences divisées de la façon suivante :

- La première colonne contient les $n + 1$ valeurs $(x_i)_{0 \leq i \leq n}$, classées dans l'ordre des indices (qui n'est pas nécessairement l'ordre des valeurs).
- La deuxième colonne contient les $n + 1$ différences divisées d'ordre 0, c'est-à-dire $(f[x_i])_{0 \leq i \leq n}$, chacune d'elles étant égale à $f(x_i)$ d'après (2.38).
- La troisième colonne contient les n différences divisées d'ordre 1, c'est-à-dire $(f[x_i, x_{i+1}])_{0 \leq i \leq n-1}$, données par (2.39).
- La quatrième colonne contient les $n-1$ différences divisées d'ordre 2, c'est-à-dire $(f[x_i, x_{i+1}, x_{i+2}])_{0 \leq i \leq n-2}$, données par (2.40)

- De façon générale, la $k+2$-ième colonne pour tout $k \in \{1, ..., n\}$ contient les $n-k+1$ différences divisées d'ordre k, c'est-à-dire $(f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}])_{0 \leq i \leq n-k}$ données par (2.41).
- L'avant-dernière colonne contient les 2 différences divisées d'ordre $n-1$, c'est-à-dire, $f[x_0, x_1, \ldots, x_{n-1}]$ et $f[x_1, x_2, \ldots, x_n]$ définie par (2.41) pour $k = n-1$ et $i \in \{0, 1\}$, c'est-à-dire

$$f[x_0, x_1, \ldots, x_{n-1}] = \frac{f[x_1, \ldots, x_{n-1}] - f[x_0, x_1, \ldots, x_{n-2}]}{x_{n-1} - x_0},$$

$$f[x_1, x_2, \ldots, x_n] = \frac{f[x_2, \ldots, x_n] - f[x_1, x_2, \ldots, x_{n-1}]}{x_n - x_1}.$$

- La dernière colonne contient l'unique différence divisée d'ordre n, c'est-à-dire, $f[x_0, x_1, \ldots, x_{n-1} x_n]$ définie (2.41) pour $k = n$ et $i = 0$, c'est-à-dire

$$f[x_0, x_1, \ldots, x_{n-1}, x_n] = \frac{f[x_1, \ldots, x_{n-1}, x_n] - f[x_0, x_1, \ldots, x_{n-2}, x_{n-1}]}{x_n - x_0}.$$

Voir les tableaux 2.2 page suivante et 2.3 page 26. La double flèche $\begin{matrix} \searrow \\ \nearrow \end{matrix}$ signifie que chaque différence divisée $f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}]$ pour $k \in \{1, ..., n\}$ et pour $i \in \{0, ..., n-k\}$ est déterminée de la façon suivante

$$\begin{matrix} A \searrow \\ B \nearrow \end{matrix} \frac{B - A}{x_{i+k} - x_i} = f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}].$$

Enfin, on se sert de ce tableau et uniquement des différences divisées encadrées dans les tableaux 2.2 et 2.3 pour calculer Π_n donné par (2.44)

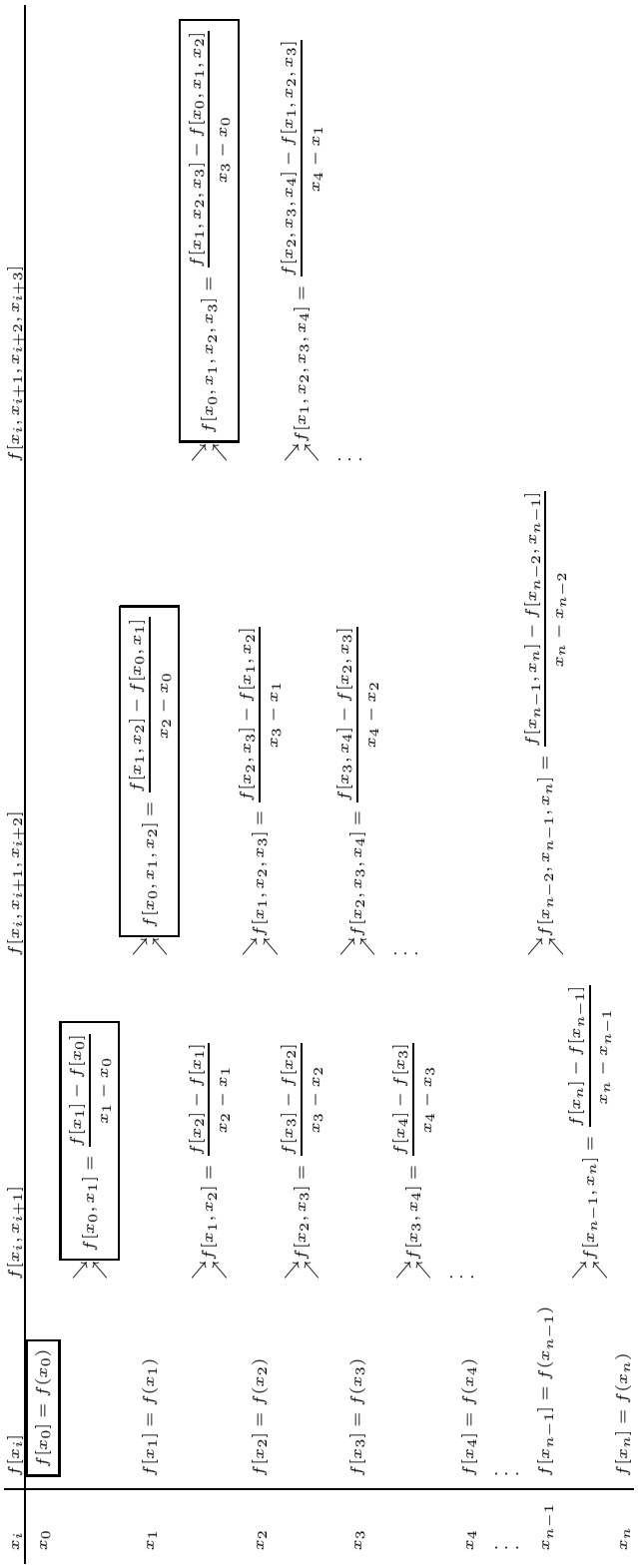


TABLE 2.2. Construction des différences divisées (premières colonnes).

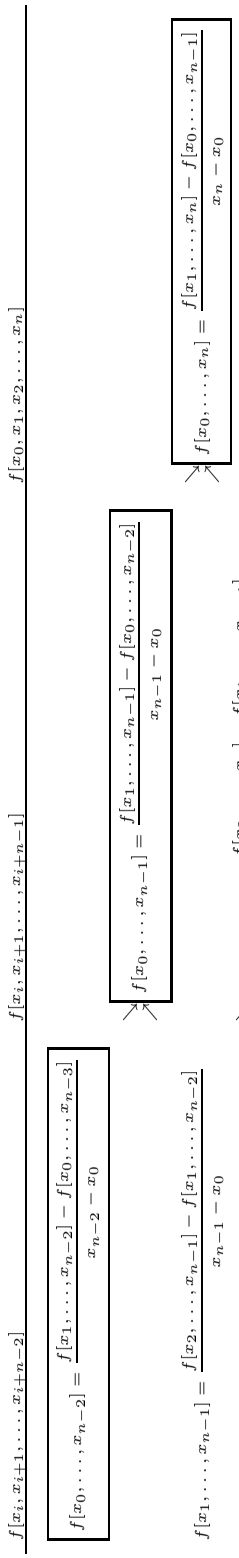


TABLE 2.3. Construction des différences divisées (dernières colonnes).

Nous avons aussi la relation de récurrence fondamentale qui découle tout simplement de l'écriture de Π_n :

LEMME 2.19. *Si, pour $n \geq 1$, Π_{n-1} est le polynôme interpolateur de f sur le support $\{x_0, ..., x_{n-1}\}$ alors, Π_n, le polynôme interpolateur de f sur le support $\{x_0, ..., x_n\}$ vérifie*

$$\Pi_n(x) = \Pi_{n-1}(x) + f[x_0, x_1, ..., x_n](x - x_0)(x - x_1)...(x - x_{n-1}). \tag{2.47}$$

DÉMONSTRATION. Faire en exercice. □

REMARQUE 2.20. Grâce au lemme 2.19, si le tableau des différences divisées sur le support d'interpolation $\{x_0, ..., x_{n-1}\}$ est connu, on en déduit facilement le tableau des différences divisées sur le support d'interpolation $\{x_0, ..., x_n\}$. Il suffit de rajouter les points x_n et la donnée $f(x_n)$ en bas de tableau et d'en déduire progressivement les nouvelles différences divisées comme le montre le tableau 2.4 page suivante. Si le polynôme Π_{n-1} est connu, on déduit alors Π_n en utilisant la nouvelle différence divisée calculée $f[x_0, ..., x_n]$ et (2.47).

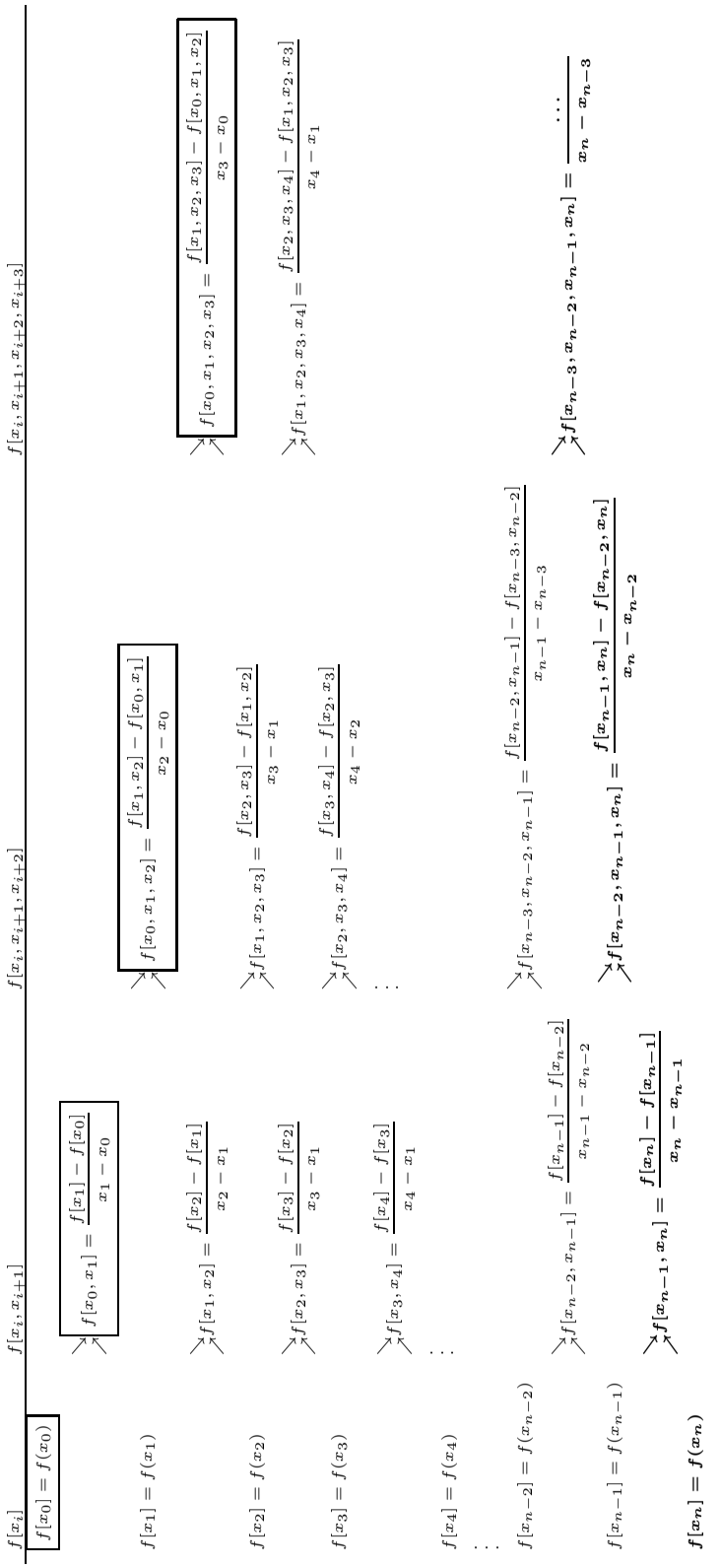


TABLE 2.4. Construction des différences divisées : rajout d'un point (calculs en gras) x_n en utilisant les calculs relatifs à $x_0, \ldots, x_{n-1}$.

2.2.2.5. Bilan sur le choix de la méthode.
La méthode de newton, sauf indication, sera toujours à privilégier, cela pour deux raisons :

(1) Sur le plan algorithmique d'abord ; la méthode de Lagrange force à tout recalculer les polynômes l_0, ..., l_n à chaque rajout de nouveau point. Au contraire, la méthode de Newton, à l'ajout d'un point, permet de calculer le nouveau polynôme en se servant du précédent, grâce au lemme 2.19. C'est d'ailleurs pour cela qu'à été inventée la notion de différences divisées, qui se calculent aisément de façon récurrente grâce à la proposition 2.14. Voir la définition 2.18, la remarque 2.20 et l'exercice de TD 2.2.

(2) Sur le plan numérique ensuite ; grâce à l'algorithme d'évaluation d'Horner (voir [BM03, Algorithme 2.1 p. 38]), la forme de Newton (2.44) permet de calculer de façon plus correcte Π_n qu'en utilisant les polynômes de Lagrange l_i, notamment quand n grandit. Voir le corrigé de l'exercice de TD 2.4.

Exemple 2.21. Reprenons l'exemple 2.2 page 13, dans le cas où $x_A \neq x_B$, avec les différentes méthodes vues dans ce chapitre. On renvoie à l'annexe C.

◇

Exemple 2.22. Reprenons les données de l'exemple 2.1 page 12. Pour calculer le polynôme sous la forme

$x_i \setminus k$	0	1	2	3
$x_0 = 1$	$\boxed{-2}$			
		$\boxed{5/3}$		
$x_1 = 4$	3		$\boxed{-2/3}$	
		1		$\boxed{-\frac{17}{30}}$
$x_2 = 2$	1		$\frac{7}{15}$	
		$-4/3$		
$x_3 = -1$	5			

Table 2.5. Différences divisées de f.

de Newton, on détermine tout d'abord les différences divisées $f[x_i, ..., x_{i+k}]$ données dans le tableau 2.5. Ensuite, on n'utilise plus que les différences divisées qui sont encadrées et le polynôme interpolateur de degré 3, Π_3, est donné par la formule (2.44). Ici, on a donc :

$$\Pi_3(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2).$$

On a successivement

$$\begin{aligned} x - x_0 &= x - 1, \\ (x - x_0)(x - x_1) &= x^2 - 5\,x + 4, \\ (x - x_0)(x - x_1)(x - x_2) &= x^3 - 7\,x^2 + 14\,x - 8. \end{aligned}$$

Après calculs, il vient :

$$\Pi_3(x) = -\frac{17}{30}\,x^3 + \frac{33}{10}\,x^2 - \frac{44}{15}\,x - 9/5. \tag{2.48}$$

On retrouve bien (2.19d).

Exemple 2.23. Voir de nouveau l'exemple de la section 2.1.

2.2.3. Erreur en interpolation polynômiale

On rappelle que, d'après (2.2), on dit que Π_n (noté parfois en cas d'ambiguïté $\Pi_n(f)$) interpole f sur le support $\{x_0, ..., x_n\}$ si :

$$\forall i \in \{0, ..., n\}, \quad \Pi_n(x_i) = f(x_i). \tag{2.49}$$

THÉORÈME 2.24 (Erreur d'interpolation pour des nœuds quelconques). *Soient $(x_i)_{0\leq i\leq n}$ $n+1$ nœuds distincts dans $[a, b]$ et soit $f \in \mathcal{C}^{n+1}([a, b])$. Alors*

$$\forall x \in [a, b], \quad \exists \xi \in]a, b[, \quad f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \tag{2.50}$$

où ω_{n+1} est défini par

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i). \tag{2.51}$$

DÉMONSTRATION. On montre tout d'abord que pour tout réel x de $[a, b]$ on peut écrire :

$$f(x) - \Pi_n(x) = f[x_0, ..., x_n, x]\left[\prod_{j=0}^{n}(x - x_j)\right]. \tag{2.52}$$

Dans cette proposition, $f[x_0, ..., x_n, x]$ devra avoir un sens, que les points soient distincts ou non : voir [BM03, exercice 2.8 et TP 2.F]. On considère deux cas.

- Si $x \notin \{x_0, ..., x_n\}$, alors $\{x_0, ..., x_n, x\}$ constitue un ensemble de $n+2$ points distincts de I. Notons Π_{n+1} le polynôme d'interpolation de f sur $\{x_0, ..., x_n, x\}$. Par suite $\Pi_{n+1}(x) = f(x)$; donc $\Pi_{n+1}(t) - \Pi_n(t)$ désigne le terme qu'il faut ajouter à $\Pi_n(x)$ pour obtenir $\Pi_{n+1}(x)$, c'est-à-dire l'expression (2.52).
- Si $x \in \{x_0, ..., x_n\}$ alors $f(x) - \Pi_n(x) = 0$. Comme le produit qui apparaît dans l'expression (2.52) est nul, celle-ci reste valide en admettant que $f[x_0, ..., x_n, x]$ a un sens.

On montre ensuite qu'il existe $\xi \in]a, b[$ tel que :

$$f[x_0, ..., x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Ce calcul est assez technique et repose sur une utilisation itérée du théorème de Rolle. Voir par exemple [CB81]. □

♢

Si on peut trouver une borne supérieure pour $f^{(n+1)}(\xi)$, soit $M_{n+1} = \max_{x_0\leq x\leq x_n} |f^{(n+1)}(x)|$, on peut alors majorer l'erreur sous la forme suivante :

THÉORÈME 2.25 (Erreur d'interpolation pour des nœuds quelconques). *Sous les hypothèses du théorème 2.24, on note*

$$M_{n+1} = \max_{a\leq x\leq b}\left|f^{(n+1)}(x)\right|. \tag{2.53}$$

Alors

$$\forall x \in [a, b], \quad |f(x) - \Pi_n f(x)| \leq \frac{M_{n+1}}{(n+1)!}\max_{x\in[a,b]}|\omega_{n+1}(x)|. \tag{2.54}$$

DÉFINITION 2.26 (Nœuds équirépartis). Pour $a < b$ et $n \in \mathbb{N}^*$, $n+1$ nœuds équirépartis $(x_i)_{0\leq i\leq n}$ définissent n sous-intervalles de $[a, b]$ de la même taille et sont définis par

$$h = \frac{b-a}{n}, \tag{2.55a}$$

$$\forall i \in \{0, ..., n\}, \quad x_i = a + ih. \tag{2.55b}$$

THÉORÈME 2.27 (Erreur d'interpolation pour des nœuds équirépartis). *Pour des nœuds équirépartis, c'est-à-dire donnés par la définition 2.26, on a, sous les hypothèses du théorème 2.24,*

$$E_n(f) = \max_{x\in[a,b]}|f(x) - \Pi_n f(x)| \leq \frac{1}{4(n+1)}\left(\frac{b-a}{n}\right)^{n+1}\max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|. \tag{2.56}$$

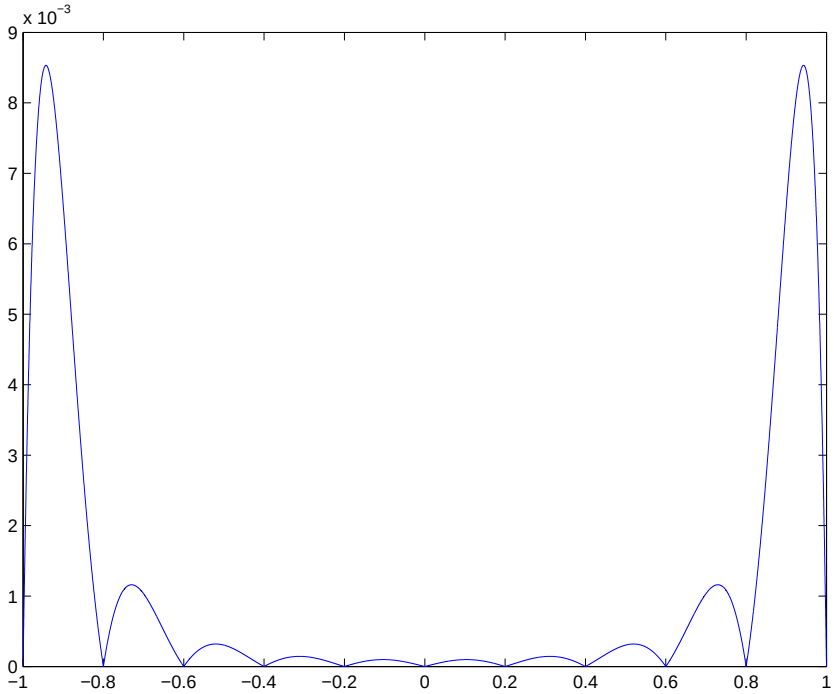


FIGURE 2.8. Fonction $|\omega_{11}|$ pour 11 nœuds équirépartis dans $[-1, 1]$.

DÉMONSTRATION. On peut montrer que le maximum de $|\omega_{n+1}(x)|$ est atteint toujours dans un des deux intervalles extrêmes $[x_0, x_1]$ ou $[x_{n-1}, x_n]$ (voir figure 2.8). On prend $x \in [x_0, x_1]$ (l'autre cas est similaire), et on a

$$|(x - x_0)(x - x_1)| \leq \frac{(x_1 - x_0)^2}{4}.$$

On peut aussi montrer que le maximum de $|(x - x_0)(x - x_1)|$ est atteint en $\dfrac{x_0 + x_1}{2}$ et vaut $\dfrac{(x_1 - x_0)^2}{4}$. On a donc

$$|(x - x_0)(x - x_1)| \leq \frac{(x_1 - x_0)^2}{4} = \frac{h^2}{4},$$

où on a noté le pas $h = (b - a)/n$. De plus, pour tout $i > 1$,

$$|(x - x_i)| \leq ih.$$

Donc,

$$\max_{x \in I} \prod_{i=0}^{n} |(x - x_i)| \leq \frac{h^2}{4}(2h) \times (3h) \times \cdots \times (nh) = \frac{h^{n+1} n!}{4}.$$

□

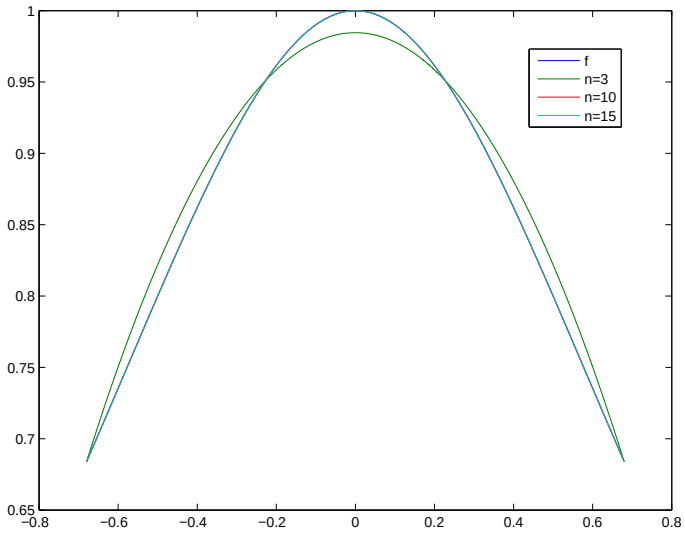
♢

REMARQUE 2.28. Attention, l'erreur $E_n(f)$ ne tend pas nécessairement vers zéro quand n tend vers l'infini.

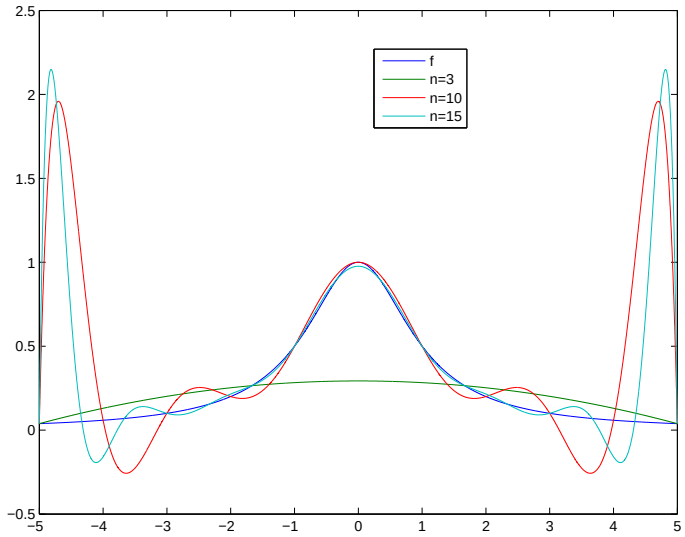
EXEMPLE 2.29. Soit f définie par

$$\forall x \in \mathbb{R}, \quad f(x) = \frac{1}{1 + x^2}. \tag{2.57}$$

Si on interpole avec des points équirépartis sur un intervalle $[-a, a]$, l'erreur peut tendre en théorie vers zéro ou pas quand n tend vers l'infini, selon la valeur de a. Si $2a < e$, il y a convergence vers zéro, en particulier pour $a = 1/4e$, comme l'indique la figure 9(a). Si $2a \geq e$, on peut montrer que l'erreur tend vers l'infini, en



(a) $a = 1/4\, e^1$



(b) $a = 5$

FIGURE 2.9. Interpolation de la fonction de Runge. Dans le cas de la figure 9(a), la convergence a lieu quand n tend vers l'infini. Au contraire, les polynômes d'interpolation présentent des oscillations qui augmentent avec le degré du polynôme pour la figure 9(b).

particulier pour $a = 5$, comme l'indique la figure 2.9, où l'interpolant présente des oscillations. Voir [CM84, p. 10].

REMARQUE 2.30. Attention aussi au mauvais comportement numérique des évaluations polynômiales ; voir le corrigé de l'exercice de TD 2.4.

Afin de palier le problème de l'absence de convergence ou les instabilités numériques quand le degré grandit, deux méthodes alternatives sont proposées : la méthode de Tchebycheff (voir section 2.4) ou l'interpolation par intervalles ou par morceaux (dite aussi interpolation composée ou composite, voir section 2.5).

2.3. Un exercice type à savoir traiter parfaitement (Interpolation de Lagrange)

Énoncé

On connaît les valeurs d'une fonction g aux points $x_0 = -1$, $x_1 = 3$ et $x_2 = 5$:

$$g(x_0) = 0, \quad g(x_1) = 1, \quad g(x_2) = 4.$$

(1) Construire le polynôme de degré au plus 2 (noté $\Pi_2 g$), interpolant la fonction g aux nœuds x_0, x_1 et x_2.

(2) Pour $\alpha = 1$, donner une valeur approchée de $g(\alpha)$.

Corrigé

(1) Il est naturellement préférable d'utiliser la forme de Newton pour le calcul de l'interpolation, néanmoins, le calcul utilisant la forme de Lagrange est aussi présenté.

(a) Chacun des polynômes de Lagrange l_i (de degré 2) est donné par la formule :

$$\forall i \in \{0, ..., n\}, \quad l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{2.58}$$

On a donc successivement

$$l_0(x) = \frac{(x-3)\,(x-5)}{(-1-3)\,(-1-5)},$$
$$l_1(x) = \frac{(x+1)\,(x-5)}{(3+1)\,(3-5)},$$
$$l_2(x) = \frac{(x+1)\,(x-3)}{(5+1)\,(5-3)}.$$

soit encore après calculs :

$$l_0(x) = 1/24\,x^2 - 1/3\,x + 5/8, \tag{2.59a}$$
$$l_1(x) = -1/8\,x^2 + 1/2\,x + 5/8, \tag{2.59b}$$
$$l_2(x) = 1/12\,x^2 - 1/6\,x - 1/4. \tag{2.59c}$$

Ensuite, le polynôme interpolateur de degré 2, $\Pi_2(g)$, est donné par la formule :

$$\Pi_2(g)(x) = \sum_{i=0}^{n} g(x_i) l_i(x). \tag{2.60}$$

Ici, on a donc :

$$\Pi_2(g)(x) = g(x_0) l_0(x) + g(x_1) l_1(x) + g(x_2) l_2(x).$$

Après calculs, il vient :

$$\Pi_2(g)(x) = \frac{5}{24}\,x^2 - 1/6\,x - 3/8. \tag{2.61}$$

(b)

Pour calculer le polynôme sous la forme de Newton, on détermine tout d'abord les différences divisées $g[x_i, ..., x_{i+k}]$ données dans le tableau 2.6. Ensuite, on n'utilise plus que les différences divisées qui sont encadrées et le polynôme interpolateur est donné par la formule :

$$\Pi_2(g)(x) = \sum_{i=0}^{n} g[x_0, ..., x_i](x - x_0)...(x - x_{i-1}). \tag{2.62}$$

$x_i \setminus k$	0	1	2
$x_0 = -1$	$\boxed{0}$		
		$\boxed{1/4}$	
$x_1 = 3$	1		$\boxed{\frac{5}{24}}$
		3/2	
$x_2 = 5$	4		

TABLE 2.6. Différences divisées de g.

Ici, on a donc :

$$\Pi_2(g)(x) = g[x_0] + g[x_0, x_1](x - x_0) + g[x_0, x_1, x_2](x - x_0)(x - x_1).$$

On a successivement

$$x - x_0 = x + 1,$$
$$(x - x_0)(x - x_1) = x^2 - 2\,x - 3.$$

Après calculs, on retrouve donc bien le polynôme déterminé par la méthode de Lagrange (voir équation (2.61)).

(2) Pour $\alpha = 1$, on obtient alors :

$$\Pi_2(g)(\alpha) = -1/3 \approx -0.333333,$$

ce qui constitue une valeur approchée de $g(\alpha)$.

2.4. Interpolation de Tchebycheff (ou Chebyshev)

DÉFINITION 2.31. Soit $n \in \mathbb{N}^*$. Sur l'intervalle $[a, b]$, les $n + 1$ points $x_{i0 \leq i \leq n}$ de Tchebycheff sont donnés par

$$\forall i \in \{0, ..., n\}, \quad z_i = -\cos\left(\frac{\pi i}{n}\right), \tag{2.63a}$$

et

$$\forall i \in \{0, ..., n\}, \quad x_i = \frac{a+b}{2} + \frac{b-a}{2} z_i, \tag{2.63b}$$

REMARQUE 2.32. Notons aussi l'interprétation géométrique des formules (2.63) : on considère le demi-cercle de centre d'abscisse $(a + b)/2$ et de rayon $(b - a)/2$, c'est-à-dire passant par les points d'abscisses a et b. On divise ce demi-cercle en en n parties égales ; on a donc $n + 1$ points définis par les angles $i\pi/n$ pour $0 \leq i \leq n$. Voir la figure 2.10 page suivante. Les $n + 1$ abscisses de ces points ne sont autres que les points définis par (2.63).

REMARQUE 2.33. Attention, parfois une définition légèrement différente de (2.63a) est donnée (c'est le cas dans [BM03, p. 282]) :

$$\forall i \in \{0, ..., n\}, \quad z_i = \cos\left(\frac{2i+1}{n+1}\frac{\pi}{2}\right). \tag{2.64}$$

♢

REMARQUE 2.34. On pourra démontrer les formules (2.63) à partir de l'interprétation géométrique de la figure 2.10. Voir exercice de TD 2.10.

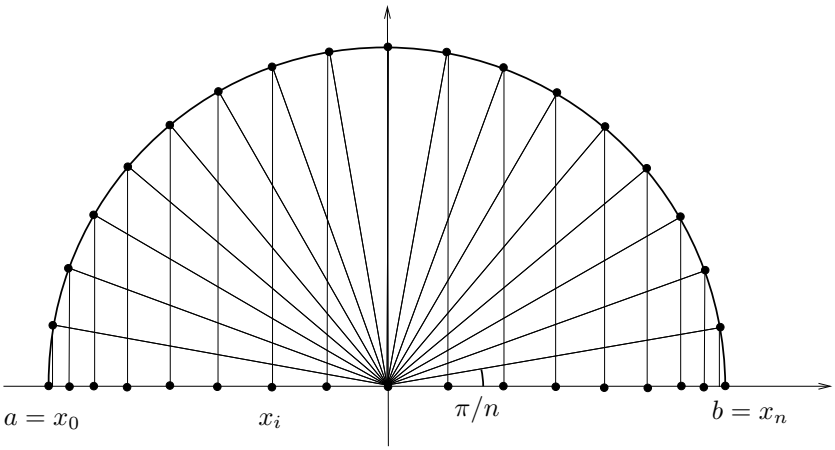


FIGURE 2.10. Les points de Tchebycheff.

THÉORÈME 2.35. *Pour une fonction $f \in \mathcal{C}^1([a,b])$, le polynôme d'interpolation $\Pi_n f$ de degré n aux noeuds $(x_i)_{0\leq i\leq n}$ de Tchebysheff converge uniformément vers f quand n tend vers l'infini, c'est-à-dire que l'on a*

$$\lim_{n\to+\infty} \max_{x\in[a,b]} |\Pi_n f(x) - f(x)| = 0. \tag{2.65}$$

DÉMONSTRATION. Voir [CM84, p. 20]. □

♢

EXEMPLE 2.36. Reprenons l'exemple 2.29 où au lieu de prendre des nœudes équirépartis, on choisit ceux de Tchebycheff. Soit f définie par (2.57) et $a = 5$

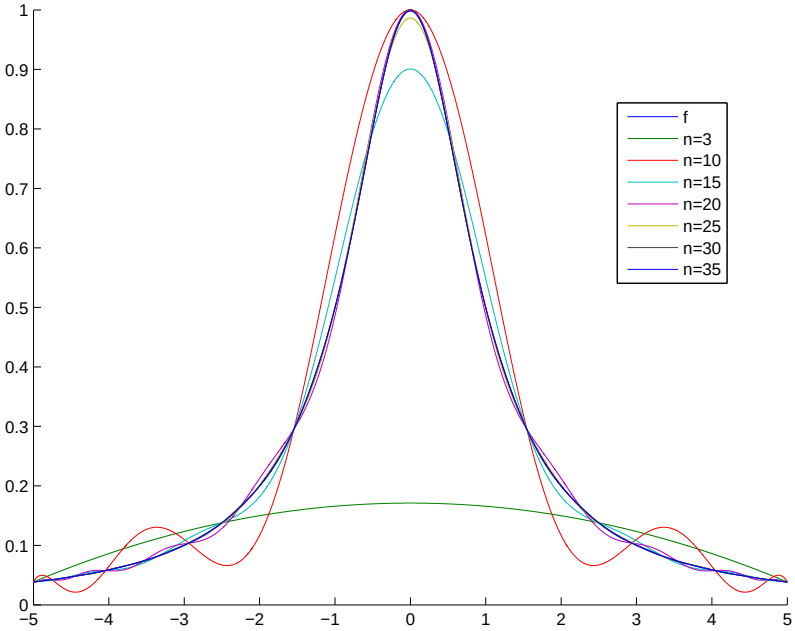


FIGURE 2.11. Interpolation de la fonction de Runge avec les points de Tchebycheff pour $a = 5$.

Dans ce cas, l'interpolant ne présente plus d'oscillations. Voir figure 2.11, à comparer avec la figure 9(b).

2.5. Interpolation par intervalles ou par morceaux (dite aussi interpolation composée ou composite)

Pour éviter d'avoir un degré trop élevé et néanmoins permettre une approximation correcte d'une fonction, on découpe l'intervalle en un grand nombre de sous-intervalle et on interpole sur chacun d'eux avec un degré fixé. La précision de l'interpolation proviendra du grand nombre de sous intervalles.

THÉORÈME 2.37 (Erreur d'interpolation composite). *Soient $x_0 = A < x_1 < \cdots < x_N = B$ des points qui divisent $I = [A, B]$. On note $I_j = [x_{j-1}, x_j]$ les sous-intervalles de longueur h_j et $h = \max_{1\leq j\leq N} h_j$. Sur chaque sous-intervalle I_j, on interpole $f_{|_{I_j}}$ par un polynôme de degré*[4] *n avec des points équirépartis. Le polynôme par morceaux est noté $\Pi_n^h f(x)$. Si $f \in \mathcal{C}^{n+1}([A, B])$, alors, on a*

$$E_n^h(f) = \max_{x\in[A,B]} |f(x) - \Pi_n^h f(x)| \leq \frac{1}{4(n+1)n^{n+1}} \max_{x\in[A,B]} \left|f^{(n+1)}(x)\right| h^{n+1}. \tag{2.66}$$

Si de plus, les points sont équirépartis, on a

$$E_n^h(f) = \max_{x\in[A,B]} |f(x) - \Pi_n^h f(x)| \leq \frac{(B-A)^{n+1}}{4(n+1)n^{n+1}} \max_{x\in[A,B]} \left|f^{(n+1)}(x)\right| \frac{1}{N^{n+1}}. \tag{2.67}$$

DÉMONSTRATION. On applique le théorème 2.27 à chaque intervalle. □

♢

REMARQUE 2.38. À n fixé, c'est le fait que h tende vers 0 qui assure que l'erreur $E_n^h(f)$ tend vers 0.

EXEMPLE 2.39. On considère A et B définis par

$$A = 0, \quad B = 1. \tag{2.68}$$

et la fonction f définie par

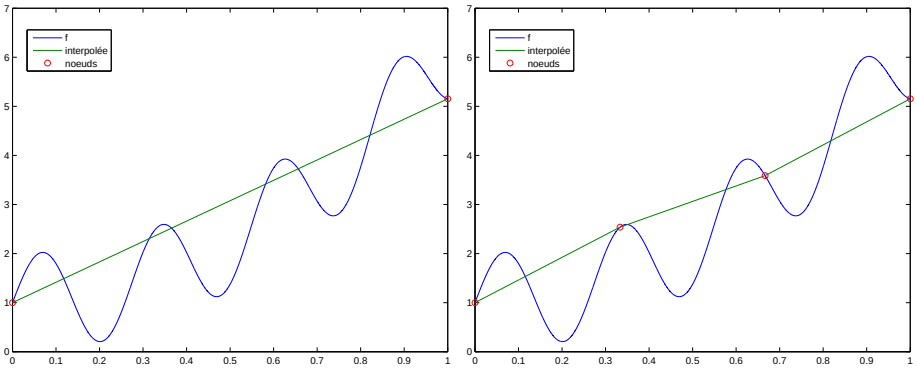
$$\forall x \in [A, B], \quad f(x) = \sin(23\,x) + 1 + 5\,x^2. \tag{2.69}$$

On détermine par la méthode d'interpolation composite, les interpolations de f de degré 1 et 2, par morceaux, en considérant différentes valeurs de N, nombre de sous-intervalle décrivant l'ensemble

$$J = \{1, 3, 5, 10, 20, 50\}. \tag{2.70}$$

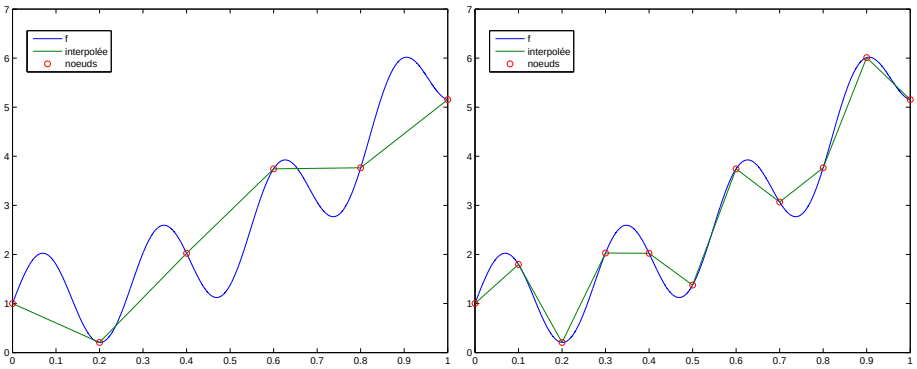
Voir les figures 2.12 et 2.13. Nous reviendrons sur cette méthode lors du chapitre 3 (voir exemple 3.26 page 61), où pour calculer l'intégrale approchée de f, nous remplacerons la fonction f par l'interpolation composite ainsi définie.

4. Attention, N le nombre de sous-intervalles ou morceaux n'a aucun lien avec n le degré des polynômes !



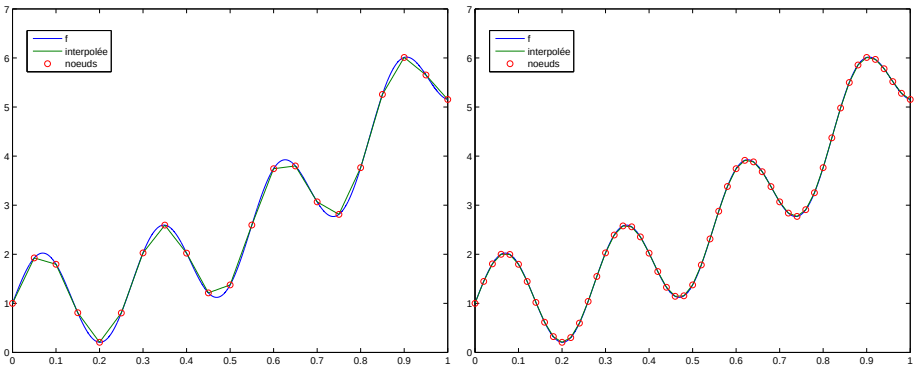
(a) $N = 1$.

(b) $N = 3$.



(c) $N = 5$.

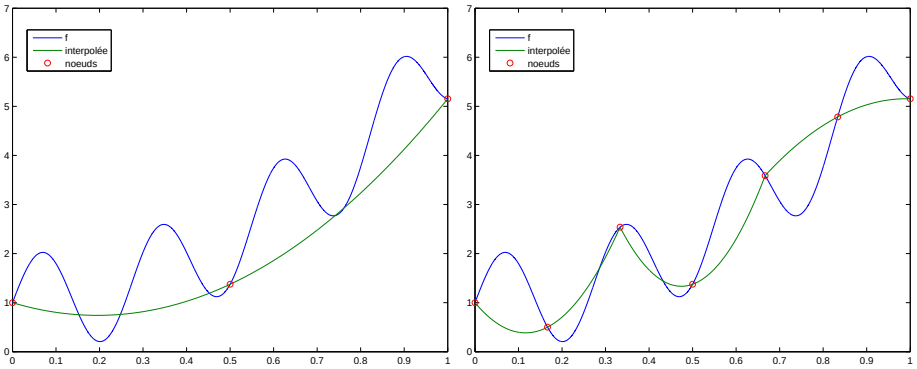
(d) $N = 10$.



(e) $N = 20$.

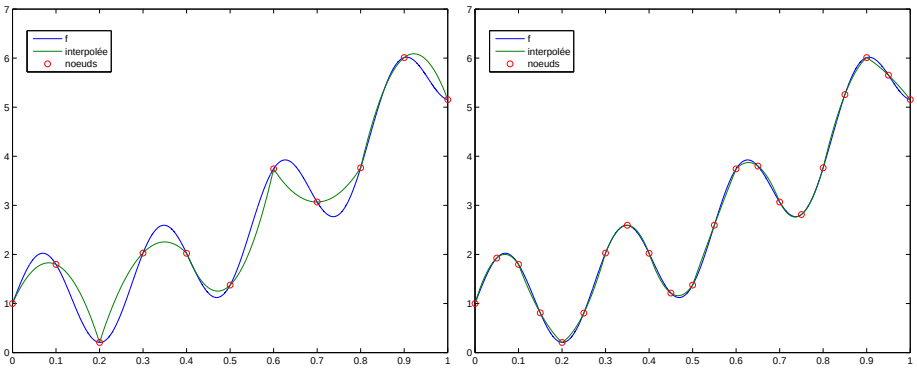
(f) $N = 50$.

FIGURE 2.12. Différentes illustrations de l'interpolation composite, avec un degré égal à 1 et différentes valeurs de N.



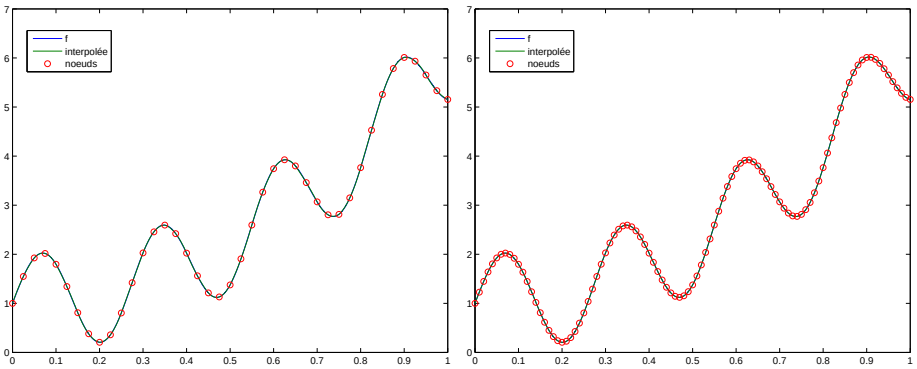
(a) $N = 1$.

(b) $N = 3$.



(c) $N = 5$.

(d) $N = 10$.



(e) $N = 20$.

(f) $N = 50$.

FIGURE 2.13. Différentes illustrations de l'interpolation composite, avec un degré égal à 2 et différentes valeurs de N.

2.6. Splines cubiques et autres courbes d'ajustement

2.6.1. Splines cubiques

On pourra consulter : `https://fr.wikipedia.org/wiki/Spline`

On considère des points x_j d'un intervalle. On approxime une fonction f par un polynôme de degré trois, sur chacun des intervalles $[x_j, x_{j+1}]$. Sur chaque intervalle, ce polynôme est connu grâce aux deux valeurs de f en x_j et x_{j+1} ce qui n'est pas suffisant. On impose alors la continuité de la dérivée première et seconde du polynôme de degré trois de part et d'autre de chaque point x_j. Manquent alors deux conditions pour déterminer complément la spline et plusieurs choix sont possibles.

Exemple 2.40.

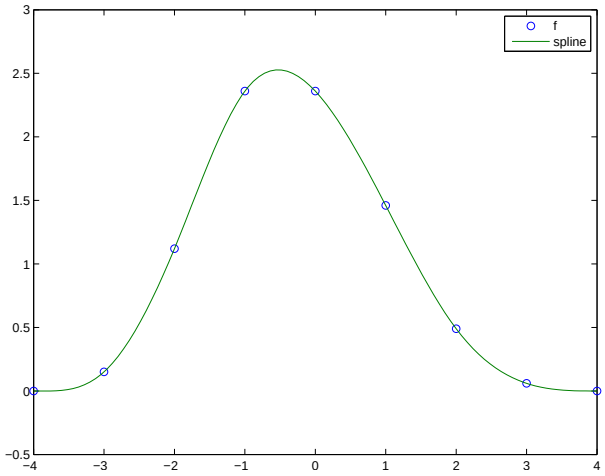


Figure 2.14. Spline cubique.

Voir l'exemple de la figure 2.14, directement inspirée de l'aide de matlab.

Exemple 2.41. Voir de nouveau l'exemple de la section 2.1.

2.6.2. Autres courbes d'ajustement

On pourra consulter `https://fr.wikipedia.org/wiki/Courbe_de_Bézier` et `https://fr.wikipedia.org/wiki/Algorithme_de_Casteljau` dont sont extraites et adaptées les lignes suivantes. On pourra aussi consulter [HM13].

De nombreuses courbes permettent de relier des points entre eux. Par exemple, les courbes de Bézier sont des courbes polynomiales paramétriques décrites pour la première fois en 1962 par Pierre Bézier (ingénieur Arts et Métiers et Supélec à la régie Renault dans les années 1950) qui les utilisait pour concevoir des pièces d'automobiles. L'algorithme de construction de ces courbes avait été aussi mis au point par Paul de Casteljau (Ingénieur chez Citroën).

Elles ont de nombreuses applications dans la synthèse d'images et le rendu de polices de caractères. Elles ont donné naissance à de nombreux autres objets mathématiques. Les courbes de Bézier cubiques, les plus utilisées, se retrouvent en graphisme et dans de multiples systèmes de synthèse d'images, tels que PostScript, Adobe (pdf), Metafont et GIMP, pour dessiner des courbes "lisses" et vectorielles joignant des points ou des polygones de Bézier. Voir par exemple la figure[5] 2.16. Ces courbes n'utilisent en effet que peu de données, moins lourdes à stocker que les images des lettres faisant apparaître la trame du dessin. On comparera les figures 15(a) et 15(b).

5. Merci à Matthieu Cortat, `www.nonpareille.net`, Musée de l'imprimerie, Lyon.

(a) Vieux fichier sans courbes de Bézier (b) Fichier plus récent avec courbes de Bézier

FIGURE 2.15. Grossissement à 6400 % d'un fichier au format pdf de notes de cours.

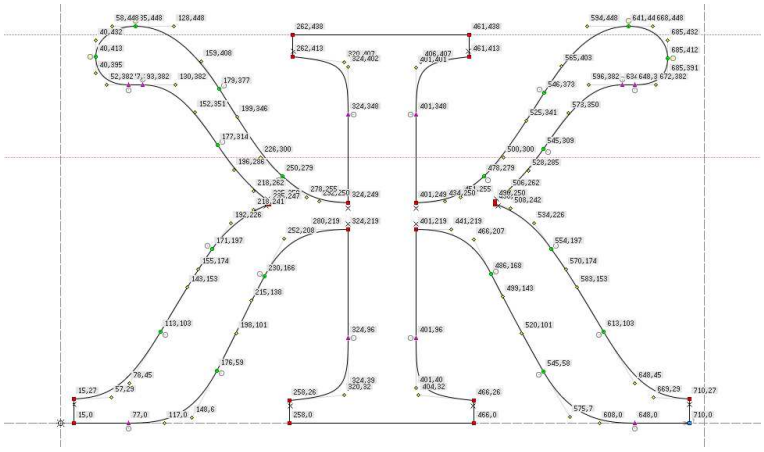


FIGURE 2.16. Le "J" cyrillique défini avec des courbes de Bézier.

2.7. Interpolation d'Hermite

Reprenons l'exemple 2.16 (ou 2.2), dans lequel on a supposé que $x_1 \neq x_0$. À x_0 fixé, faisons tendre x_1 vers x_0 : les équations (2.46a) et (2.46c) restent vraies, mais (2.46b) n'est plus valable puisque $x_1 = x_0$. Cependant, si on suppose f dérivable en x_0, on sait que

$$\lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(x_0), \tag{2.71}$$

et donc (2.46b) et (2.46c) deviennent

$$f[x_0, x_1] = f'(x_0), \tag{2.72}$$

et

$$\Pi_1(x) = f(x_0) + f'(x_0)(x - x_0), \tag{2.73}$$

ce qui est l'équation de la tangente à la courbe f au point x_0. Dans ce cas, les équations (2.2) sont réduites à la seule équation

$$\Pi_1(x_0) = f(x_0), \tag{2.74}$$

et Π_1 correspond à la fonction polynômiale associée à la tangente à la courbe en x_0, et on a

$$\Pi_1'(x_0) = f'(x_0). \tag{2.75}$$

Ainsi, le cas[6] $x_1 = x_0$ peut être encore traité et la notion de différences divisées est encore valable, définies par (2.46a), (2.72) et Π_1 est défini par (2.73) et vérifie (2.74) et (2.75).

Dans un cas plus général, il est tout à fait possible de considérer des nœuds d'interpolation qui ne soient pas deux à deux distincts, mais qui peuvent se confondre par des passages à la limite des cas distincts. Dans ce cas, les équations de la proposition 2.14 sont toujours valables quand les dénominateurs sont non nuls. Si les dénominateurs sont nuls, cela signifie des points x_i sont confondus ; dans ce cas, en généralisant ce qu'on a vu ci-dessous, on considère les dérivées successive de f. De même, les équations (2.2) sont complétées par les valeurs des dérivées successives de f en certains x_i. Plus de détails dans [CB81] et [BM03, Exercice 2.8 et TP 2.F]. On pourra aussi consulter l'exercice de TD 2.7.

Exemple 2.42. Dans [BM03, TP 2.F], on considère le support $\{0, 0, 1, 1\}$. Le polynôme Π_3 coïncide donc avec f en 0 et 1 et la dérivée Π_3' coïncide avec f' en 0 et 1. Ce polynôme peut être donc utilisé pour construire les slines cubiques de la section 2.6.1.

Exemple 2.43. Reprenons une des questions de [BM03, Exercice 2.8]. Déterminons le polynôme d'interpolation de $f : x \mapsto e^x$ sur le support $\{1, 2, 2, 2, 4, 4, 5\}$. Avec les notations précédentes $x_0 = 1$, $x_1 = x_2 = x_3 = 2$, $x_4 = x_5 = 4$ et $x_6 = 5$. Il vient

$$f[x_0] = f(x_0) = e^{x_0} \approx 2.718$$

De même

$$\begin{aligned} f[x_1] = f[x_2] = f[x_3] &\approx 7.389 \\ f[x_4] = f[x_5] &\approx 54.60 \\ f[x_6] &\approx 148.4 \end{aligned}$$

On en déduit par exemple

$$\begin{aligned} f[x_0, x_1] &= \frac{f(x_0) - f(x_1)}{x_1 - x_0} \approx 4.671, \\ f[x_1, x_2] &= \frac{f'(x_1)}{1!} = e^2 \approx 7.389 \end{aligned}$$

et

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_1} \approx 2.718 \\ f[x_1, x_2, x_3] &= \frac{f''(x_1)}{2!} \approx 3.694 \end{aligned}$$

On remplit successivement la table et on obtient les valeurs des différences divisées dans le tableau 2.7.

On utilisant la diagonale descendante (en gras), on en déduit le polynôme p_6 d'interpolation sous la forme de Newton :

$$\begin{aligned} \Pi_6(x) = 2.718 + 4.671(x-1) + 2.718(x-1)(x-2) + 0.975(x-1)(x-2)^2 \\ + 0.410(x-1)(x-2)^3 + 0.112(x-1)(x-2)^3(x-4) + 0.027(x-1)(x-2)^3(x-4)^2. \end{aligned}$$

Voir la figure 2.17 montrant que f, f' et f'' correspondent bien avec Π_6, Π_6' et Π_6'' aux nœuds.

♢

2.8. Approximation au sens des moindres carrés

Une autre façon d'avoir à utiliser un grand degré de polynôme, notamment quand les données sont très nombreuses consiste à écrire l'égalité (2.5) non pas au sens exact, mais au sens des moindres carrés, c'est-à-dire

6. Historiquement, la de différence divisée a été mise au point par Newton, et ce, avant la notion de dérivée. Celle-ci serait née de la considération du cas limite x_1 tendant vers x_0 par Newton, parallèlement aux travaux de Leibniz.

x	$f^{[0]}$	$f^{[1]}$	$f^{[2]}$	$f^{[3]}$	$f^{[4]}$	$f^{[5]}$	$f^{[6]}$
$x_0 = 1$	**2.718**						
		4.671					
$x_1 = 2$	7.389		**2.718**				
		7.389		**0.975**			
$x_2 = 2$	7.389		3.694		**0.410**		
		7.389		2.206		**0.112**	
$x_3 = 2$	7.389		8.108		0.774		**0.027**
		23.60		3.694		0.220	
$x_4 = 4$	54.60		15.49		1.404		
		54.60		7.908			
$x_5 = 4$	54.60		39.20				
		93.80					
$x_6 = 5$	148.4						

TABLE 2.7. Différences divisées correspondant à l'interpolation de la fonction $f : x \mapsto e^x$ sur le support $\{1, 2, 2, 2, 4, 4, 5\}$

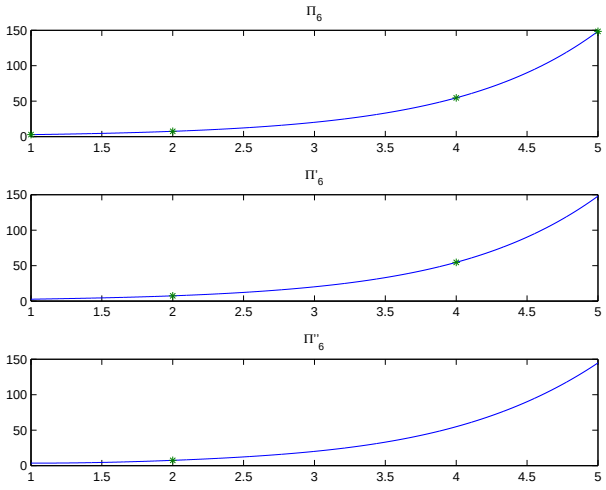


FIGURE 2.17. Fonctions Π_6, Π_6' et Π_6''.

que l'on cherche un polynôme p passant le plus près possible d'un nuage de points donnés ; on minimise la somme des carrés des écarts entre la valeur du polynôme en x_i et la valeurs y_i, c'est-à-dire

$$S = \sum_{i=0}^{n} \left(p(x_i) - y_i\right)^2.$$

Ce polynôme est unique. Il coïncide avec le polynôme d'interpolation si le degré du polynôme recherché est au plus égal au nombre de points moins un.

On pourra consulter par exemple [LT93, Chapitre 6], [Cia82, Sections 3.7 p. 69 et 7.4 p. 157] [Bas22a, section 4.3. "Étude d'un exemple concret"] disponible sur `http://utbmjb.chez-alice.fr/Polytech/MFI/coursMFI.pdf` ou l'annexe D. Attention, dans cette annexe, il convient de remplacer dans (D.5) et (D.6), b_i par y_i et $\sum_{j=1}^{p} a_{ij}x_j$ par $\sum_{j=0}^{n} \alpha_j x_i^j$, où

$$p(x) = \sum_{j=0}^{n} \alpha_j x^j.$$

Prendre garde aussi au fait que dans cette annexe, les points ici notés x_0, ..., x_n et les valeurs notées y_0, ..., y_n sont respectivement notés x_1, ..., x_n et y_1, ..., y_n. Dans cette annexe, sont montrés l'existence et l'unicité du polynôme p, défini au sens des moindres carrés. Voir le lemme D.2 et la proposition D.4.

Remarque 2.44. L'hypothèse de la proposition D.4 (la matrice A est de rang p) est vérifiée dés que, parmi les valeurs des x_i, il y a au moins p valeurs distinctes. En effet, dans ce cas, on raisonne comme dans la remarque D.3 : on peut extraire de A, une sous-matrice de type

$$\begin{pmatrix} 1 & x_{i_1} & x_{i_1}^2 & \dots & x_{i_1}^p \\ & \vdots & & \vdots & \\ 1 & x_{i_p} & x_{i_p}^2 & \dots & x_{i_p}^p \end{pmatrix}$$

où les indices i_q, pour $q \in \{1, ..., p\}$ sont deux à deux distincts, les x_{i_q} sont deux à deux distincts et l'on sait que cette matrice est inversible (voir par exemple la section 2.2.2.1).

♢

Exemple 2.45.

On considère un nuage [7] de 200 points. Dans ce cas, le lemme D.2 s'applique. Pour quelques valeurs de N décrivant l'ensemble $\{1, 3, 5, 7, 10, 20, 40\}$, on construit et trace le polynôme de degré N interpolant les (x_i, y_i) au sens des moindres carrés.

Voir figure 2.18.

On prendra garde au fait, que pour la dernière valeur de N choisie, le polynôme prend de "grandes valeurs" au voisinage de $x = 1$, comme le montre la figure 2.19.

Exemple 2.46. Voir de nouveau l'exemple de la section 2.1.

7. Il est défini en fait à partir de la fonction f de l'exemple 2.39 : on considère 200 points $(x_i)_{1 \leq i \leq 200}$ aléatoires dans $[0, 1]$. Pour tout i, on pose $y_i = F(x_i) + \varepsilon_i$ où les ε_i sont des nombres aléatoires dans $[-R, R]$ avec $R = 0.50$.

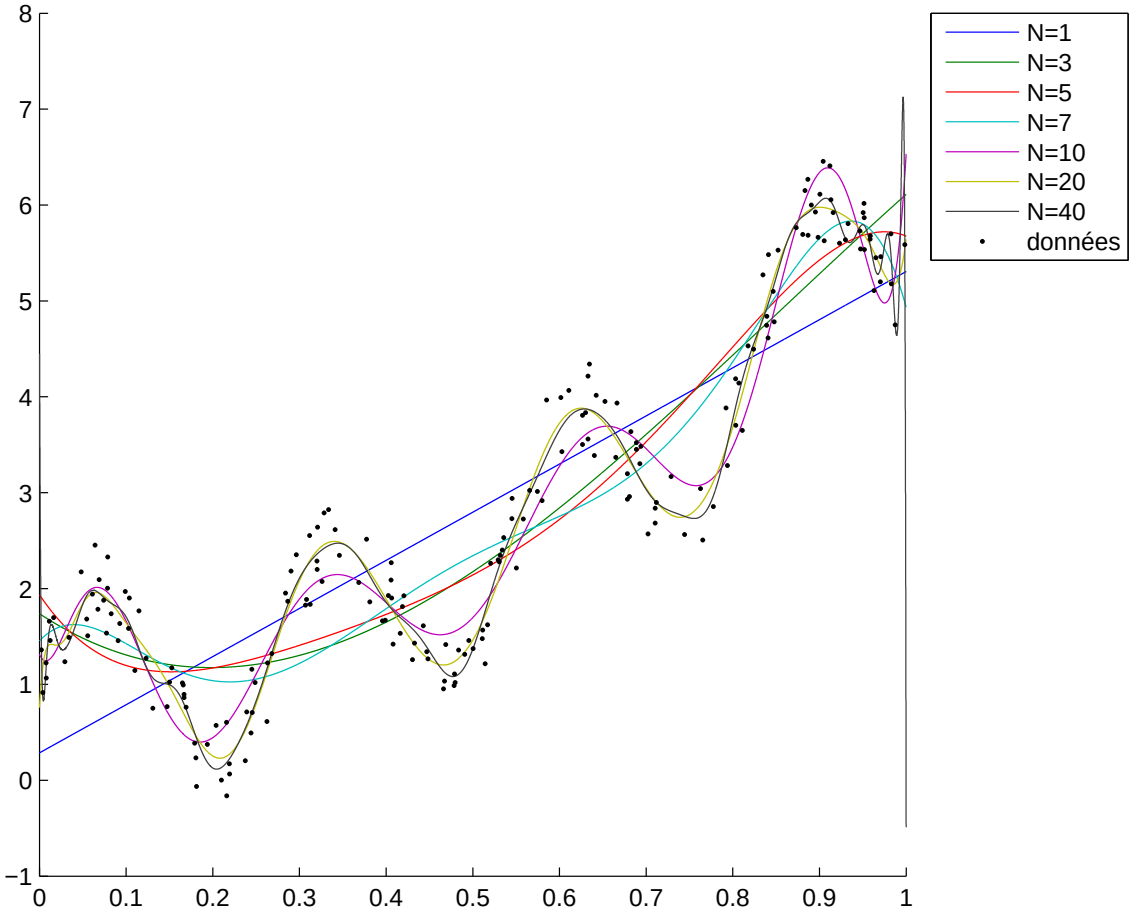


FIGURE 2.18. Interpolation au sens des moindres carrés.

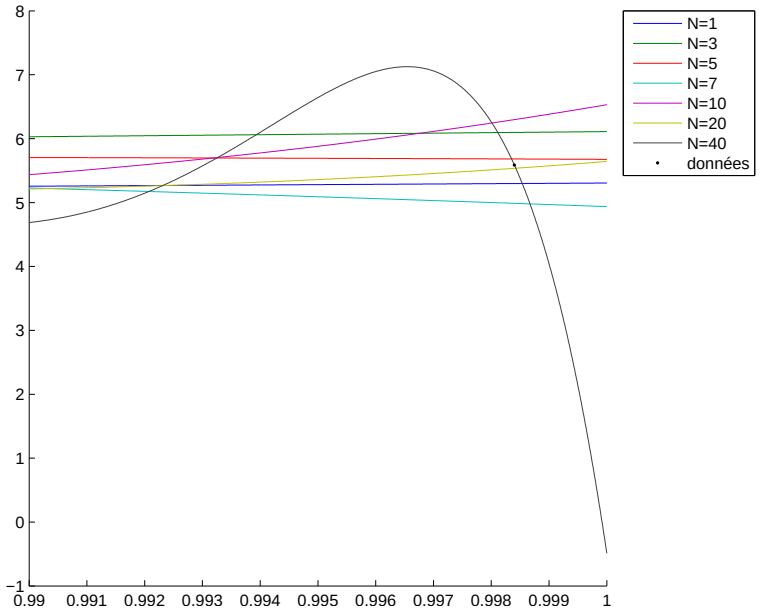


FIGURE 2.19. Interpolation au sens des moindres carrés, au voisinage de $x = 1$.

Chapitre 3

Intégration

3.1. Motivation

La chaleur spécifique C_v (en J/K/mol) d'un solide monoatomique varie en fonction de la température absolue T suivant la loi :

$$C_v = \frac{9R}{x_m^3} \int_0^{x_m} \frac{e^x x^4}{\left(e^x - 1\right)^2} dx, \tag{3.1}$$

où $R = 8.3140$ J/K/mol et $x_m = \Theta_D/T$, D étant la température de Debye, qui dépend du solide considéré. On souhaite tracer l'évolution de cette chaleur spécifique en fonction de la température T pour $T \in [T_{\min}, T_{\max}]$: où

$$T_{\min} = 10, \tag{3.2a}$$

$$T_{\max} = 500. \tag{3.2b}$$

Considérons f définie (prolongée par continuité) par

$$\forall x \in \mathbb{R}_+, \quad f(x) = \begin{cases} 0, & \text{si } x = 0, \\ \frac{e^x x^4}{(e^x-1)^2}, & \text{si } x \neq 0. \end{cases} \tag{3.3}$$

On a donc

$$C_v = \frac{9R}{x_m^3} \int_0^{x_m} f(x) dx,$$

En posant, pour tout T non nul, $u = \Theta_D/T$, on a

$$C_v = \frac{9R}{u^3} \int_0^{u} f(x) dx.$$

♢

Il faut donc calculer :

$$\forall u \in [\Theta_D/T_{\max}, \Theta_D/T_{\min}], \quad C_v(u) = \frac{9R}{u^3} \int_0^{u} f(x) dx. \tag{3.4}$$

Déterminons la valeurs de l'intégrale pour plusieurs méthodes : en utilisant les fonctions `quadl` de matlab, puis par les méthodes des rectangles, des trapèzes, des milieux et de Simpson.

La chaleur spécifique pour $T = 300$, correspondant à $u = 1.0432$, dans le cas du cuivre, est donnée donc pour chacune des méthodes (en prenant 10000 sous-intervalles) par

$$C_v = 23.6355326661,$$
$$C_v = 23.6321131478,$$
$$C_v = 23.6355327699,$$
$$C_v = 23.6355326142,$$
$$C_v = 23.6355326661.$$

Traçons maintenant la fonction définie par (3.4) en prenant 4000 nombre de points de calcul noté u_i et sur chacun des intervalles $[u_i, u_{i+1}]$, 1000 sous-intervalles. On utilise les fonctions `quadl` et `quadv` de matlab, puis par les méthodes des rectangles, des trapèzes, des milieux et de Simpson. Voir la figure 3.1.

Les temps de calculs sont donnés dans le tableau 3.1.

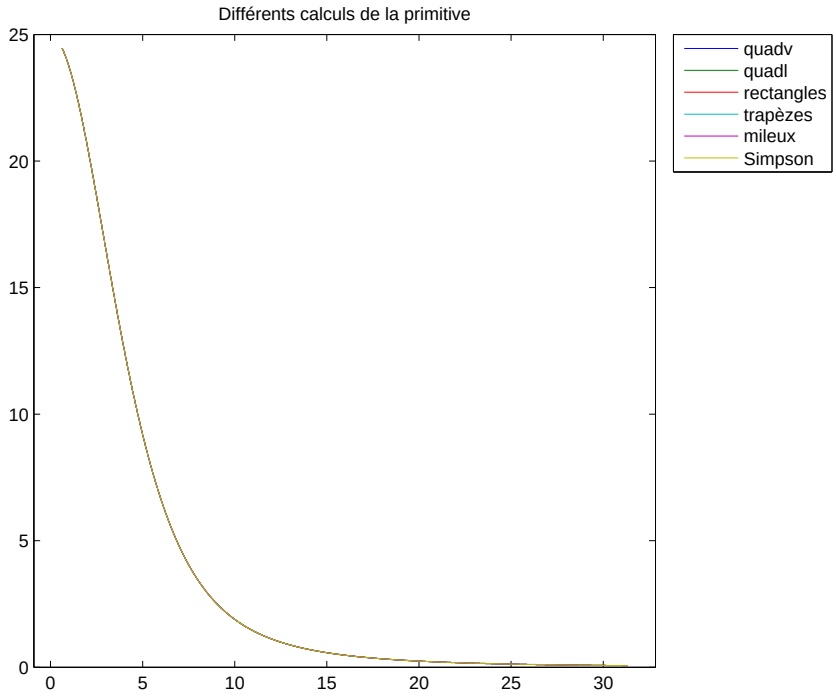


FIGURE 3.1. Différents calculs de la fonction définie par (3.4).

`quadv`	`quadl`	rectangles	trapèzes	milieux	Simpson
0.02991	0.49437	1.02362	1.15366	1.02047	2.17541

TABLE 3.1. Temps de calcul des différentes méthodes de calcul de la fonction définie par (3.4)

3.2. Introduction informelle

Citons quelques pages extraites de [Bas22a, Chapitre "Intégration (théorie)"].

Si on considère un solide évoluant sur un axe rectiligne, dont l'abscisse est notée $x(t)$, nous passons du déplacement à la vitesse et de la vitesse à l'accélération par dérivation, ce qui graphiquement, correspond à prendre la tangente à la courbe.

Rappelons que si d est la distance parcourue, pendant un temps t, la vitesse moyenne est définie par

$$v = \frac{d}{t}. \tag{3.5}$$

Cette formule définit aussi la vitesse instantanée à tout instant, si celle-ci est constante.

Si $x(t)$ est connue, la vitesse moyenne v sur l'intervalle de temps $[t, t+T]$, la vitesse moyenne sur cet intervalle de temps est définie par

$$v = \frac{x(t+T) + x(t)}{T}, \tag{3.6}$$

parfois notée sous la forme

$$v = \frac{\Delta x}{\Delta t}. \tag{3.7}$$

La vitesse instantanée à l'instant t est

$$v(t) = x'(t), \tag{3.8}$$

noté aussi sous une forme analogue à (3.7)

$$v = \frac{dx}{dt}, \tag{3.9}$$

et en confondant parfois abusivement Δx et dx, et Δt et dt, quand dt est « petit ».

On pourra consulter [Bas18, chapitre 4].

Supposons maintenant la courbe v connue et quelconque. On cherche à déterminer x.

Plus précisément, on se donne $a < b$; on se suppose connue $x(a)$, la fonction v sur l'intervalle $[a, b]$ et on cherche à calculer $x(b)$.

Pour cela, on se donne un entier N et on découpe l'intervalle $[a, b]$ en N intervalle $[t_i, t_{i+1}]$ de la façon suivante : on pose

$$\begin{aligned} &t_0 = a, \\ &h = \frac{b-a}{N}, \\ &\forall i \in \{0, ..., N\}, \quad t_i = hi + a. \end{aligned}$$

On a donc $t_N = b$.

Voir figure 3.2.

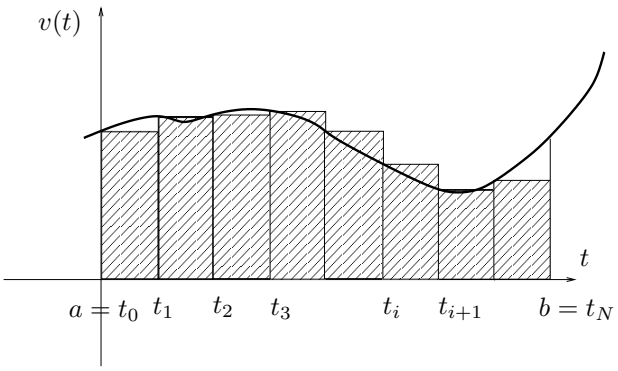


FIGURE 3.2. L'aire sous la courbe avec des rectangles

Soit i dans $\{0, ..., N-1\}$ fixé. Nous allons «tricher» et supposer que, dans l'intervalle $[t_i, t_{i+1}]$, la vitesse v varie peu, de façon à remplacer la vitesse *a priori* quelconque par la vitesse constante $v(t_i)$. Cette approximation sera d'autant meilleure que h est petit (c'est-à-dire N grand). On a donc, pour tout $t \in [t_i, t_{i+1}]$,

$$v(t) \approx v(t_i)$$

La vitesse est constante et d'après (3.5), on a

$$v(t) \approx v(t_i) = \frac{\Delta x}{\Delta t} = \frac{x(t) - x(t_i)}{t - t_i}$$

et donc

$$x(t) \approx (t - t_i)v(t_i) + x(t_i)$$

En particulier

$$x(t_{i+1}) \approx (t_{i+1} - t_i)v(t_i) + x(t_i) = hv(t_i) + x(t_i).$$

soit encore

$$x(t_{i+1}) - x(t_i) \approx (t_{i+1} - t_i)v(t_i) = A_i, \tag{3.10}$$

où

$$A_i = hv(t_i). \tag{3.11}$$

Notons que $A_i = hv(t_i)$ représente l'aire sous la courbe $v(t)$ où $v(t) \approx v(t_i)$ entre t_i et t_{i+1}. C'est l'aire du rectangle de largeur h et de hauteur $v(t_i)$. Voir figure 3.2 page précédente.

On en déduit successivement

$$\begin{aligned}
&x(t_1) \approx hv(t_1) + x(t_0) = hv(t_0) + x(a),\\
&x(t_2) \approx hv(t_2) + x(t_1) = h(v(t_0) + v(t_1)) + x(a),\\
&x(t_3) \approx h(v(t_0) + v(t_1) + v(t_2)) + x(a),\\
&\vdots\\
&x(t_{i+1}) \approx hv(t_i) + x(t_i) = h(v(t_0) + v(t_1) + v(t_2) + ... + v(t_i)) + x(a),\\
&\vdots\\
&x(t_N) \approx h(v(t_0) + v(t_1) + v(t_2) + ... + v(t_i) + ... + v(t_{N-1})) + x(a).
\end{aligned}$$

Autrement dit

$$x(b) - x(a) \approx h(v(t_0) + v(t_1) + v(t_2) + ... + v(t_i) + ... + v(t_{N-1})) = h\sum_{i=0}^{N-1} v(t_i). \tag{3.12}$$

Cette formule fait apparaître «l'aire des rectangles», hachurée sur la figure 3.2 page précédente. Quand le nombre N tend vers l'infini, cette aire tend vers l'aire qui est sous la courbe v entre a et b. Cette aire est notée $\int_a^b v(s)ds$. On a donc montré que

$$x(b) - x(a) = \int_a^b v(s)ds. \tag{3.13}$$

L'équation (3.12) pourrait constituer une définition de l'intégrale, en passant à la limite. Elle constitue aussi une approximation de cette aire.

Remarque 3.1 (Méthodes des rectangles à pas variable). Dans la méthode des rectangles, on n'est pas obligé de prendre un pas h constant. On peut découper l'intervalle en sous-intervalle de taille variable et remplacer (3.10) et (3.11) par

$$x(t_{i+1}) - x(t_i) \approx A_i, \tag{3.14}$$

où

$$A_i = (t_{i+1} - t_i)v(t_i). \tag{3.15}$$

Remarque 3.2 (Méthodes des trapèzes à pas variable).

On peut aussi utiliser la méthode des trapèzes à pas variable, plus précise que celle des rectangles : (Voir figure 3.3 page suivante) on remplace, sur chaque intervalle $[t_i, t_{i+1}]$, $v(t)$ par une vitesse v linéaire. De sorte que l'aire approchée est remplacée par l'aire des trapèzes. On a donc

$$x(t_{i+1}) - x(t_i) \approx A_i, \tag{3.16}$$

où

$$A_i = \frac{1}{2}(t_{i+1} - t_i)(v(t_i) + v(t_{i+1})). \tag{3.17}$$

On a enfin

$$x(b) - x(a) = \int_a^b v(s)ds \approx \sum_{i=0}^{N-1} \frac{1}{2}(t_{i+1} - t_i)(v(t_i) + v(t_{i+1})). \tag{3.18}$$

Souvent, cette formule est utilisée pour h constant.

Remarque 3.3. Se rappeler que la primitive ainsi informellement définie correspond à l'opération inverse de la dérivation. On pourra consulter l'annexe F.

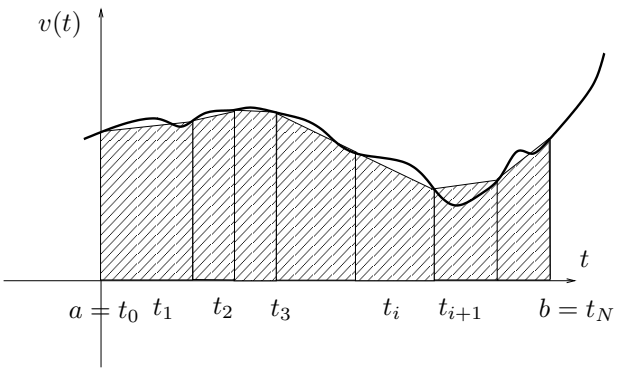


FIGURE 3.3. L'aire sous la courbe avec des trapèzes

3.3. Méthodes élémentaires et composites (composées)

3.3.1. Définition et propriétés des méthodes élémentaires

Dans la section 3.2, on a défini une surface approchée sous la courbe en remplaçant cette dernière par une fonction constante sur chaque intervalle (voir remarque 3.1) ou affine sur chaque intervalle (voir remarque 3.2). En fait, cela peut se généraliser et se formaliser en utilisant la théorie de l'interpolation polynômiale du chapitre 2.

Cette section correspond à [BM03, Section 3.2.1].

3.3.1.1. Notations.

Soit f une fonction régulière sur le fermé borné $[a, b]$ de $\mathbb{R}$; f sera au moins de classe C^1 et au plus de classe C^4.

Pour n dans $\mathbb{N}$, considérons le support $\{x_0, ..., x_n\}$ formé de points quelconques et deux à deux distincts de $[a, b]$. Soit Π_n le polynôme d'interpolation de f associé. D'après la proposition 2.24, si f est de classe $\mathcal{C}^{n+1}$ sur $[a, b]$, alors

$$\int_a^b f(x)dx = \int_a^b \Pi_n(x)dx + \int_a^b E_n(x)dx, \tag{3.19}$$

où E_n est donnée par l'expression (2.52) figurant dans la preuve, soit

$$E_n(x) = f(x) - \Pi_n(x) = f\,[x_0, ..., x_n, x]\,\omega_{n+1}(x), \tag{3.20}$$

où ω_{n+1} est défini par (2.51).

Nous notons

- I_n la valeur approchée de I définie par

$$I_n = \int_a^b \Pi_n(x)dx\,; \tag{3.21}$$

- $\mathcal{E}_n$ l'erreur d'intégration qui est l'intégration de l'erreur d'interpolation fournie par :

$$\mathcal{E}_n = \int_a^b f\,[x_0, ..., x_n, x]\,\omega_{n+1}(x)dx, \tag{3.22}$$

Ainsi, nous avons

$$\int_a^b f(x)dx = I_n + \mathcal{E}_n. \tag{3.23}$$

REMARQUE 3.4. L'application $x \mapsto f\,[x_0, ..., x_n, x]$ est définie pour tout x, même égal à l'un des x_i, comme déjà écrit dans la preuve de la proposition 2.25. Voir [BM03, exercice 2.8 et TP 2.F]. De plus, cette application est dérivable un certain nombre de fois, si f l'est aussi.

♢

REMARQUE 3.5. On peut aussi remarquer que Π_n étant aussi défini par (2.27), on a aussi

$$I_n = \int_a^b \Pi_n(x)dx = \int_a^b \sum_{i=0}^{n} f(x_i)l_i(x),$$

soit

$$I_n = \sum_{i=0}^{n} W_i f(x_i), \tag{3.24}$$

où

$$W_i = \int_a^b l_i(x)dx. \tag{3.25}$$

♢

Les méthodes élémentaires d'intégration vont simplement s'interpréter comme des cas particuliers selon le choix du degré n du polynôme d'interpolation.

3.3.1.2. Interpolation par une fonction Π_n de degré $n = 0$.

PROPOSITION 3.6. *Dans le cas de degré $n = 0$, le support d'interpolation est réduit à $\{x_0\}$. Si f est de classe C^1 sur $[a, b]$, la valeur approchée est :*

$$I_0 = (b - a)f(x_0),$$

et l'erreur d'intégration est

$$\mathcal{E}_0 = \int_a^b f[x_0, x]\,(x - x_0)\,dx.$$

DÉMONSTRATION.

(1) Ici $I_0 = \int_a^b \Pi_0(x)dx$ avec $\Pi_0(x)$ défini par $\Pi_0(x) = f[x_0] = f(x_0)$ d'où le résultat.

(2) Le caractère C^1 de f garantit la continuité de l'application : $x \mapsto f[x_0, x]$ et donc l'existence de

$$\mathcal{E}_0 = \int_a^b f[x_0, x]\,\omega_1(x)dx,$$

avec $\omega_1(x)$ égal à

$$\omega_1(x) = (x - x_0).$$

□

♢

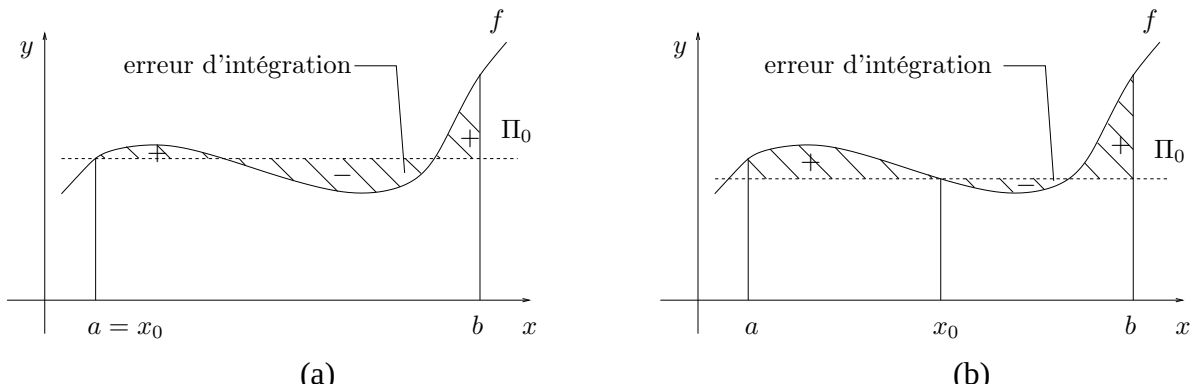


FIGURE 3.4. Les fonctions f et Π_0 et l'erreur d'intégration pour la méthode du rectangle, à gauche, (a) (si $x_0 = a$) et la méthode du point milieu (b).

En choisissant x_0 comme extrémité de l'intervalle $[a, b]$, il vient (voir figure 3.4a)

COROLLAIRE 3.7 (Méthode du rectangle). *Si f est de classe C^1 sur $[a,b]$, alors, pour $x_0 = a$, la valeur approchée vaut*

$$I^R = (b-a)f(a), \tag{3.26}$$

l'erreur d'intégration vaut

$$\mathcal{E}^R = \frac{(b-a)^2}{2} f'(\eta) \ \textit{avec } \eta \in]a,b[\,, \tag{3.27}$$

et, pour $x_0 = b$,

$$I^R = (b-a)f(b), \tag{3.28}$$

l'erreur d'intégration vaut

$$\mathcal{E}^R = \frac{(b-a)^2}{2} f'(\eta) \ \textit{avec } \eta \in]a,b[\,. \tag{3.29}$$

DÉMONSTRATION. Les deux cas sont similaires ; on traite le cas $x_0 = a$.
Présentons deux preuves :

(1) L'expression de la valeur approchée I^R provient immédiatement de la proposition 3.6. Voir aussi l'annexe H. De même l'erreur d'intégration s'écrit ici

$$\mathcal{E}^R = \int_a^b f\,[a,x]\,(x-a)\,dx.$$

On remarque que $(x-a)$ garde un signe constant sur $[a,b]$; ainsi, d'après [BM03, le corollaire 3.4 (cas 1 : (3.18))] avec $n=0$ et $r=f$ sur $[a,b]$)

$$\mathcal{E}^R = \frac{f'(\eta)}{(1)!}\int_a^b \omega_1(x)dx = f'(\eta)\int_a^b (x-a)dx,$$

avec $\eta \in]a,b[$ et donc

$$\mathcal{E}^R = \frac{(b-a)^2}{2} f'(\eta). \tag{3.30}$$

(2) Plus simplement, dans le particulier étudié ici, on peut écrire

$$\mathcal{E}^R = \int_a^b f(x) - \Pi_0(x)dx = \int_a^b f(x) - f(a)dx,$$

et d'après le théorème des accroissements finis, pour tout $x \in [a,b]$, il existe $c_x \in]a,x|$ tel que

$$f(x) = f(a) + f'(c_x)(x-a),$$

et donc, d'après ce qui précède :

$$\mathcal{E}^R = \int_a^b f'(c_x)(x-a)dx,$$

et d'après le théorème de la moyenne, puisque $f'(c_x)$ est continu par rapport à x et $x-a$ est de signe constant sur $[a,b]$ (voir [BM03, Théorème A.8 p.327]), il existe $\eta \in]a,b[$ tel que

$$\mathcal{E}^R = f'(\eta)\int_a^b (x-a)dx,$$

et on conclue comme dans (3.30).

□

♢

En choisissant pour x_0 le milieu de l'intervalle $[a,b]$, il vient (voir figure 3.4b) :

COROLLAIRE 3.8 (Méthode du point milieu). *Si f est de classe C^2 sur $[a,b]$, alors la valeur approchée vaut*

$$I^M = (b-a)f\left(\frac{a+b}{2}\right), \tag{3.31}$$

et l'erreur d'intégration vaut

$$\mathcal{E}^M = \frac{(b-a)^3}{24} f''(\eta) \ \textit{avec } \eta \in]a,b[. \tag{3.32}$$

Démonstration. On applique encore la proposition 3.6 d'où l'expression de I^M et

$$\mathcal{E}^M = \int_a^b f\,[m,x]\,(x-m)\,dx \text{ avec } m = \frac{a+b}{2}.$$

Ici, $(x-m)$ change de signe sur $[a,b]$. On peut montrer (voir [BM03]) que

$$\mathcal{E}^M = \int_a^b f\,[x_0,x_1,x]\,(x-x_0)\,(x-x_1)\,dx = \int_a^b f\,[m,m,x]\,(x-m)^2 dx.$$

On a alors, puisque, f est de classe C^2 sur $[a,b]$,

$$\mathcal{E}^M = \frac{f^{(2)}(\eta)}{(2)!}\int_a^b \omega_2(x)dx = \frac{f''(\eta)}{2}\int_a^b (x-m)^2 dx, \text{ avec } \xi \in]a,b[\,.$$

On conclut en calculant explicitement l'intégrale. □

♢

Remarque 3.9. Ce phénomène de «doublement» du point $x_0 = (a+b)/2$ est très important ; certes nous avons interpolé f en un seul point x_0 mais il «travaille comme deux»... Il convient de remarquer que l'erreur d'intégration commise a gagné un ordre [1] : elle passe d'une forme $\alpha(b-a)^2$ à $\beta(b-a)^3$.

3.3.1.3. Interpolation par une fonction Π_n de degré $n = 1$.

Proposition 3.10. *Dans le cas de degré $n = 1$, le support d'interpolation est égal à $\{x_0, x_1\}$. Si f est de classe C^2 sur $[a,b]$, la valeur approchée est :*

$$I_1 = \int_a^b (f[x_0] + f[x_0,x_1](x-x_0))\,dx,$$

et l'erreur d'intégration est

$$\mathcal{E}_1 = \int_a^b f[x_0,x_1,x]\,(x-x_0)\,(x-x_1)\,dx.$$

Démonstration. On a $I_1 = \int_a^b \Pi_1(x)dx$ où Π_1 désigne le polynôme d'interpolation de f sur $\{x_0,x_1\}$; d'où le résultat.
Le caractère C^2 de f garantit l'existence dans tous les cas de $f\,[x_0,x_1,x]$ puis l'écriture proposée de $\mathcal{E}_1$, puisque $\omega_2(x) = (x-x_0)\,(x-x_1)$. □

♢

En choisissant $x_0 = a$ et $x_1 = b$, il vient (voir figure 3.5a)

Corollaire 3.11 (Méthode du trapèze). *Si f est de classe C^2 sur $[a,b]$, alors la valeur approchée vaut*

$$I^T = (b-a)\left(\frac{f(a)+f(b)}{2}\right), \tag{3.33}$$

et l'erreur d'intégration vaut

$$\mathcal{E}^T = -\frac{(b-a)^3}{12} f''(\eta) \quad \textit{avec } \eta \in]a,b[. \tag{3.34}$$

1. Voir définition 3.35.

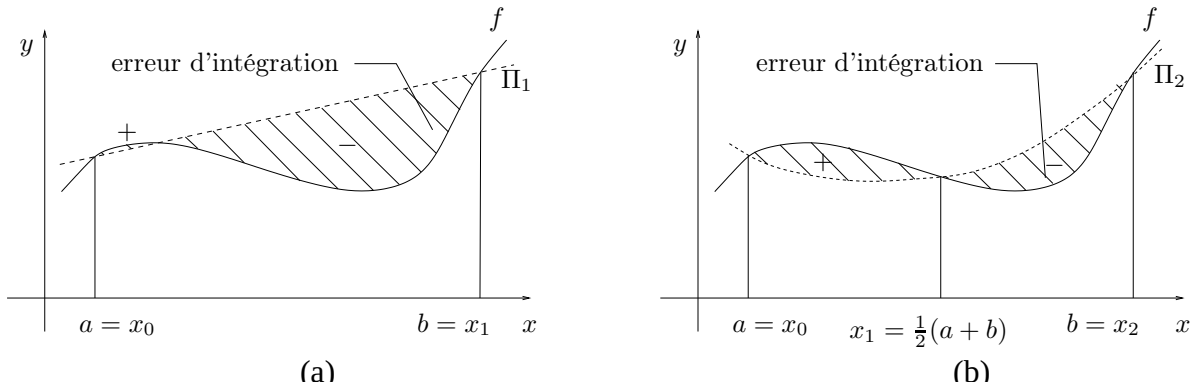


FIGURE 3.5. Les fonctions f et Π_n et l'erreur d'intégration pour la méthode du trapèze (a) (avec $n = 1$) et la méthode de Simpson (b) (avec $n = 2$).

DÉMONSTRATION. En termes de valeur approchée, $I^T = \int_a^b \Pi_1(x)dx$ où Π_1 désigne le polynôme d'interpolation de f sur $\{a, b\}$. Ainsi

$$I^T = \int_a^b f[a]\,dx + \int_a^b f[a,b]\,(x-a)dx = (b-a)f(a) + \frac{f(b)-f(a)}{b-a}\frac{(b-a)^2}{2}, \tag{3.35}$$

d'où le résultat annoncé qu'on aurait pu déduire d'une interprétation géométrique, via l'aire d'un trapèze. Voir aussi l'annexe H.

Pour l'erreur d'intégration,

$$\mathcal{E}_1 = \int_a^b f[a,b,x]\,(x-a)\,(x-b)\,dx,$$

on peut montrer (voir [BM03]) que

$$\mathcal{E}_1 = \frac{f''(\eta)}{(2)!}\int_a^b (x-a)\,(x-b)\,dx, \text{ avec } \eta \in]a,b[\,.$$

On conclut en calculant l'intégrale. □

♢

3.3.1.4. Interpolation par une fonction Π_n de degré $n = 2$.

PROPOSITION 3.12. *Dans le cas de degré $n = 2$, le support d'interpolation est égal à à $\{x_0, x_1, x_2\}$. Si f est de classe C^3 sur $[a, b]$, la valeur approchée est :*

$$I_2 = \int_a^b \left(f[x_0] + f[x_0,x_1]\,(x-x_0) + f[x_0,x_1,x_2]\,(x-x_0)\,(x-x_1)\right)dx,$$

et l'erreur d'intégration est

$$\mathcal{E}_2 = \int_a^b f[x_0,x_1,x_2,x]\,(x-x_0)\,(x-x_1)\,(x-x_2)\,dx.$$

DÉMONSTRATION. Similaire à celle de la proposition 3.10. □

♢

En choisissant $x_0 = a$, $x_1 = \frac{a+b}{2}$ et $x_2 = b$. (voir figure 3.5b), il vient

COROLLAIRE 3.13 (Méthode de Simpson). *Si f est de classe C^4 sur $[a, b]$, alors la valeur approchée vaut*

$$I^S = \frac{(b-a)}{6}\big(f(a) + 4f(m) + f(b)\big) \text{ où } m = \frac{a+b}{2}, \tag{3.36}$$

et l'erreur d'intégration vaut

$$\mathcal{E}^S = -\frac{(b-a)^5}{2880}f^{(4)}(\eta). \tag{3.37}$$

DÉMONSTRATION.

(1) *Etude de la valeur approchée*

Notons $m = (a+b)/2$. On désigne par I^S l'intégrale $\int_a^b \Pi_2(x)dx$ où Π_2 interpole f sur $\{a, m, b\}$. Voir le calcul de l'annexe H. On remarquera que I^S est le barycentre de

$$f(a)\left(\frac{(b-a)}{6}\right), \quad f(m)\left(\frac{4(b-a)}{6}\right) \text{ et } f(b)\left(\frac{(b-a)}{6}\right).$$

(2) *Étude de l'erreur d'intégration*

On a

$$\mathcal{E}_2 = \int_a^b f\,[a, m, b, x]\,(x-a)\,(x-m)\,(x-b)\,dx.$$

On peut montrer (voir [BM03]) que

$$\mathcal{E}_2 = \frac{f^4(\eta)}{(4)!}\int_a^b (x-a)\,(x-m)^2\,(x-b)\,dx.$$

On conclut en calculant explicitement l'intégrale.

□

♢

REMARQUE 3.14. Ici encore, il n'y a que trois points dans le support mais ils interviennent «comme quatre» d'où la bonne performance numérique de la méthode de Simpson ; cette particularité explique que l'ordre [2] soit supérieur à celui que laissait prévoir le nombre de points : la formule de Simpson est exacte pour toute fonction polynôme de degré trois.

REMARQUE 3.15. Notons d'après les formules (3.27), (3.29), (3.32), (3.34) et (3.37) que

- la formule du rectangle est exacte pour des polynômes de degré 0,
- la formule du milieu est exacte pour des polynômes de degré 1 ;
- la formule du trapèze est exacte pour des polynômes de degré 1 ;
- la formule de Simpson est exacte pour des polynômes de degré 3.

Notons dans le tableau suivant le degré des formules simples d'intégration, c'est-à-dire le degré du plus grand polynôme intégré exactement par la formule et le nombre de points du support :

méthode	degré	nombre de points
rectangle	0	1
milieu	1	1
trapèze	1	2
Simpson	3	3

REMARQUE 3.16. Remarquons aussi que chacune des formules (3.26), (3.28), (3.31), (3.33) et (3.36) peut s'écrire sous la forme

$$\int_a^b f(x)dx \approx \sum_{i=0}^{m} W_i f(x_i), \tag{3.38}$$

où m est un entier naturel, les x_i éléments de $[a, b]$ et les W_i des réels. Une telle relation est appelée formule de quadrature. Voir aussi la remarque 3.5 et la section 3.4.

Les formules d'intégration élémentaires et les erreurs correspondantes sont résumées dans les tableaux 3.2 et 3.3.

3.3.1.5. Interpolation par une fonction Π_n de degré $n = 3$.

Cette formule n'est pas utilisée en pratique, néanmoins présentés dans l'annexe I.

♢

2. Voir définition 3.35.

méthode	formule
rectangle (gauche)	$(b-a)f(a)$
rectangle (droite)	$(b-a)f(b)$
milieu	$(b-a)f((a+b)/2)$
trapèze	$\frac{1}{2}(b-a)(f(a)+f(b))$
Simpson	$\frac{1}{6}(b-a)(f(a)+4f((a+b)/2)+f(b))$

TABLE 3.2. Méthodes élémentaires sur $[a,b]$.

méthode	erreur
rectangle	$\frac{(b-a)^2}{2}f'(\eta)$
milieu	$\frac{(b-a)^3}{24}f''(\eta)$
trapèze	$-\frac{(b-a)^3}{12}f''(\eta)$
Simpson	$-\frac{(b-a)^5}{2880}f^{(4)}(\eta)$

TABLE 3.3. Erreurs des méthodes élémentaires sur $[a,b]$. Dans ce tableau, η appartient à $]a,b[$.

3.3.1.6. Interpolation par une fonction Π_n de degré n quelconques équirépartis : formule de Newton-Cotes.

Voir section 3.4.1.

♢

3.3.2. Définition et propriétés des méthodes composites (composées) à pas constant.

Cette section correspond à [BM03, Section 3.2.2].

3.3.2.1. Problématique et principe.

Soit f une fonction régulière sur $[A,B]$.

Nous souhaitons fournir une valeur approchée de $\int_A^B f(x)dx$ ainsi que l'expression de l'erreur d'intégration commise.

Soient N un entier naturel non nul et n un entier naturel.

Découpons $[A,B]$ en sous-intervalles à pas constant h ($h \in \mathbb{R}_+^*$), notés $[x_i, x_{i+1}]$. Ainsi

$$x_0 = A, \quad x_N = B, \quad \forall i \in \{0,...,N-1\} \quad x_{i+1} - x_i = h \quad \text{d'où} \quad h = \frac{B-A}{N}. \tag{3.39}$$

Par suite pour tout i de $\{0,...,N\}$

$$x_i = A + ih. \tag{3.40}$$

Sur chaque sous-intervalle $[x_i, x_{i+1}]$ (pour $0 \leq i \leq N-1$) considérons la fonction $p_{i,n}$ qui interpole f en des points $\{x_{i,0},...,x_{i,n}\}$ à préciser. Effectuons une intégration approchée sur chaque intervalle élémentaire $[x_i, x_{i+1}]$ pour $0 \leq i \leq N-1$:

$$\int_A^B f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx,$$
$$= \sum_{i=0}^{N-1} \left(\int_{x_i}^{x_{i+1}} p_{i,n}(x)dx \right) + \sum_{i=0}^{N-1} \left(\int_{x_i}^{x_{i+1}} f\left[x_{i,0}, ..., x_{i,n}, x\right] \left(\prod_{j=0}^{n} (x - x_{i,j}) \right) dx \right).$$

Nous en déduisons

(1) la valeur approchée I_N de $\int_A^B f(x)dx$. Elle est la somme de N valeurs approchées élémentaires

$$I_N = \sum_{i=0}^{N-1} \left(\int_{x_i}^{x_{i+1}} p_{i,n}(x)dx \right).$$

(2) De même, l'erreur d'intégration globale E_N est la somme de N valeurs d'erreurs élémentaires $E_{i,N}$ commises sur chacun des intervalles $[x_i, x_{i+1}]$:

$$\mathcal{E}_N = \sum_{i=0}^{N-1} \mathcal{E}_{i,N} = \sum_{i=0}^{N-1} \left(\int_{x_i}^{x_{i+1}} f\left[x_{i,0}, ..., x_{i,n}, x\right] \left(\prod_{j=0}^{n} (x - x_{i,j}) \right) dx \right).$$

Nous fournissons ci-dessous les résultats relatifs aux quatre formules simples classiques étudiées précédemment ; seule la première preuve est détaillée car les autres en sont très proches.

3.3.2.2. Intégration par la méthode des rectangles.

PROPOSITION 3.17. *Soit f de classe C^1 sur $[A, B]$. Avec les notations* (3.39) *et* (3.40)*, l'intégrale approchée de f sur $[A, B]$ par la méthode composée des rectangles et l'erreur d'intégration sont données par :*

$$I_N^R = h \sum_{i=0}^{N-1} f(x_i), \tag{3.41}$$

$$\mathcal{E}_N^R = h \frac{(B-A)}{2} f'(\eta) \quad \textit{avec } \eta \in [A, B]. \tag{3.42}$$

REMARQUE 3.18. Dans l'écriture de la valeur approchée, nous avons choisi pour chaque intervalle $[x_i, x_{i+1}]$ d'interpoler sur le support $\{x_i\}$, c'est-à-dire une formule des rectangles à gauche . Il était possible de prendre une formule des rectangles à droite ; ceci eût fourni une formule légèrement différente

$$I_N^R = h \sum_{i=1}^{N} f(x_i), \tag{3.43}$$

et la même expression de l'erreur.

DÉMONSTRATION.

(1) La valeur approchée I_N^R a évidemment la forme indiquée vu l'étude menée dans le cas d'intervalles élémentaires et puisque pour tout i on a $x_{i+1} - x_i = h$.

(2) D'après le corollaire 3.7 appliqué à la fonction f sur chacun des intervalles $[x_i, x_{i+1}]$, il vient :

$$\mathcal{E}_N^R = \sum_{i=0}^{N-1} \frac{(x_{i+1} - x_i)^2}{2} f'(\eta_i) = \sum_{i=0}^{N-1} \frac{h^2}{2} f'(\eta_i) \quad \text{avec } \eta_i \in]x_i, x_{i+1}[\text{ pour tout } i ;$$

Mais comme f est de classe C^1 sur $[A, B]$, f' atteint son minimum m et son maximum M sur $[A, B]$; par conséquent, on a pour tout x de $[A, B]$

$$m \leq f'(x) \leq M \quad \text{donc} \quad \frac{h^2}{2} m \leq \frac{h^2}{2} f'(x) \leq \frac{h^2}{2} M ;$$

ainsi on peut écrire les N inégalités suivantes

$$\begin{aligned}
\frac{h^2}{2}m \leq \frac{h^2}{2}f'(\eta_0) \leq \frac{h^2}{2}M,\\
\frac{h^2}{2}m \leq \frac{h^2}{2}f'(\eta_1) \leq \frac{h^2}{2}M,\\
\vdots\\
\frac{h^2}{2}m \leq \frac{h^2}{2}f'(\eta_{N-1}) \leq \frac{h^2}{2}M,
\end{aligned}$$

qui, ajoutées membre à membre, conduisent à

$$m \leq \frac{\mathcal{E}_N^R}{Nh^2/2} \leq M.$$

Ainsi, le théorème des valeurs intermédiaires appliqué à la fonction f', continue sur $[A,B]$, implique qu'il existe η dans $[A,B]$ tel que

$$\frac{\mathcal{E}_N^R}{Nh^2/2} = f'(\eta),$$

c'est-à-dire

$$\mathcal{E}_N^R = N\frac{h^2}{2}f'(\eta) = h\frac{(B-A)}{2}f'(\eta).$$

□

♢

Remarque 3.19. Les preuves conduisant à l'expression des erreurs d'intégration relatives aux autres méthodes étudiées sont similaires et reposent sur le [BM03, Lemme 3.25].

Remarque 3.20. Le [BM03, Lemme 3.25] exprime une idée très générale en analyse numérique : si chacune des erreurs locales$E_{i,N}$ est d'ordre [3] $k+1$globale $\mathcal{E}_N$ est d'ordre k. Cette perte d'un ordre de précision se retrouvera plus bas : voir remarque 3.38 page 70 ou pour les équations différentielles (cf. chapitre 5).

3.3.2.3. Intégration par la méthode des milieux.

Proposition 3.21. *Soit f de classe C^2 sur $[A,B]$. Avec les notations (3.39) et (3.40), l'intégrale approchée de f sur $[A,B]$ par la méthode composée du point milieu et l'erreur d'intégration sont données par :*

$$I_N^M = h\sum_{i=0}^{N-1} f\left(x_i + \frac{h}{2}\right), \tag{3.44}$$

$$\mathcal{E}_N^M = h^2\frac{(B-A)}{24}f''(\eta), \text{ avec } \eta \in [A,B]. \tag{3.45}$$

Démonstration. Analogue à celle développée en méthode des rectangles. □

♢

3.3.2.4. Intégration par la méthode des trapèzes.

Proposition 3.22. *Soit f de classe C^2 sur $[A,B]$. Avec les notations (3.39) et (3.40), l'intégrale approchée de f sur $[A,B]$ par la méthode composée des trapèzes et l'erreur d'intégration sont données par :*

$$I_N^T = \frac{h}{2}\big(f(A)+f(B)\big) + h\sum_{i=1}^{N-1} f(x_i), \tag{3.46}$$

$$\mathcal{E}_N^T = -h^2\frac{(B-A)}{12}f''(\eta) \text{ avec } \eta \in [A,B]. \tag{3.47}$$

Démonstration. De même type que les précédentes. On notera seulement que dans l'expression de I_N^T les points extrêmes A et B interviennent une seule fois au lieu de deux pour les autres x_i (une comme extrémité finale d'intervalle élémentaire, une comme extrémité initiale). □

3. Voir définition 3.35.

♢

3.3.2.5. Intégration par la méthode de Simpson.

PROPOSITION 3.23. *Soit f de classe C^4 sur $[A,B]$. Avec les notations (3.39) et (3.40), l'intégrale approchée de f sur $[A,B]$ par la méthode composée des rectangles et l'erreur d'intégration sont données par :*

$$I_N^S = \frac{h}{6}\left(f(A)+f(B)+2\sum_{i=1}^{N-1} f(x_i)+4\sum_{i=0}^{N-1} f\left(x_i+\frac{h}{2}\right)\right), \tag{3.48}$$

$$\mathcal{E}_N^S = -h^4\frac{(B-A)}{2880}f^{(4)}(\eta) \quad \textit{avec } \eta\in[A,B]\,. \tag{3.49}$$

DÉMONSTRATION. Analogue encore aux précédentes. On remarquera seulement que dans l'expression de I_N^S les points extrêmes A et B interviennent une seule fois, les autres points x_i deux fois encore et les milieux de chaque intervalle $[x_i, x_{i+1}]$ quatre fois en raison de l'étude sur les intervalles élémentaires.

Le lecteur aura remarqué que pour appliquer la méthode de Simpson, f doit être connue en $2N+1$ points, espacés d'un pas $h/2$: les $N+1$ points x_i d'une part et les N milieux des intervalles $[x_i, x_{i+1}]$ d'autre part.

On pourra aussi trouver une preuve alternative de ce résultat dans l'annexe J. □

♢

Les formules d'intégration composites et les erreurs correspondantes sont résumées dans les tableaux 3.4 et 3.5.

méthode	formule
rectangle (gauche)	$h\sum_{i=0}^{N-1} f(x_i)$
rectangle (droite)	$h\sum_{i=1}^{N} f(x_i)$
milieu	$h\sum_{i=0}^{N-1} f(x_i+h/2)$
trapèze	$\frac{h}{2}\big(f(A)+f(B)\big)+h\sum_{i=1}^{N-1} f(x_i)$
Simpson	$\frac{h}{6}\left(f(A)+f(B)+2\sum_{i=1}^{N-1} f(x_i)+4\sum_{i=0}^{N-1} f\left(x_i+\frac{h}{2}\right)\right)$

TABLE 3.4. Méthodes composites (composées) sur $[A,B]$. Dans ce tableau, $N\in\mathbb{N}^*$, $h=(B-A)/N$ et, pour tout $i\in\{0,...,N\}$, $x_i=a+ih$.

REMARQUE 3.24. Comme le montre l'exemple 3.26 page 61, l'intégration composite n'est rien d'autre que l'intégration de l'interpolation composite (voir section 2.5 page 36).

EXEMPLE 3.25. On renvoie à l'annnexe G qui montre comment, dans un cas particulier, on peut approcher l'intégrale d'une fonction sur $[0,+\infty[$ grâce à la méthode de Simpson.

♢

3.3.3. Définition et propriétés des méthodes composites (composées) à pas variable.

Toutes les méthodes composites de la section 3.3.2 sont données à pas variables (les points x_i de la formule (3.39) sont équirépartis). On peut aussi les donner dans le cas où les x_i sont quelconques. Ces formules ne sont pas présentées ici.

méthode	erreur
rectangle	$h\frac{B-A}{2}f'(\eta)$
milieu	$h^2\frac{B-A}{24}f''(\eta)$
trapèze	$-h^2\frac{B-A}{12}f''(\eta)$
Simpson	$-h^4\frac{B-A}{2880}f^{(4)}(\eta)$

TABLE 3.5. Erreurs des méthodes composites (composées) sur $[A, B]$. Dans ce tableau, η appartient à $[A, B]$.

3.3.4. Exemples de simulations numériques sur des Méthodes composites (composées)

EXEMPLE 3.26. On considère A, B et f définis dans l'exemple 2.39 par (2.68) et (2.69). Une primitive de f est donnée par

$$\int f(x)dx = -1/23\,\cos(23\,x) + x + 5/3\,x^3. \tag{3.50}$$

On en déduit la valeur exacte de

$$I = \int_0^1 f(x)dx = \frac{187}{69} - 1/23\,\cos(23) \approx 2.733311580594. \tag{3.51}$$

Reprenons l'exemple 2.39 page 36. Nous calculons quelques approximations de l'intégrale (2.70) selon les quatre méthodes composées vues dans ce chapitre pour différentes valeurs de N, nombre de sous-intervalle décrivant l'ensemble défini par (2.70). Voir les figures 3.6 page suivante à 3.9 page 65. Rappelons à ce propos que l'intégration composite revient à intégrer l'interpolation composite. Les figures 2.12 sont identiques aux figures 3.7 et les figures 2.13 sont identiques aux figures 3.9.

N	rectangles (gauche)	trapèzes	milieux	Simpson
1	1	3,076 889 797 9	1,374 547 825 3	1,941 995 149 5
3	2,375 404 573 4	3,067 701 172 7	2,219 867 631 8	2,502 478 812 1
5	2,147 665 791 6	2,563 043 751 2	2,855 554 821 8	2,758 051 131 6
10	2,501 610 306 7	2,709 299 286 5	2,746 466 563 1	2,734 077 470 9
20	2,624 038 434 9	2,727 882 924 8	2,736 089 093 4	2,733 353 703 9
50	2,690 927 780 6	2,732 465 576 5	2,733 736 146 6	2,733 312 623 3

TABLE 3.6. Différentes valeurs approchées de l'intégrale I donnée par (3.51), fournies par différentes méthodes composites et différentes valeurs de N

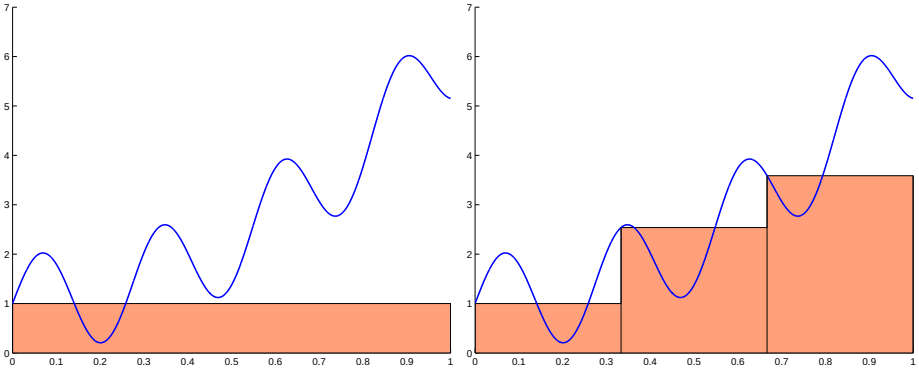
Dans le tableau 3.6, sont données les différentes valeurs obtenues en fonctions de N et des méthodes.

N	rectangles (gauche)	trapèzes	milieux	Simpson
1	1,733 30	$3,435\ 77 \times 10^{-1}$	1,358 76	$7,913\ 16 \times 10^{-1}$
3	$3,579\ 07 \times 10^{-1}$	$3,343\ 90 \times 10^{-1}$	$5,134\ 43 \times 10^{-1}$	$2,308\ 33 \times 10^{-1}$
5	$5,856\ 46 \times 10^{-1}$	$1,702\ 68 \times 10^{-1}$	$1,222\ 42 \times 10^{-1}$	$2,473\ 95 \times 10^{-2}$
10	$2,317\ 00 \times 10^{-1}$	$2,401\ 23 \times 10^{-2}$	$1,315\ 49 \times 10^{-2}$	$7,658\ 90 \times 10^{-4}$
20	$1,092\ 72 \times 10^{-1}$	$5,428\ 65 \times 10^{-3}$	$2,777\ 50 \times 10^{-3}$	$4,212\ 32 \times 10^{-5}$
50	$4,238\ 38 \times 10^{-2}$	$8,460\ 03 \times 10^{-4}$	$4,245\ 66 \times 10^{-4}$	$1,042\ 66 \times 10^{-6}$

TABLE 3.7. Différentes erreurs correspondant aux différentes approximations de l'intégrale I donnée par (3.51), fournies par différentes méthodes composites et différentes valeurs de N

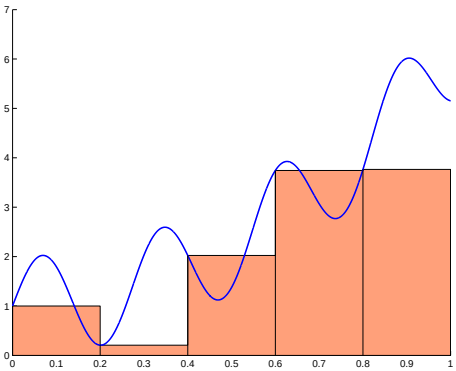
Dans le tableau 3.7, sont données les différentes erreurs correspondantes.

Nous constatons que les erreurs diminuent avec N et que la méthode des rectangles est moins précise que celles des trapèzes et des milieux, du même ordre de précision, elles-mêmes moins précises que celles de Simpson, comme le prévoient les résultats de la section 3.3.2.

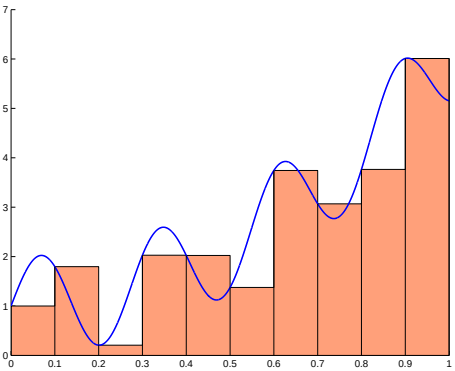


(a) $N = 1$.

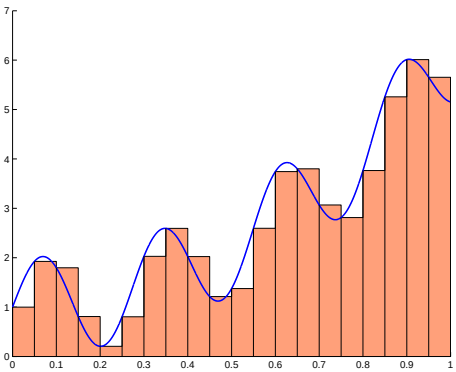
(b) $N = 3$.



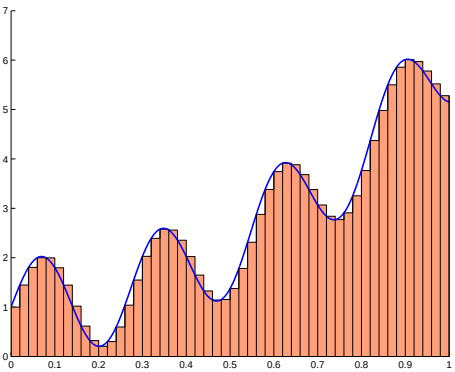
(c) $N = 5$.



(d) $N = 10$.

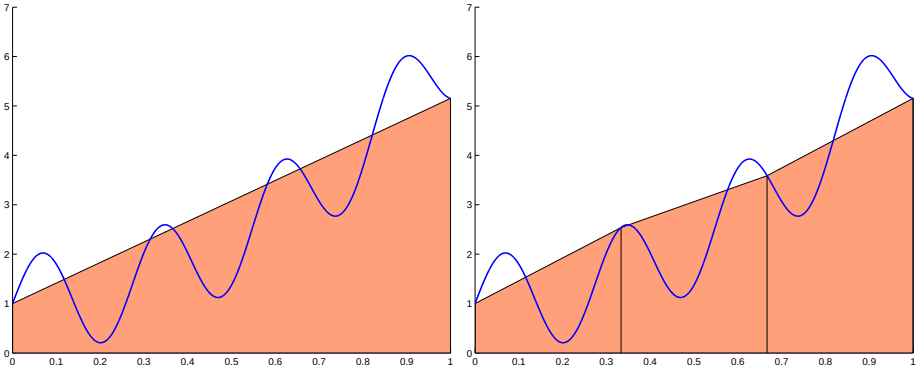


(e) $N = 20$.



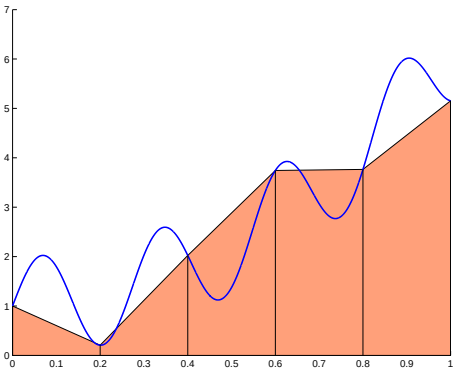
(f) $N = 50$.

FIGURE 3.6. Différentes illustrations de l'approximation de l'intégrale I donnée par (3.51), fournie par la méthode des rectangles (à gauche) avec N sous-intervalles.

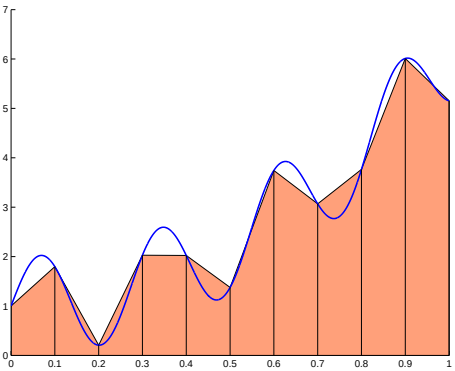


(a) $N = 1$.

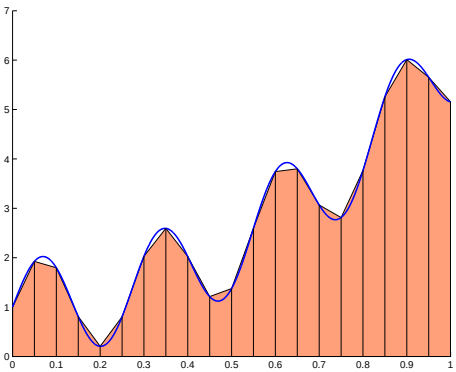
(b) $N = 3$.



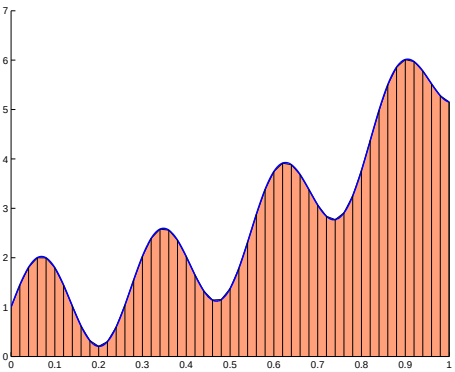
(c) $N = 5$.



(d) $N = 10$.

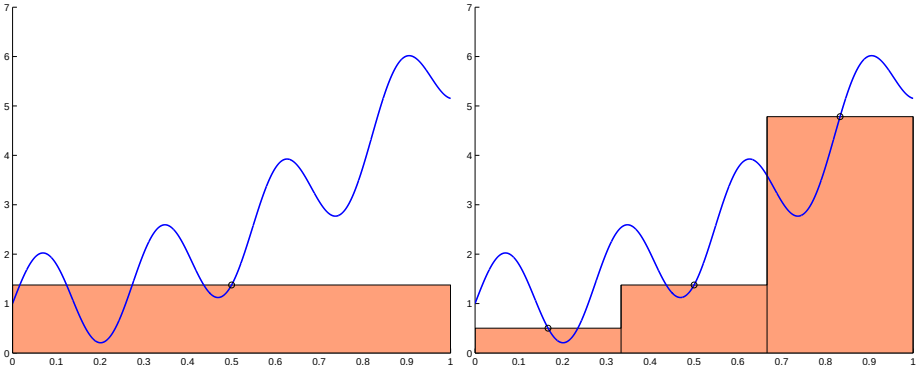


(e) $N = 20$.



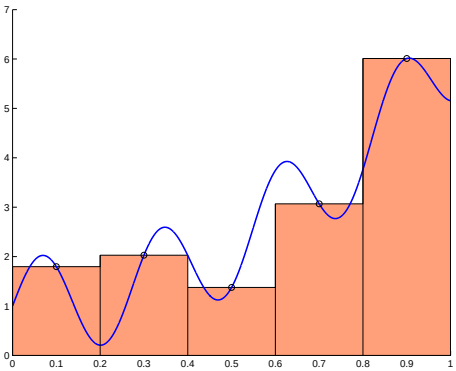
(f) $N = 50$.

FIGURE 3.7. Différentes illustrations de l'approximation de l'intégrale I donnée par (3.51), fournie par la méthode des trapèzes avec N sous-intervalles. Figures à comparer avec les figures 2.12 page 37 !

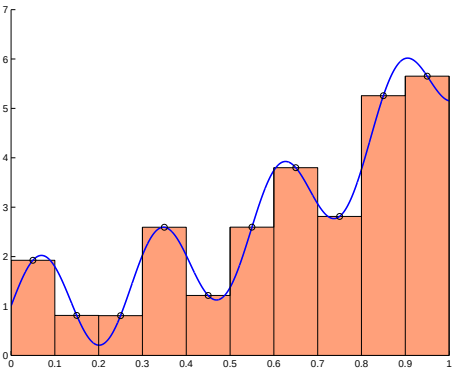


(a) $N = 1$.

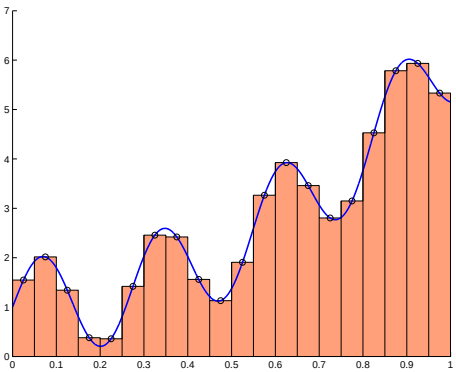
(b) $N = 3$.



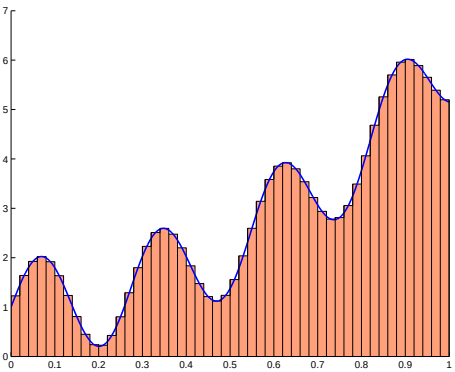
(c) $N = 5$.



(d) $N = 10$.

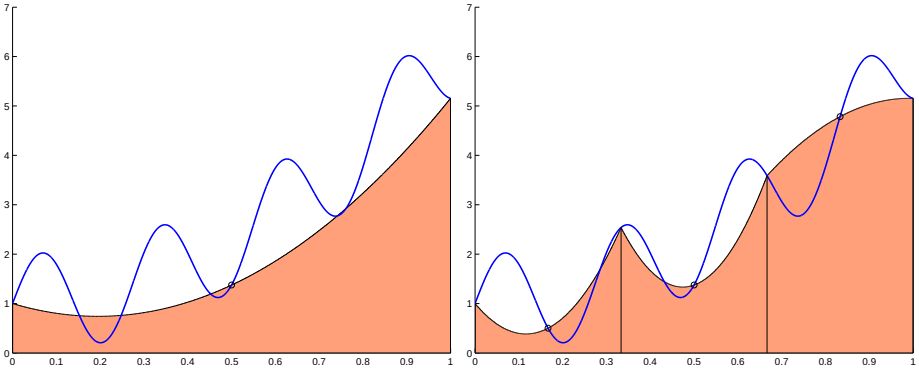


(e) $N = 20$.



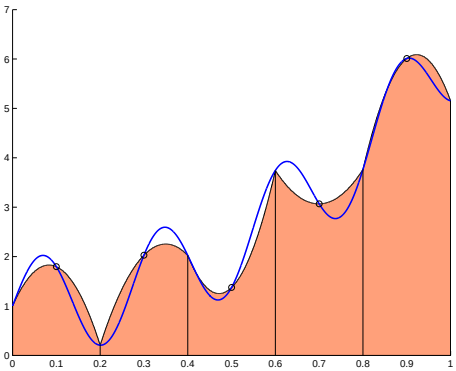
(f) $N = 50$.

FIGURE 3.8. Différentes illustrations de l'approximation de l'intégrale I donnée par (3.51), fournie par la méthode des milieux avec N sous-intervalles.

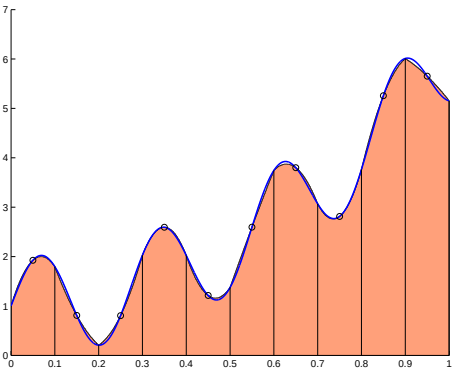


(a) $N = 1$.

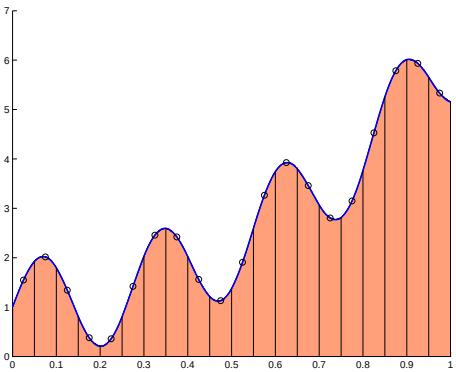
(b) $N = 3$.



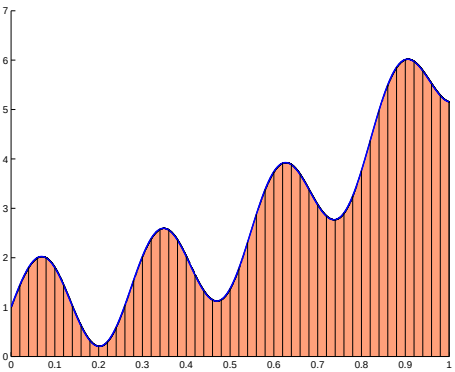
(c) $N = 5$.



(d) $N = 10$.



(e) $N = 20$.



(f) $N = 50$.

FIGURE 3.9. Différentes illustrations de l'approximation de l'intégrale I donnée par (3.51), fournie par la méthode de Simpson avec N sous-intervalles. Figures à comparer avec les figures 2.13 page 38 !

EXEMPLE 3.27.

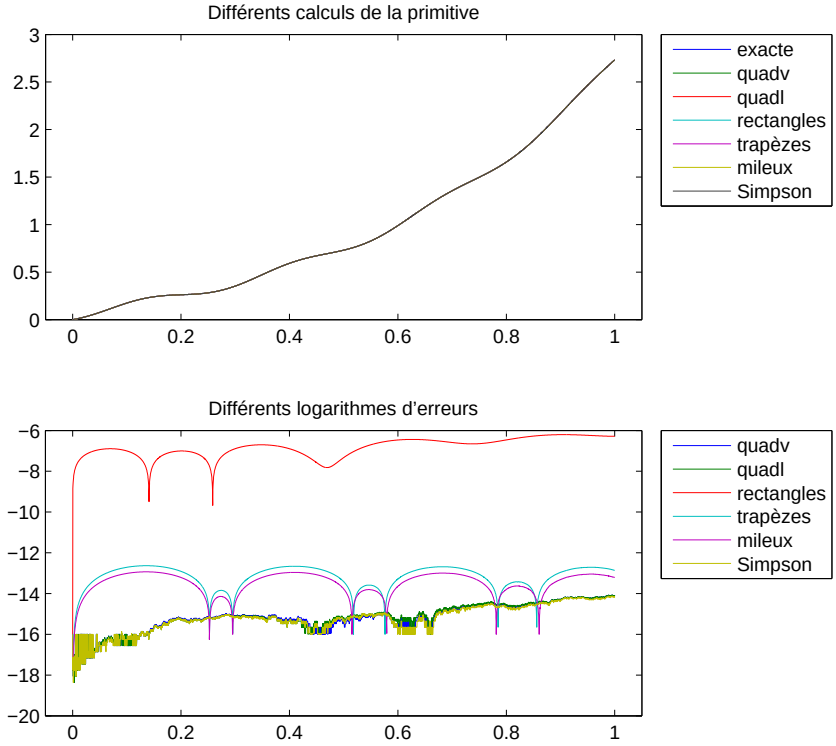


FIGURE 3.10. Différents calculs de la primitive de la fonction de l'exemple 3.26.

Reprenons l'exemple 3.26 page 61. Traçons maintenant la primitive de f sur $[A, B]$ en prenant 4000 points de calcul noté u_i et sur chacun des intervalles $[u_i, u_{i+1}]$, 1000 sous-intervalles. On utilise les fonctions `quadl` et `quadv` de matlab, puis par les méthodes des rectangles, des trapèzes, des milieux et de Simpson. Voir la figure 3.10. On y constate que la primitive calculée est visuellement identique pour les six méthodes. Comme prévu, la méthode la moins précise est celle des rectangles, puis celles des milieux et des trapèzes, du même ordre et enfin, les plus précises celles de Simpson et celles utilisant les fonctions `quadv` de `quadl` de matlab.

`quadv`	`quadl`	rectangles	trapèzes	milieux	Simpson
0.07982	0.41554	0.17777	0.22372	0.16589	0.37050

TABLE 3.8. Temps de calcul des différentes méthodes de calcul de la fonction définie dans l'exemple 3.26 page 61

Les temps de calculs sont donnés dans le tableau 3.8.

EXEMPLE 3.28. Reprenons l'exemple 3.26 page 61. On trace le logarithme (décimal, en base 10) de l'erreur entre les approximations par différentes méthodes (des rectangles, des trapèzes, des milieux et de Simpson) en fonction du logarithme du pas $h = (B - A)/N$ où N est le nombre de sous-intervalles. D'après les résultats du tableau 3.5, pour chacune des ces quatre méthodes, on a

$$|E| = C \left| f^{(q)}(\eta) \right| h^p,$$

où C dépend de A et de B. Si on admet que η varie peu en fonction de h et en écrivant abusivement, on a

$$\log(|E|) \approx p \log(h) + D,$$

où D ne dépend pas de h. Ainsi, la pente du nuage de points constitué par $(\log(h), (\log(|E|)))$ correspond à la valeur de p.

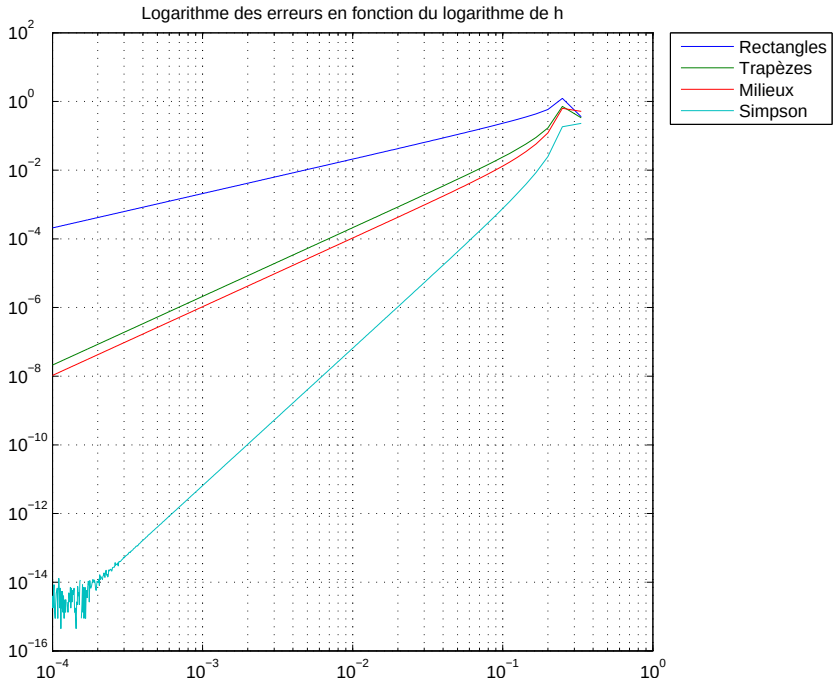


FIGURE 3.11. Le nuage de points $(\log(h), (\log(|E|)))$ pour f données dans l'exemple 3.26.

	exposant théorique	pente expérimentale
Rectangles	1	1.0072318
Trapèzes	2	2.0120085
Milieux	2	2.0207604
Simpson	4	3.9765418

TABLE 3.9. Pentes théoriques et calculées pour f donnée dans l'exemple 3.26

On observe alors pour N variant (de façon telle que les valeurs des log de N soit équirépartis) de 3 à 10000, en prenant 666 valeurs, la figure 3.11 et les pentes données par le tableau 3.9.

EXEMPLE 3.29.

Si maintenant, sur le même intervalle $[A, B]$ que dans l'exemple 3.28, on considère la fonction f définie par

$$f(x) = \sqrt{x}, \tag{3.52}$$

on obtient alors la figure 3.12 et les pentes données par le tableau 3.10.

Cela s'explique que f n'est plus dérivable en zéro et que les résultats du tableau 3.5 ne sont donc plus valables.

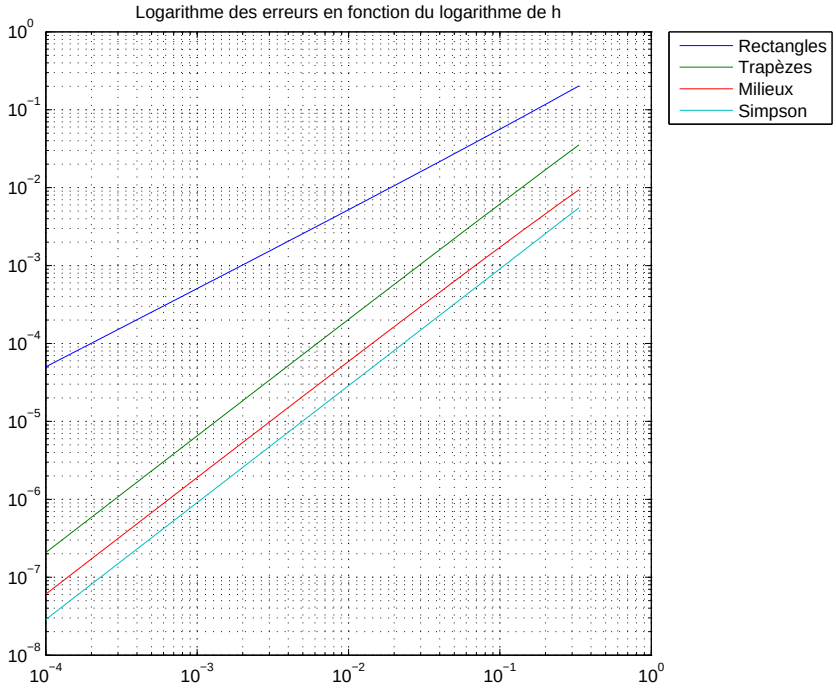


FIGURE 3.12. Le nuage de points $(\log(h), (\log(|E|)))$ pour f données par (3.52).

	exposant théorique	pente expérimentale
Rectangles	1	1.0119003
Trapèzes	2	1.4935411
Milieux	2	1.4887678
Simpson	4	1.4999982

TABLE 3.10. Pentes théoriques et calculées pour f donnée par (3.52)

EXEMPLE 3.30. On pourra de nouveau consulter l'exemple de la section 3.1 page 46.

3.4. Formules de quadrature

DÉFINITION 3.31. Soit $n \in \mathbb{N}$. Étant donnés $n+1$ points deux à deux distincts dans $[a,b]$, on appelle formule de quadrature toute formule de la forme

$$Q(f) = \sum_{i=0}^{n} W_i f(x_i), \tag{3.53}$$

permettant d'approcher

$$\mathcal{I}(f) = \int_a^b f(x)dx. \tag{3.54}$$

Les points $(x_i)_{0\leq i\leq n}$ sont appelés les nœuds ou les points et les coefficients réels $(W_i)_{0\leq i\leq n}$ les poids de la formule de quadrature.

Toutes les formules vues plus haut sont des formules de quadrature. Voir par exemple les formules des tableaux 3.2 (méthode de quadrature sur $[a, b]$) ou 3.4 (méthode de quadrature sur $[A, B]$) ou (3.38). Voir la remarque 3.16 page 55.

Définition 3.32. Une formule de quadrature $Q(f)$ est exacte pour une fonction f si $Q(f) = \mathcal{I}(f)$. Elle est dite exacte de degré r

- si elle est exacte pour tout polynôme p de degré r au plus, c'est-à-dire :

$$\forall p \in P_r, \quad Q(p) = \mathcal{I}(p), \tag{3.55}$$

  où P_r est l'ensemble des polynômes de degrés au plus p ;
- et si elle n'est pas exacte pour un moins un polynôme de degré $r + 1$.

r est le degré d'exactitude de la formule de quadrature.

Attention, dans certains corrigés, sont confondues abusivement les notions de degré et d'ordre, défini plus bas de façon légèrement différente ! Il suffira de prêter garde au contexte. ♢

Lemme 3.33. *Une formule de quadrature $Q(f)$ de degré d'exactitude r ssi on a*

$$\forall i \in \{0, ..., r\}, \quad Q\left(x \mapsto x^i\right) = \mathcal{I}\left(x \mapsto x^i\right), \tag{3.56a}$$

et

$$Q\left(x \mapsto x^{r+1}\right) \neq \mathcal{I}\left(x \mapsto x^{r+1}\right). \tag{3.56b}$$

Les égalités (3.56a) *forment un système linéaire faisant intervenir la matrice de Vandermonde (voir section 2.2.2.1).*

Démonstration. Résultats admis. Voir exercices de TD 3.3 et 3.2. □

♢

Voir de nouveau la remarque 3.15 page 55.

Lemme 3.34. *Soit r un entier supérieur à n. La formule de quadrature $Q(f)$ est de degré d'exactitude au moins $r \in \mathbb{N}$ ssi les poids $(W_i)_{0 \leq i \leq n}$ sont donnés par la formule*

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_a^b l_i(x)dx, \tag{3.57}$$

où l_i sont les polynômes de Lagrange relatifs au support $\{x_0, ..., x_n\}$.

Démonstration.

- Si la formule de quadrature $Q(f)$ est de degré d'exactitude au moins $r \geq n$, alors elle est exacte pour l_j, pour j fixé, le polynôme de Lagrange de f sur le support $\{x_0, ..., x_n\}$ puisqu'il est de degré n. On a donc

$$\mathcal{I}(l_j) = Q(l_j),$$

  et donc

$$\int_a^b l_j(x)dx = \sum_{i=0}^{n} W_i l_j(x_i).$$

  On utilise le fait que

$$\forall j \in \{0, ..., n\}, \quad l_i(x_j) = \delta_{ij},$$

  et on obtient donc

$$\int_a^b l_j(x)dx = \sum_{i=0}^{n} W_i \delta_{ij},$$

  et donc

$$\int_a^b l_j(x)dx = W_j.$$

- Réciproquement, soit P un polynôme de degré au plus n. Montrons que $\mathcal{I}(P) = Q(P)$. On a successivement

$$\begin{aligned} Q(p) &= \sum_{i=0}^{n} W_i P(x_i), \\ &= \sum_{i=0}^{n} \left(\int_a^b l_i(x)dx \right) P(x_i), \\ &= \int_a^b \left(\sum_{i=0}^{n} P(x_i) l_i(x) \right) dx. \end{aligned}$$

D'après l'équation

$$\Pi_n(P)(x) = \sum_{i=0}^{n} P(x_i) l_i(x),$$

on a donc

$$= \int_a^b \Pi_n(P)(x)dx.$$

Puisque P est un polynôme de degré au plus n, on a donc $\Pi_n(P) = P$, ce qui achève la preuve.

□

♢

Définition 3.35. Soit $q \in \mathbb{N}^*$. On dit que la formule de quadrature $Q(f)$ est d'ordre q si, pour toute fonction assez régulière f, il existe une constante M telle que

$$|Q(f) - \mathcal{I}(f)| \leq M(b-a)^q. \tag{3.58}$$

Exemple 3.36. D'après le tableau 3.3, on constate que les méthodes élémentaires du rectangle, du milieu, du trapèze et de Simpson, sont d'ordre respectifs 2, 3, 3 et 5, si f est respectivement de classe $\mathcal{C}^1$, $\mathcal{C}^2$, $\mathcal{C}^2$ et $\mathcal{C}^4$. La notion d'ordre est liée à celle de degré (remarque 3.15 page 55) mais pas toujours à celle du nombre de points. Par exemple, pour 1 point d'intégration, la méthode du rectangle est de degré 0 et d'ordre 2, tandis que pour le même nombre de point la méthode du milieu est de degré 1 et d'ordre 3.

Remarque 3.37. On pourra aussi consulter l'annexe J.

Remarque 3.38. Comme dans la remarque 3.20 page 58, on peut aussi parler d'ordre pour les méthodes composites, mais, dans ce cas, (3.58) est remplacé par

$$|Q(f) - \mathcal{I}(f)| \leq Mh^q, \tag{3.59}$$

où ici $h = (B-A)/N$. Ainsi, comme annoncé dans la remarque 3.20 et comme le montrent les tableaux 3.3 page 56 et 3.5, les méthodes élémentaires du rectangle, du milieu, du trapèze et de Simpson, sont d'ordre respectifs 2, 3, 3 et 5, tandis que leurs homologues composites sont d'ordre un de moins. En revanche, les degrés des méthodes de quadrature élémentaires sont identiques à leurs homologues composées.

Exemple 3.39. Consulter de nouveau l'exemple 3.28 page 66 qui montre que les ordres des méthodes composées peuvent se mesurer de façon graphique.

3.4.1. Formules de Newton-Cotes

En prenant n points quelconque équirépartis, on peut engendrer des formules de quadrature avec la formule (3.38), dites de Newton-Cotes. Elles sont toutes de degré au moins $n = 1$. Si n est pair, elle sont de degré n. On gagne donc un degré pour les n paires ! La méthode du milieu et de Simpson, qui sont des méthodes de newton-Cotes particulières, seront donc à privilégier par rapport à celle des rectangles et des trapèzes. En pratique, la méthode d'ordre plus élevé, celle de Simpson, sera à privilégier. En revanche, si n augmente, les poids W_i deviennent grands et de signes mélangés, ce qui rend ces formules sensibles aux erreurs d'arrondis.

Voir [CM84, p. 37, 38] et annexe K. ♢

3.4.2. Formules de Gauss-Legendre

$n=0$	$x_0=0$		
$n=1$	$x_0=-\frac{\sqrt{3}}{3}$	$x_1=\frac{\sqrt{3}}{3}$	
$n=2$	$x_0=-\frac{\sqrt{15}}{5}$	$x_1=0$	$x_2=\frac{\sqrt{15}}{5}$

TABLE 3.11. Expression des premiers points x_i pour la méthode Gauss-Legendre.

$n=0$	$W_0=2$		
$n=1$	$W_0=1$	$W_1=1$	
$n=2$	$W_0=\frac{5}{9}$	$W_1=\frac{8}{9}$	$W_2=\frac{5}{9}$

TABLE 3.12. Expression des premiers poids W_i pour la méthode Gauss-Legendre.

Plutôt que de chercher des nœuds équirépartis, on considère, pour $n \in \mathbb{N}$ fixé, $n+1$ points $(x_i)_{0\leq i\leq n}$ dans l'intervalle $[A,B]$ et $n+1$ poids $(W_i)_{0\leq i\leq n}$. On les détermine de façon à obtenir le degré le plus élevé possible. On peut montrer que le degré optimal est $2n+1$ et peut être obtenu en choisissant les points et des poids de la méthode dite de Gauss-Legendre, qui sont donnés dans les tableaux 3.11 et 3.12. Cette méthode est traditionnellement étudiée sur $[-1,1]$. À partir de cette intervalle, par changement de variable, on se ramène à tout intervalle. La méthode du milieu est une méthode de Gauss et c'est la seule, parmi les méthodes de Newton-Cotes à bénéficier de cette double qualification !

Voir aussi les exercices de TD 3.4 et 3.5

Cette méthode peut aussi être utilisée pour des intervalles non bornés et fournit des méthodes très précises. Voir [CM84, p. 45] et [BM03, Section 3.3]. ♢

3.5. Un exercice type à savoir traiter parfaitement

Énoncé

Soit f donnée par

$$\forall x \in [0,2]\,, \quad f(x)=\sin(x)\,, \tag{3.60a}$$

et l'intégrale I

$$I=\int_0^2 f(x)dx. \tag{3.60b}$$

(1) (a) Déterminer I^T, l'approximation de I par la méthode élémentaire du trapèze.

(b) Déterminer f' et f'', puis donnez l'expression de l'erreur commise avec la méthode élémentaire du trapèze et fournissez-en une majoration.

(c) (i) Calculer la valeur exacte de I.

(ii) En déduire l'erreur commise réelle, c'est-à-dire $\left|I^T-I\right|$ et vérifier qu'elle est inférieure au majorant de l'erreur donné plus haut.

(2) (a) Déterminer I_3^T, l'approximation de I par la méthode composite des trapèzes avec $N = 3$ sous-intervalles.

(b) Donnez l'expression de l'erreur commise avec la méthode composite des trapèzes puis fournissez-en une majoration.

(c) Déterminer l'erreur réelle erreur commise, c'est-à-dire $\left|I_3^T - I\right|$ et vérifier qu'elle est inférieure au majorant de l'erreur donné plus haut.

(3) Déterminer le nombre N de sous-intervalles qu'il faudrait utiliser pour avoir une approximation de I par la méthode composite des trapèzes avec une erreur inférieure à

$$\varepsilon = 10^{-8}. \tag{3.61}$$

Corrigé

(1) (a) En utilisant le tableau 3.2, on détermine la valeur approchée avec la méthode élémentaire du trapèze :

$$I^T = \sin(2) \tag{3.62}$$

soit

$$I^T = 0.90929742682568. \tag{3.63}$$

(b) On obtient les dérivées successives de f :

$$f'(x) = \cos(x)\ ; \tag{3.64a}$$

$$f''(x) = -\sin(x)\,. \tag{3.64b}$$

On majore $|\sin(x)|$ par 1. On en déduit

$$M_2 = \max_{x\in[0,2]} \left|f^{(2)}(x)\right|, \tag{3.65}$$

le maximum de la valeur absolue de la dérivée 2-ième de f sur l'intervalle d'étude, donné numériquement par

$$M_2 = 1. \tag{3.66}$$

On note

$$a = 0, \quad b = 2. \tag{3.67}$$

Le tableau 3.3 fournit l'expression de l'erreur commise avec la méthode élémentaire du trapèze :

$$\mathcal{E}^T = -\frac{(b-a)^3}{12} f''(\eta), \tag{3.68}$$

où η appartient à $]a, b[$. On vérifie que f est bien de classe $\mathcal{C}^2$. On majore la valeur absolue de $f''(\eta)$, par le maximum de la valeur absolue de la dérivée correspondant et la majoration de l'erreur commise est donc donnée par

$$\mathcal{E}^T \leq \frac{(b-a)^3}{12} M_2 \tag{3.69}$$

Grâce à (3.67) et (3.66), on déduit donc la majoration de l'erreur commise suivante :

$$\mathcal{E}^T \leq 0.6666666667. \tag{3.70}$$

(c) (i) On obtient

$$I = 1 - \cos(2)\,, \tag{3.71a}$$

soit encore

$$I = 1.4161468365471. \tag{3.71b}$$

(ii) L'erreur réelle commise est égale à

$$\left|I^T - I\right| = |1.4161468365471 - 0.9092974268257| = 0.5068494097215$$

qui est inférieure à celle donnée par (3.70).

(2) (a) En utilisant le tableau 3.4, on détermine la valeur approchée avec la méthode composite des trapèzes avec $N = 3$:

$$I_3^T = 1/3 \sin(2) + 2/3 \sin(2/3) + 2/3 \sin(4/3) \tag{3.72}$$

soit

$$I_3^T = 1.36330427856393. \tag{3.73}$$

(b) On note maintenant

$$A = 0, \quad B = 2. \tag{3.74}$$

Le tableau 3.5 fournit l'expression de l'erreur commise avec la méthode composite des trapèzes :

$$\mathcal{E}_3^T = -h^2 \frac{B-A}{12} f''(\eta), \tag{3.75}$$

où η appartient à $[A, B]$ et

$$h = \frac{B-A}{N}, \tag{3.76}$$

soit

$$h = \frac{(2)-(0)}{3},$$

et donc

$$h = 0.6666666666667. \tag{3.77}$$

On peut donc écrire

$$\left|\mathcal{E}_3^T\right| \leq h^2 \frac{B-A}{12} M_2. \tag{3.78}$$

En utilisant de nouveau (3.66), on déduit donc la majoration de l'erreur commise suivante :

$$\mathcal{E}_3^T \leq 7.407407\,10^{-2}. \tag{3.79}$$

(c) L'erreur réelle commise est égale à

$$\left|I_3^T - I\right| = |1.4161468365471 - 1.3633042785639| = 5.284256\,10^{-2}$$

qui est inférieure à celle donnée par (3.79).

(3) Pour que

$$\left|\mathcal{E}_3^T\right| \leq \varepsilon,$$

il suffit, d'après (3.78) que l'on ait :

$$h^2 \frac{B-A}{12} M_2 \leq \varepsilon,$$

soit, d'après (3.76),

$$\left(\frac{B-A}{N}\right)^2 \frac{B-A}{12} M_2 \leq \varepsilon,$$

soit encore

$$\frac{(B-A)^3}{12\varepsilon} M_2 \leq N^2,$$

et donc

$$N \geq \sqrt{\frac{M_2 (B-A)^3}{12\varepsilon}}.$$

Il suffit donc de prendre

$$N = \left\lceil \sqrt{\frac{M_2 (B-A)^3}{12\varepsilon}} \right\rceil . \tag{3.80}$$

où pour tout réel X,

$\lceil X \rceil$ est le plus petit entier supérieur ou égal à X.

Numériquement, on a donc en utilisant de nouveau (3.66),

$$N = 8165. \tag{3.81}$$

Remarque 3.40. Avec cette valeur de N, on a

$$\mathcal{E}_{8165}^{T} = 1.416146829466465,$$

et l'erreur réelle

$$\left| \mathcal{E}_{8165}^{T} - I \right| = 7.0806769\,10^{-9},$$

quantité qui est inférieure à ε donné par l'équation (3.61) de l'énoncé.

Chapitre 4

Équations non-linéaires

4.1. Motivation

On veut calculer le taux de rente moyen i d'un fonds de placement sur plusieurs années. On a investi dans le fonds $v = 500$ euros chaque année et on se retrouve après 5 ans avec un montant de $p = 3000$ euros. On sait que la relation qui lie p, v et i et le nombre d'années n est

$$p = v \sum_{k=1}^{n} (1+i)^k = v \frac{1+i}{i} \left((1+i)^n - 1 \right) \tag{4.1}$$

On pose

$$\Psi(i) = p - v \frac{1+i}{i} \left((1+i)^n - 1 \right). \tag{4.2}$$

Ce problème est donc ramené à trouver $i \in \mathbb{R}_+^*$ tel que

$$\Psi(i) = 0, \tag{4.3}$$

équation non linéaire, dont on n'est pas capable de trouver une solution exacte.

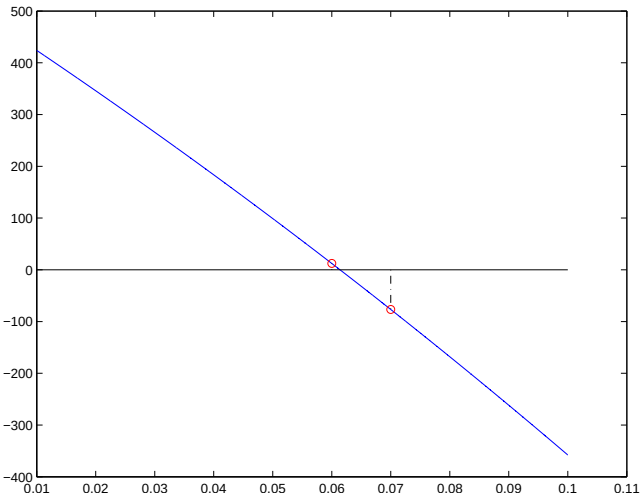


FIGURE 4.1. Le graphique de la fonction Ψ sur $[0.010, 0.100]$.

On constate graphiquement que Ψ semble posséder une racine, notée α, dans l'intervalle $[0.060000, 0.070000]$. Les résultats de plusieurs méthodes sont présentés ici :

(1)

La méthode de dichotomie (voir section 4.3). On définit une suite d'intervalles $[a_n, b_n]$ dont le milieu x_n tend vers α. Voir le tableau 4.1. Pour une précision

$$\varepsilon = 1.\,10^{-12}, \tag{4.4}$$

n	x_n	a_n	b_n
0	0.0550000000000	0.0100000000000	0.1000000000000
1	0.0775000000000	0.0550000000000	0.1000000000000
2	0.0662500000000	0.0550000000000	0.0775000000000
3	0.0606250000000	0.0550000000000	0.0662500000000
4	0.0634375000000	0.0606250000000	0.0662500000000
5	0.0620312500000	0.0606250000000	0.0634375000000
6	0.0613281250000	0.0606250000000	0.0620312500000
7	0.0616796875000	0.0613281250000	0.0620312500000
29	0.0614024115447	0.0614024114609	0.0614024116285
30	0.0614024115028	0.0614024114609	0.0614024115447
31	0.0614024115237	0.0614024115028	0.0614024115447
32	0.0614024115342	0.0614024115237	0.0614024115447
33	0.0614024115395	0.0614024115342	0.0614024115447
34	0.0614024115368	0.0614024115342	0.0614024115395
35	0.0614024115355	0.0614024115342	0.0614024115368
36	0.0614024115362	0.0614024115355	0.0614024115368

TABLE 4.1. Valeurs des extrémités a_n et b_n des intervalles et des milieux x_n

on obtient, en 36 itérations,

$$x_n \approx 0.0614024115362, \tag{4.5}$$

en lequel

$$\Psi(x_n) \approx 3.01951\,10^{-9}. \tag{4.6}$$

(2)

La méthode de point fixe (voir section 4.4). On pose

$$\Phi(i) = \frac{\Psi(i)}{K} + i; \tag{4.7}$$

avec

$$K = 1.\,10^4. \tag{4.8}$$

Il est clair que (4.3) est équivalent à

$$\Phi(i) = i. \tag{4.9}$$

On définit une suite x_n par

$$\forall n \in \mathbb{N}, \quad x_{n+1} = \Phi(x_n), \tag{4.10}$$

avec

$$x_0 = 0.0600000000000. \tag{4.11}$$

Voir le tableau 4.2 et la figure 4.2. Pour une précision donnée par (4.4) on obtient, en 11 itérations,

$$x_n \approx 0.0614024115364, \tag{4.12}$$

en lequel

$$\Phi(x_n) - x_n \approx 6.14024\,10^{-6}, \tag{4.13}$$

et donc des valeurs proches de (4.5) et (4.6).

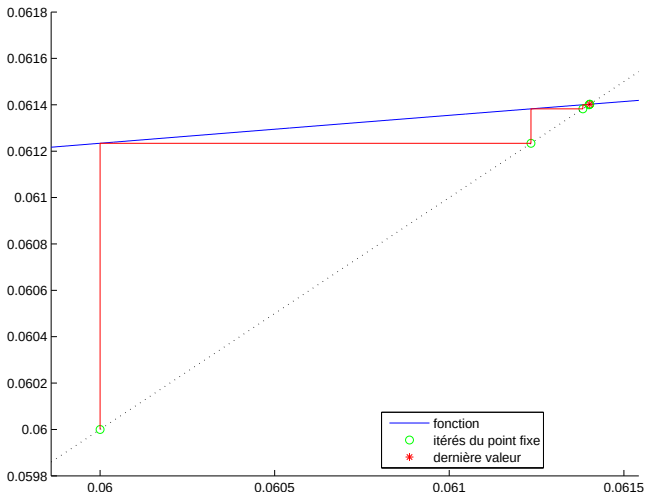


FIGURE 4.2. Le graphique de la fonction Φ sur $[0.010, 0.100]$ et les premiers termes de la suite des itérés x_n.

n	x_n
0	0.0600000000000
1	0.0612340731200
2	0.0613824426482
3	0.0614000461496
4	0.0614021313956
5	0.0614023783591
6	0.0614024076073
7	0.0614024110712
8	0.0614024114814
9	0.0614024115300
10	0.0614024115358
11	0.0614024115364

TABLE 4.2. Valeurs des itérés du point fixe x_n

(3)

La méthode de Newton (voir section 4.5). On construit une autre suite x_n.

Voir le tableau 4.3 et la figure 4.3. Pour une précision donnée par (4.4) on obtient, en 4 itérations (beaucoup moins que précédemment donc),

$$x_n \approx 0.0614024115365, \tag{4.14}$$

en lequel

$$\Psi(x_n) \approx 5.00222\,10^{-12}, \tag{4.15}$$

et donc des valeurs proches de (4.5) et (4.6).

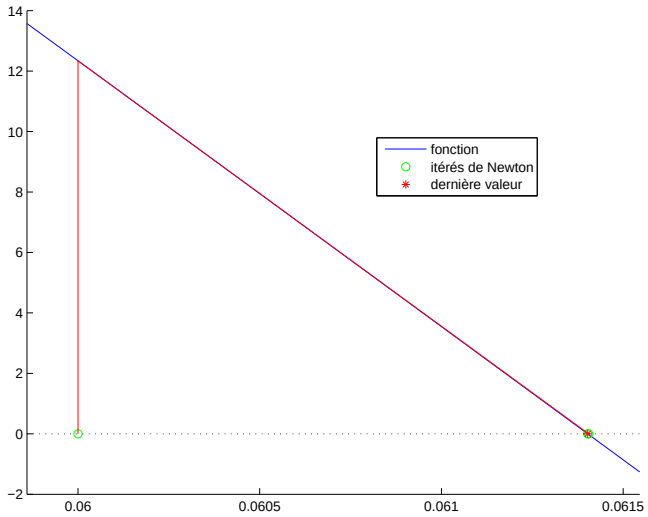


FIGURE 4.3. Le graphique de la fonction Ψ sur $[0.010, 0.100]$ et les premiers termes de la suite des itérés x_n de la méthode de Newton.

n	x_n
0	0.0600000000000
1	0.0614049702763
2	0.0614024115450
3	0.0614024115365
4	0.0614024115365

TABLE 4.3. Valeurs des itérés de la méthode de Newton x_n

4.2. Généralités

On se donne une fonction f, une fonction de $\mathbb{R}$ dans $\mathbb{R}$ dont on cherche un zéro ou une racine, noté r, c'est-à-dire vérifiant

$$f(r) = 0. \tag{4.16}$$

On fera l'hypothèse minimale que f est continue, hypothèse qui pourra être renforcée. Nous allons construire, par différentes méthodes, dites itératives, une suite $(x_n)_{n\in\mathbb{N}}$ convergeant vers r.

4.3. Méthode de bissection ou dichotomie

Cette méthode est la plus simple, aussi bien sur le plan théorique, que sur le plan pratique (mise en œuvre informatique). Elle garantit aussi que l'on reste dans l'intervalle de départ.

On a le lemme immédiat suivant, qui est une conséquence directe du théorème des valeurs intermédiaires.

LEMME 4.1. *Soit f continue sur un intervalle $[a, b]$ telle que*

$$f(a)f(b) \leq 0. \tag{4.17}$$

Alors, il existe un zéro de f sur $[a, b]$.

REMARQUE 4.2. Attention, f peut posséder un zéro sur un intervalle sans changer de signe, comme par exemple $f(x) = x^2$ sur $[-1, 1]$. f peut posséder aussi plusieurs zéro sur $[a, b]$. Les méthodes itératives fourniront une suite qui convergera vers l'un des zéros de f. Si $f(a)f(b) = 0$, alors a ou b est un zéro de f et on s'arrête là. On remplacera donc souvent en pratique (4.17) par

$$f(a)f(b) < 0, \tag{4.18}$$

ce qui assurera l'existence un zéro de f sur $]a, b[$.

On renvoie à l'annexe L. On en déduit le lemme suivant

LEMME 4.3. *Soit f vérifiant les hypothèses du lemme 4.1. La méthode de dichotomie est convergente. Autrement dit, la suite (x_n) construite converge vers un des zéros de f.*

L'algorithme 4.1 permet de déterminer les premières valeurs des suites (a_n), (b_n) et (x_n), en reprenant les résultat de l'annexe L, qui implique que cet algorithme est bien écrit, c'est-à-dire, qu'à chaque étape, a_n et b_n sont constructibles. On peut considérer que les deux suite a_n et b_n sont des approximations d'un zéro recherché (comme dans dans la section 1.2 page 2) ou alors que c'est la suite x_n qui en est une approximation. Cet algorithme sera affiné en pratique, comme dans la fonction `http://utbmjb.chez-alice.fr/Polytech/MNBif/fichiers_matlab/bisection.m`.

REMARQUE 4.4. Cet algorithme de dichotomie n'a pas besoin en fait de l'existence d'un zéro de f sur $[a, b]$. Il permet en fait d'en montrer l'existence et donc de montrer le lemme 4.1 !

On a le résultat suivant :

PROPOSITION 4.5. *Avec l'algorithme de dichotomie (sous les hypothèses du lemme 4.1), si on cherche la racine r avec une erreur absolue ε (précision recherchée), il faut effectuer n itérations telles que*

$$\frac{b-a}{2^n} \leq \varepsilon, \tag{4.19}$$

soit encore

$$n \geq \frac{\ln\left((b-a)/\varepsilon\right)}{\ln 2}, \tag{4.20}$$

soit encore,

$$n = \left\lceil \frac{\ln\left((b-a)/\varepsilon\right)}{\ln 2} \right\rceil, \tag{4.21}$$

Algorithme 4.1 Algorithme de dichotomie *dichotomie*(a, b, ε, $f \longrightarrow n$, x)

Entrée :

a, b réels tels que $a < b$.

ε, réel strictement positif.

f une fonction continue sur $[a, b]$ telle que $f(a)f(b) \leq 0$.

Sortie :

n, entier positif ou nul.

x tel que $|\alpha - x| \leq \varepsilon$ ou $x = \alpha$ où α est un zéro de f sur $[a, b]$ ($x = x_n$, après n itérations de l'algorithme).

$n \longleftarrow 0$
$x \longleftarrow (a + b)/2$
tant que $b - a > \varepsilon$ et $f(x) \neq 0$ **faire**
  $n \longleftarrow n + 1$
  $x \longleftarrow (a + b)/2$
  si $f(a)f(x) \geq 0$ **alors**
    $a \longleftarrow x$
  sinon
    $b \longleftarrow x$
  fin si
fin tant que

où[1]*, pour tout réel* X*,*

$$\lceil X \rceil \text{ est le plus petit entier supérieur ou égal à } X. \tag{4.22}$$

Démonstration. On sait que $x_n \in [a_n, b_n]$, pour tout n. À chaque itération, la longueur de l'intervalle $[a_n, b_n]$ est divisée par 2. À la n-ième itération, sa longueur est donc égale à $(b - a)/2^n$. D'après l'annexe L, l'intervalle $[a_n, b_n]$ contient à la fois un zéro de f, noté r et x_n. On a donc

$$|x_n - r| \leq \frac{b - a}{2^n},$$

et si $(b-a)/2^n \leq \varepsilon$, on a par transitivité $|x_n - r| \leq \varepsilon$. On conclue en prenant le logarithme de (4.19) qui fournit (4.20) et (4.21). □

Exemple 4.6. On pourra de nouveau consulter l'exemple de la section 4.1, point 1 page 75.

Remarque 4.7. Une conséquence rarement écrite du principe de dichotomie est qu'il fournit l'écriture en base de 2 de $(r - a)/(b - a)$. Voir la section M.4 page 202.

♢

4.4. Méthode de point fixe

4.4.1. Définition

On peut transformer la recherche d'un zéro de f vérifiant (4.16) en la recherche d'un nombre r vérifiant

$$g(r) = r, \tag{4.23}$$

où g peut être déterminée en fonction de f. L'équation (4.23) est appelée équation de point fixe de g.

On associe alors à cette équation, la suite du point fixe définie par

1. En anglais, on parle de "plafond" : ceil.

Définition 4.8 (Méthode du point fixe). La suite $(x_n)_{n\in\mathbb{N}}$ est définie par la donnée de $x_0 \in \mathbb{R}$ et

$$\forall n \in \mathbb{N}, \quad x_{n+1} = g(x_n). \tag{4.24}$$

Attention, cette définition ne permet pas *a priori* de définir x_n, pour tout n. Il faut en effet, que pour tout n, x_n soit dans le domaine de définition de f, ce qui sera assuré par les conditions de convergence de la proposition 4.19 page 87.

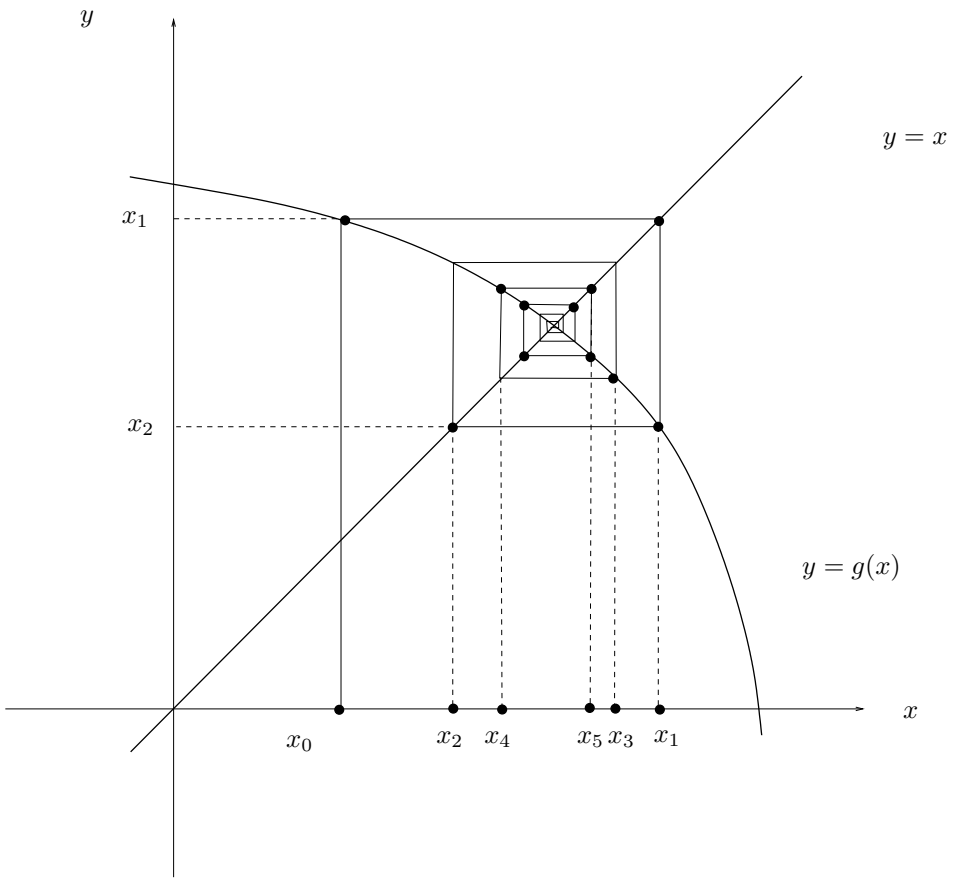


Figure 4.4. fonction g avec un point fixe et les premiers itérés de la suite définie par $x_{n+1} = g(x_n)$ (dans le cas où la fonction g est décroissante, on parle du "colimaçon").

Graphiquement, on peut lire les valeurs de x_n comme le montrent les figures 4.4 et 4.5.

Remarque 4.9. Si la suite x_n tend vers l et si g est continue en l, alors nécessairement l vérifie $l = g(l)$.

4.4.2. Convergence

Tous les choix de g ne mènent pas nécessairement à une convergence. Voir l'exercice de TD 4.5 ou l'exemple 4.10 qui sera justifié en TD (voir exercice 4.2).

Exemple 4.10. On considère le polynôme P donné par

$$P(x) = x^2 - 3 - 2\,x.$$

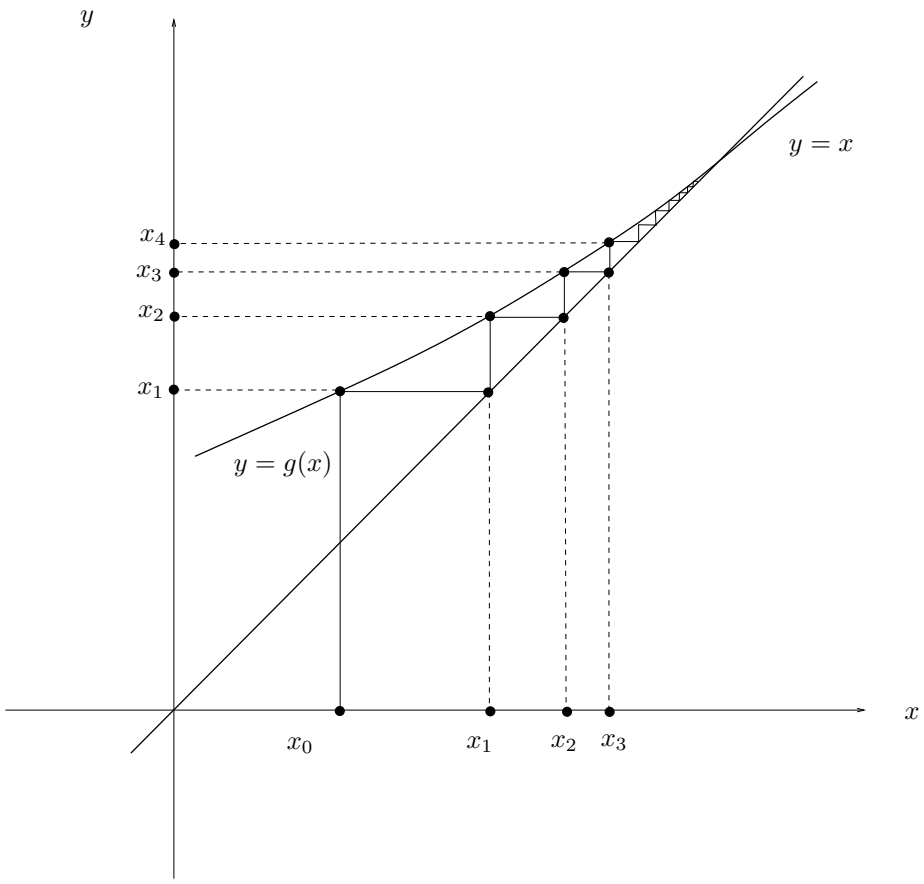


FIGURE 4.5. fonction g avec un point fixe et les premiers itérés de la suite définie par $x_{n+1} = g(x_n)$ (dans le cas où la fonction g est croissante, on parle de l'"escalier").

dont les racines sont $\{3, -1\}$. On transforme l'équation $P(x) = 0$ en une équation de point fixe de trois façons différentes. Elle est équivalente à $g_i(x) = x$, pour $i \in \{1, 2, 3\}$ avec

$$g_1(x) = \sqrt{2\,x + 3},$$
$$g_2(x) = 3\ (x - 2)^{-1},$$
$$g_3(x) = 1/2\,x^2 - 3/2.$$

Les comportements de convergence dépendent du choix de g_i comme le montre le tableau 4.4.

Les deux premiers choix présente une convergence vers l'une des racines de P, tandis que le dernier choix correspond à une divergence.

n	g_1	g_2	g_3
0	4	4	4.
1	3.3166247903554	1.5000000000000	6.500000000000
2	3.1037476670488	-6	$1.962499999999\ 10^{1}$
3	3.0343854953017	-0.3750000000000	$1.910703125000\ 10^{2}$
4	3.0114400194265	-1.2631578947368	$1.825243215942\ 10^{4}$
5	3.0038109192912	-0.9193548387097	$1.665756383672\ 10^{8}$
6	3.0012700375978	-1.0276243093923	$1.387372164872\ 10^{16}$
7	3.0004233159999	-0.9908759124088	$9.624007619305\ 10^{31}$
8	3.0001411020150	-1.0030506406345	$4.631076132821\ 10^{63}$
9	3.0000470336363	-0.9989841527834	$1.072343307399\ 10^{127}$
10	3.0000156778378	-1.0003387304383	$5.749600844623\ 10^{253}$
11	3.0000052259414	-0.9998871026012	$+\infty$
12	3.0000017419800	-1.0000376338825	$+\infty$
13	3.0000005806599	-0.9999874555299	$+\infty$
14	3.0000001935533	-1.0000041815075	$+\infty$
15	3.0000000645178	-0.9999986061661	$+\infty$
16	3.0000000215059	-1.0000004646115	$+\infty$
17	3.0000000071686	-0.9999998451295	$+\infty$
18	3.0000000023895	-1.0000000516235	$+\infty$
19	3.0000000007965	-0.9999999827922	$+\infty$
20	3.0000000002655	-1.0000000057359	$+\infty$
21	3.0000000000885	-0.9999999980880	$+\infty$
22	3.0000000000295	-1.0000000006373	$+\infty$
23	3.0000000000098	-0.9999999997876	$+\infty$
24	3.0000000000033	-1.0000000000708	$+\infty$

TABLE 4.4. Valeurs des itérés du point fixe x_n pour plusieurs choix de g_i

Rappelons que r est le point fixe, solution de (4.23). On suppose que g est définie sur un intervalle du type $[r-\varepsilon_0, r+\varepsilon_0]$. La valeur de $|g'(r)|$ est très importante pour la convergence de la méthode du point fixe. Si [2] $|g'(r)|>1$, la convergence du point fixe est impossible, sauf si $x_0 = r$ (auquel cas, la suite est constante et vaut x_0). Voir propositions 4.11 et 4.12. Au contraire, la condition $|g'(r)|<1$ ne suffit pas à assurer la convergence du point fixe. Mais associée à une autre condition, cette condition implique à la fois la convergence de la méthode du point fixe et l'existence et l'unicité de ce point fixe. Voir proposition 4.19. Dans le cas $|g'(r)|=1$, on ne peut conclure (voir remarque 4.22).

PROPOSITION 4.11 (condition suffisante de divergence de la méthode du point fixe). *Supposons que g soit dérivable en r, qu'*

$$\text{il existe une infinité d'indices } n \text{ tels que } x_n \neq r, \tag{4.25a}$$

et que

$$|g'(r)|>1. \tag{4.25b}$$

Alors, il existe un intervalle du type $I=[r-\varepsilon_0, r+\varepsilon_0]$, tel que, pour tout x_0 dans $I\setminus\{r\}$, la méthode du point fixe (définition 4.8) ne converge pas.

DÉMONSTRATION. Supposons que (4.25b) ait lieu. Raisonnons par l'absurde : il existe un intervalle du type $I=[r-\varepsilon_0, r+\varepsilon_0]$, tel qu'il existe x_0 dans $I\setminus\{r\}$ tel que la méthode du point fixe (définition 4.8) converge. On a donc

$$\forall \varepsilon>0, \quad \exists N\in\mathbb{N}^*, \quad \forall n\geq N, \quad |x_n - r|\leq \varepsilon. \tag{4.26}$$

Par ailleurs, la définition de la dérivée de $g'(r)$ fournit

$$\forall \varepsilon>0, \quad \exists \eta>0, \quad \forall x\in[r-\eta, r+\eta]\setminus\{r\}, \quad \left|\frac{g(x)-g(r)}{x-r} - g'(r)\right|\leq\varepsilon, \tag{4.27}$$

ce qui s'écrit aussi

$$\forall \varepsilon>0, \quad \exists \eta>0, \quad \forall x\in[r-\eta, r+\eta]\setminus\{r\}, \quad g'(r)-\varepsilon\leq\frac{g(x)-g(r)}{x-r}\leq g'(r)+\varepsilon \tag{4.28}$$

Supposons sans perte de généralité que $g'(r)>1$. Dans (4.28), on choisit $\varepsilon=(g'(r)-1)/2>0$ de telle sorte que

$$\alpha_0 = g'(r)-\varepsilon-1 = \frac{1}{2}\left(2g'(r)-g'(r)+1-2\right) = \frac{1}{2}\left(g'(r)-1\right)>0.$$

On a donc

$$\exists \eta_0>0, \quad \forall x\in[r-\eta_0, r+\eta_0]\setminus\{r\}, \quad 1+\alpha_0\leq\frac{g(x)-g(r)}{x-r},$$

ce qui implique, d'après (4.23),

$$\exists \eta_0>0, \quad \forall x\in[r-\eta_0, r+\eta_0]\setminus\{r\}, \quad (1+\alpha_0)\,|x-r|\leq|g(x)-r|\,,$$

ce qui est aussi vrai pour $x=r$:

$$\exists \eta_0>0, \quad \forall x\in[r-\eta_0, r+\eta_0], \quad (1+\alpha_0)\,|x-r|\leq|g(x)-r|\,, \tag{4.29}$$

On choisit dans (4.26), $\varepsilon=\eta_0$, de sorte que

$$\exists N_0\in\mathbb{N}^*, \quad \forall n\geq N_0, \quad x_n\in[r-\eta_0, r+\eta_0], \tag{4.30}$$

ce qui implique donc d'après (4.29) appliqué à $x=x_n$:

$$\forall n\geq N_0, \quad (1+\alpha_0)\,|x_n-r|\leq|x_{n+1}-r|\,,$$

dont on déduit

$$(1+\alpha_0)\,\left|x_{N_0}-r\right|\leq\left|x_{N_0+1}-r\right|,$$

puis

$$(1+\alpha_0)^2\,\left|x_{N_0}-r\right|\leq\left|x_{N_0+2}-r\right|,$$

et par récurrence

$$\forall p\geq 0, \quad (1+\alpha_0)^p\,\left|x_{N_0}-r\right|\leq\left|x_{N_0+p}-r\right|. \tag{4.31}$$

D'après l'hypothèse (4.25a), il existe au moins un indice N_0 tel que $x_{N_0}-r\neq 0$. Puisque $1+\alpha_0>1$, $\lim_{p\to+\infty}|x_p-r|=+\infty$ et donc d'après (4.31), pour p assez grand, x_{N_0+p} n'est plus dans $[r-\eta_0, r+\eta_0]$, ce qui contredit (4.30). Cela qui contredit notre hypothèse de départ, qui est la convergence de la suite. Donc la suite diverge. □

2. et si d'autres conditions précisées plus bas sont vérifiées.

♢

Pour le second résultat, on renforce les hypothèses.

PROPOSITION 4.12 (condition suffisante de divergence de la méthode du point fixe). *Supposons qu'il existe un intervalle $I = [a, b]$ de $\mathbb{R}$ tel que $r \in I$ et sur lequel g vérifie les trois hypothèses suivantes :*

$$g \text{ est définie sur } I, \tag{4.32a}$$

$$\text{pour tout } x \notin I, \text{ si } g(x) \text{ est défini, il n'appartient pas à } I, \tag{4.32b}$$

$$\text{il existe un réel } k > 1 \text{ tel que } \forall x \in I, \quad |g'(x)| \geq k. \tag{4.32c}$$

Alors, pour tout x_0 dans $I \setminus \{r\}$, la méthode du point fixe ne converge pas.

REMARQUE 4.13. Remarquons que (4.32b) est équivalent à

$$g(\mathbb{R} \setminus I) \subset \mathbb{R} \setminus I. \tag{4.33}$$

♢

DÉMONSTRATION DE LA PROPOSITION 4.12. L'égalité des accroissements finis implique

$$\forall x \in I, \quad \exists z \in [r, x], \quad g(x) - g(r) = g'(z)(x - r)$$

et en particulier, pour $x = x_n$:

$$\forall n \in \mathbb{N}, \quad \left(x_n \in I \Longrightarrow \exists z_n \in [r, x_n], \quad x_{n+1} - r = g'(z_n)(x - r)\right),$$

et l'inégalité (4.32c) implique donc

$$\forall n \in \mathbb{N}, \quad (x_n \in I \Longrightarrow |x_{n+1} - r| \geq k\,|x - r|),$$

ce qui implique, comme dans la preuve de la proposition 4.11, que

$$\forall n \in \mathbb{N}, \quad (x_n \in I \Longrightarrow |x_n - r| \geq k^n\,|x_0 - r|). \tag{4.34}$$

Puisque $x_0 \neq 0$ et $k > 1$, on a donc

$$\lim_{n \to +\infty} k^n\,|x_0 - r| = +\infty. \tag{4.35}$$

On sait que x_0 appartient à I. On considère donc le plus grand entier $p \in \mathbb{N}$ tel que $x_p \in I$. Pour cela, on considère l'ensemble des entiers n tel que $x_n \in I$. Il est non vide et majoré d'après (4.35). p est donc son plus grand élément. Ainsi, on a

$$\forall n \in \{0, ..., p\}, \quad x_n \in I \tag{4.36}$$

et

$$x_{p+1} \notin I. \tag{4.37}$$

Puisque $x_p \in I$, l'hypothèse (4.32a) implique que x_{p+1} est bien défini. Ensuite, de deux choses l'une : ou bien $g(x_{p+1})$ est défini, et dans ce cas, l'hypothèse (4.32b) implique que $x_{p+2} = g(x_{p+1})$ n'est pas dans I ; ou bien $g(x_{p+1})$ n'est pas défini (c'est-à-dire que x_{p+1} n'est pas dans le domaine de définition de g) et la convergence n'a pas lieu puisque la suite n'est pas définie ! On recommence : ou bien $g(x_{p+2})$ est défini, et dans ce cas, l'hypothèse (4.32b) implique que $x_{p+3} = g(x_{p+2})$ n'est pas dans I ; ou bien $g(x_{p+2})$ n'est pas défini et la convergence n'a pas lieu puisque la suite n'est pas définie ! On montre ensuite par récurrence sur $n \geq p + 1$ que, tant que x_n est défini, il n'est pas dans I. Si la suite n'est plus définie à partir d'un rang, elle ne converge pas. Sinon, la suite (x_n) ne peut converger vers $r \in I$. Dans ce cas, en effet, il existerait un N assez grand, à partir duquel x_n appartiendrait à I. □

♢

REMARQUE 4.14. La proposition 4.11 n'est guère utilisable en l'état, car nous n'avons *a priori*, aucun renseignement sur le comportement de la suite (x_n). En pratique, on lui préfère donc la proposition 4.12. Si les hypothèses de cette proposition sont vérifiées, on a donc montré la divergence de la méthode du point fixe. Sinon, il faut étudier à la main la divergence de la suite.

REMARQUE 4.15. Classiquement, les réels x tel que $|g'(x)| < 1$ sont appelés attractifs, puisque la suite converge, d'après la proposition 4.19 et les points x tels que $|g'(x)| > 1$ sont appelés répulsifs, puisque la suite diverge, d'après les propositions 4.11 et 4.12.

Remarque 4.16. Très souvent, seules les hypothèses (4.25b) ou (4.32c), sont citées dans la littérature et *à tort*, les gens pensent que cela suffit à montrer la divergence de la suite (x_n) comme par exemple dans `https://www.univers-ti-nspire.fr/files/pdf/14-th_point_fixe-TNS21.pdf`.

Dans ce genre de preuve, l'auteur montre comme on fait ici, que si $x_0 \neq r$, alors on a (4.34) et en déduit, que si x_n est toujours défini, x_n tend vers l'infini et donc la suite diverge. Le problème de cette preuve est que cette estimation (4.34) n'est justement valable que si x_n est assez proche de r. Ainsi, si x_n s'écarte trop de r, il se peut que (4.34) soit fausse et que x_{n+} se rapproche de nouveau de r ou lui soit égal, auquel cas la suite converge de nouveau !

Par exemple, on considère g définie par

$$g(x) = 1 - \alpha(x-1)(x-2),$$

de point fixe 1 car $g(1) = 1$. De plus, $g'(1) = \alpha$ qu'il suffit de choisir strictement plus grand que 1. On peut montrer qu'il existe un choix judicieux de x_0 tel qu'il existe un p tel que $x_p = 2$, loin de 1. On a alors $x_{p+1} = g(x_p) = g(2) = 1$ et donc la suite est stationnaire à partir de ce p est donc convergente.

♢

Exemple 4.17. Prenons $\alpha > 1$ et

$$g(x) = \alpha x, \tag{4.38}$$

dont le seul point fixe est $x = 0$. La proposition 4.12 s'applique en prenant par exemple $I = [-1, 1]$. Si $x \notin I$, $|x| \geq 1$ et $|g(x)| = \alpha|x| > 1$ et donc $g(x)$ est défini et n'appartient donc pas à I. L'hypothèse (4.32b) est donc vraie. On a aussi, pour tout x, $g'(x) = \alpha$ et (4.32c) est vraie. De plus, on peut aussi directement déterminer x_n donné explicitement par $x_n = \alpha^n x_0$, qui tend bien vers $\pm\infty$ selon le signe de x_0, choisi non nul.

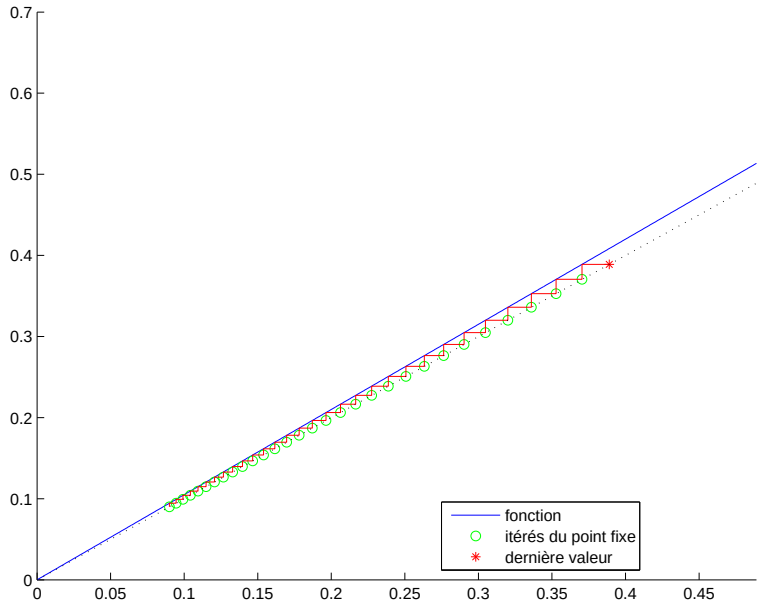


Figure 4.6. Les premiers itérés de la méthode du point fixe pour f définie par (4.38).

Voir la figure 4.6 où on a choisit $\alpha = 1.05$ et $x_0 = 0.09$.

Voir l'annexe T pour plus de détails sur un exemple illustrant les propositions 4.11 et 4.12.

Donnons la définition suivante :

Définition 4.18 (Intervalle g-stable). Si g est une fonction définie au moins sur un intervalle I de $\mathbb{R}$, on dit que I est g-stable ssi $g(I) \subset I$, ce qui est aussi équivalent à pour tout x dans I, $g(x)$ appartient à I.

Donnons alors le résultat très important suivant :

PROPOSITION 4.19 (condition suffisante de convergence de la méthode du point fixe). *Supposons qu'il existe un intervalle $I = [a, b]$ de $\mathbb{R}$ sur lequel g vérifie les deux hypothèses suivantes :*

$$g \text{ est définie sur } I \text{ et } I \text{ est } g\text{-stable ;} \tag{4.39a}$$

$$g \text{ est dérivable sur } I \text{ et il existe un réel } k \text{ de } [0, 1[\text{ tel que } \forall x \in I, \quad |g'(x)| \leq k. \tag{4.39b}$$

Alors, on a les deux résultats suivants :

(1) g admet un point fixe unique r dans $I = [a, b]$;

(2) pour tout x_0 de I, la suite (x_n) est définie et converge vers r.

DÉMONSTRATION. Voir [BM03, Les Théorèmes 4.6, 4.8 et 4.12].
Reprenons-en ici la preuve, très légèrement différemment.

- On vérifie tout d'abord que l'hypothèse de stabilité de la définition 4.18 permet que de montrer que, pour si x_0 appartient à I, pour tout n, x_n est dans I. En effet, x_0 est dans I, donc $x_1 = g(x_0)$ est dans I, et, par récurrence, pour tout n, x_n est dans I.
- Remarquons aussi que l'hypothèse (4.39b) implique l'inégalité [3]

$$\text{il existe un réel } k \text{ de } [0, 1[\text{ tel que } \forall x, y \in I, \quad |g(x) - g(y)| \leq k|x - y|. \tag{4.40}$$

  Cela provient de l'inégalité des accroissements finis, par exemple.
- Posons $I = [a, b]$ et considérons f définie par

$$\forall x \in I, \quad f(x) = g(x) - x. \tag{4.41}$$

  Puisque I est g-stable, on a $g(a) \in [a, b]$ et donc $f(a) = g(a) - a \geq 0$ et, de même, $f(b) = g(b) - b \leq 0$. Puisque g est continue, f l'est aussi. Le théorème des valeurs intermédiaires implique donc que f admet un zéro qui est un point fixe de g.
- On déduit ensuite de (4.40), l'unicité du point fixe. Supposons en effet qu'il en existe deux, distincts, notés α et β. On aurait donc

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq k\,|\alpha - \beta|\,,$$

  par division par $|\alpha - \beta| \neq 0$, on aurait

$$1 \leq k,$$

  ce qui est absurde car on a supposé $k < 1$.
- Notons donc r l'unique point fixe de g. Appliquons maintenant (4.40) à $x = r$ et $y = x_n$. On a donc, pour tout $n \in \mathbb{N}^*$

$$|x_n - r| = |g(x_{n-1}) - g(r)| \leq k\,|x_{n-1} - r|\,,$$

  ce qui implique de même :

$$|x_n - r| \leq k^2\,|x_{n-2} - r|\,,$$

  et par un récurrence immédiate (comme dans la preuve de la proposition 4.11)

$$\forall n \in \mathbb{N}, \quad |x_n - r| \leq k^n\,|x_0 - r|\,,$$

  et *a fortiori*

$$\forall n \in \mathbb{N}, \quad |x_n - r| \leq k^n\,(b - a)\,.$$

  Puisque $k \in [0, 1[$, on a

$$\lim_{n \to +\infty} k^n = 0,$$

  et donc

$$\lim_{n \to +\infty} x_n = r.$$

□

3. Qui pourrait se substituer à (4.39b).

De cette démonstration, on déduit immédiatement la proposition suivante :

PROPOSITION 4.20 (Majoration de l'erreur pour la méthode du point fixe). *Sous les hypothèses de la proposition 4.19, on a*

$$\forall n \in \mathbb{N}, \quad |x_n - r| \leq k^n (b - a). \tag{4.42}$$

On en déduit alors la proposition suivante, identique à la proposition 4.5.

PROPOSITION 4.21. *Avec la méthode du point fixe (sous les hypothèses de la proposition 4.19), si on cherche le point fixe r avec une erreur absolue ε (précision recherchée), il faut effectuer n itérations telles que*

$$k^n (b - a) \leq \varepsilon, \tag{4.43}$$

soit encore

$$n \geq -\frac{\ln((b-a)/\varepsilon)}{\ln k}, \tag{4.44}$$

soit encore,

$$n = \left\lceil -\frac{\ln((b-a)/\varepsilon)}{\ln k} \right\rceil, \tag{4.45}$$

où $\lceil . \rceil$ est défini par (4.22).

DÉMONSTRATION. Il suffit de remplacer, dans la preuve de la proposition 4.5, 1/2 par k. □

♢

REMARQUE 4.22. Pour le cas limite, $|g'(r)| = 1$, on peut avoir convergence ou non, comme le montre par exemple [BM03, Exercice 4.8]. Si on prend $g(x) = x - x^3$ pour laquelle $g'(0) = 1$, la suite des itérés converge pour tout $x_0 \in [-1, 1]$. En revanche, si $g(x) = x + x^3$ pour laquelle $g'(0) = 1$, alors il n'y a jamais convergence sauf si $x_0 = 0$.

On pourra aussi consulter l'exemple de la question 3 du corrigé de l'exercice de TD 4.2, qui met aussi en évidence une convergence très lente dans le cas où $|g'(r)| = 1$.

REMARQUE 4.23. Dans certains cas, les hypothèses de la propositions 4.19 ne sont pas assurées et cependant, la méthode du point fixe est convergente. Voir [BM03, Exercice 4.11 p. 149].

REMARQUE 4.24. On pourra consulter, pour plus de détails, l'annexe Q, en particulier la majoration (Q.6). On fera attention aux notations différentes de cette annexe !

EXEMPLE 4.25. On recherche les zéros de $f(x) = e^x - 4x^2$ par la méthode de point fixe avec fonctions d'itération suivantes

$$g_1(x) = \frac{1}{2} e^{x/2},$$
$$g_2(x) = -\frac{1}{2} e^{x/2},$$
$$g_3(x) = 2 \ln(2x).$$

Voir les figures 4.7 à 4.9 montrant la convergence ou la divergence.

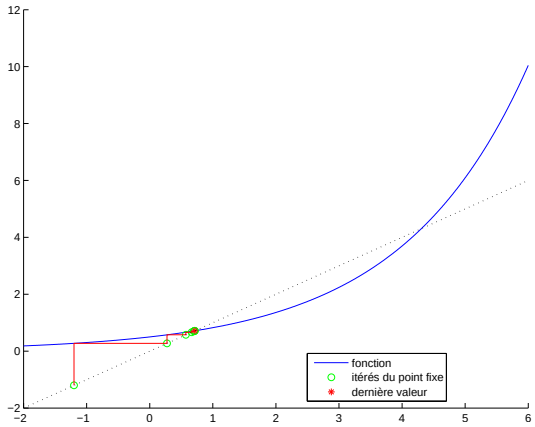
EXEMPLE 4.26. On pourra de nouveau consulter l'exemple de la section 4.1, point 2 page 76.

EXEMPLE 4.27. On pourra consulter l'annexe P page 217.

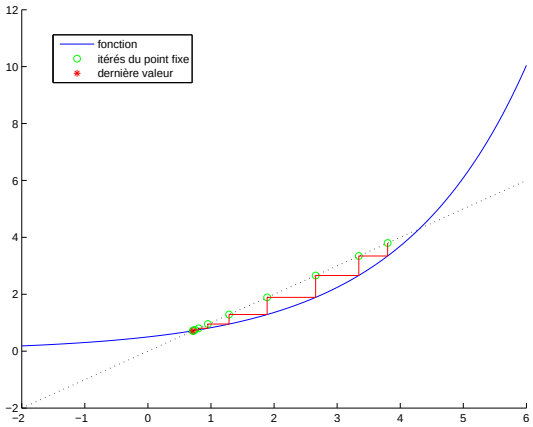
EXEMPLE 4.28. On pourra consulter l'annexe AA page 305, où sont présentés des résultats de convergence de suite vers π, fondés sur une approche géométrique, dont la célèbre méthode d'Archimède.

REMARQUE 4.29. On pourra aussi consulter l'annexe N page 211 pour des résultats de convergence un peu plus généraux.

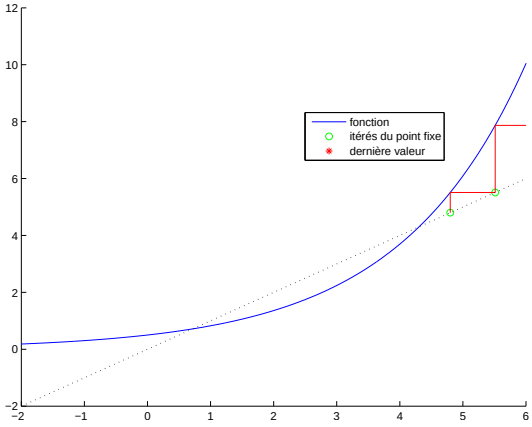
♢



(a) $x_0 = -1.20$

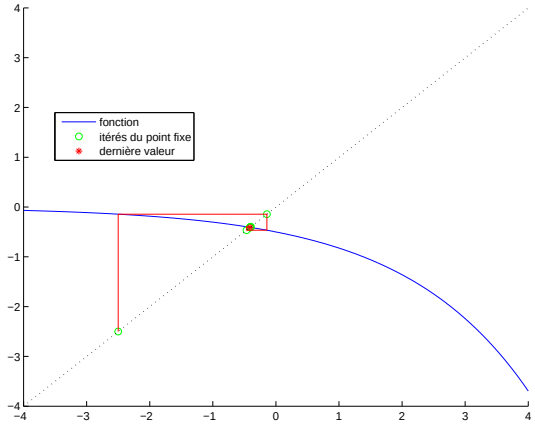


(b) $x_0 = 3.80$

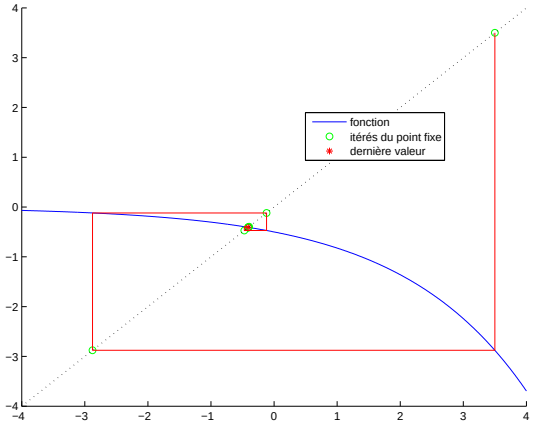


(c) $x_0 = 4.80$

FIGURE 4.7. Itérés du point fixe pour g_1.

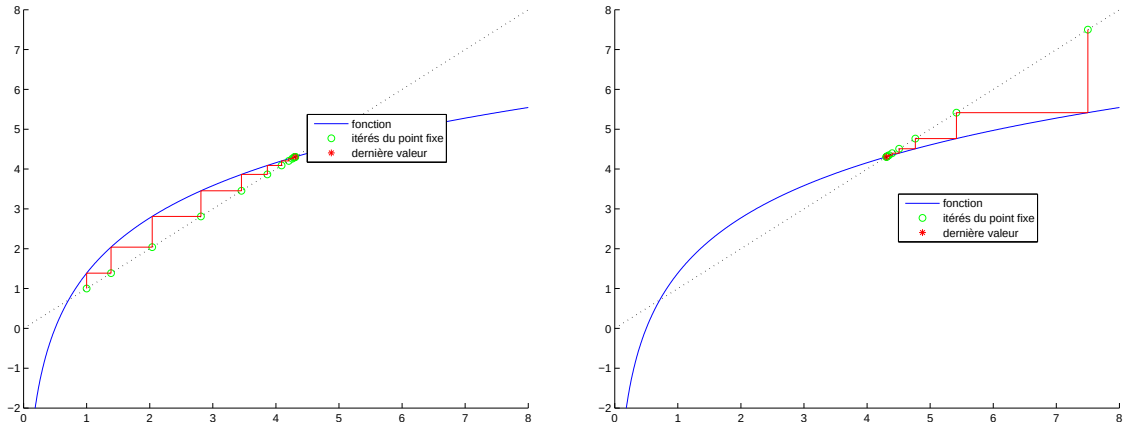


(a) $x_0 = -2.50$



(b) $x_0 = 3.50$

FIGURE 4.8. Itérés du point fixe pour g_2.



(a) $x_0 = 1$ (b) $x_0 = 7.50$

FIGURE 4.9. Itérés du point fixe pour g_3.

4.4.3. Convergence avec un ordre plus élevé

On remarque dans l'inégalité (4.42) que plus k est proche de 0, plus vite l'erreur diminue et plus rapide la convergence est. Que se passe-t-il si $g'(r) = 0$? Les majorations d'erreur en $1/2^n$ de la proposition 4.5 ou en k^n de la proposition 4.20 sont lentes (dites linéaires) et on va s'intéresser à des méthodes convergent plus rapidement. Notons néanmoins l'important avantage des méthodes de dichotomie et de point fixe, qui est qu'on est certain que chaque x_n reste dans l'intervalle de départ, ce qui ne sera pas nécessairement vrai pour des méthodes plus efficaces. Idéalement, on utilise les méthodes de dichotomie ou de point fixe pour s'approcher assez de la racine supposée. Ensuite, on utilise une méthode plus efficace. On traduit la notion d'efficacité par l'ordre, notion qui fait penser à celle des ordres des méthodes d'intégration !

DÉFINITION 4.30. Soit $p \in \mathbb{N}^*$. Nous dirons qu'une méthode de point fixe (4.24) est d'ordre au moins p si, l'erreur $e_n = x_n - r$ vérifie [4] :

$$\exists C \in \mathbb{R}_+, \quad \forall n \in \mathbb{N}, \quad \frac{|e_{n+1}|}{|e_n|^p} \leq C. \tag{4.47}$$

Elle est dite exactement d'ordre p si

$$\lim_{n \to +\infty} \frac{|e_{n+1}|}{|e_n|^p} \neq 0. \tag{4.48}$$

REMARQUE 4.31. La notion d'ordre est une notion, dite locale, qui n'assure pas la convergence globale de la suite x_n vers r, sur un intervalle donné à l'avance. Il faut supposer que x_0 est assez proche de r pour que la convergence ait lieu, avec la définition de l'ordre (4.47). Voir les annexes R et S ou [Sch01].

Donnons le résultat suivant (issu et adapté de [BM03, Annexe D]) :

LEMME 4.32. *Si une méthode de point fixe est au moins d'ordre $p \in \mathbb{N}^*$ (au sens de la définition 4.30), alors*

- *si $p = 1$, on a*

$$\forall n \geq 0, \quad |e_n| \leq |e_0| C^n, \tag{4.49}$$

4. On pourra remplacer ceci par l'inégalité un peu moins exigeante :

$$\exists C \in \mathbb{R}_+, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad \frac{|e_{n+1}|}{|e_n|^p} \leq C. \tag{4.46}$$

- *si $p > 1$, alors en définissant les nombres γ et δ par*

$$\gamma = C^{\left(\frac{1}{1-p}\right)}, \tag{4.50a}$$

$$\delta = |e_0| C^{\left(\frac{1}{p-1}\right)}, \tag{4.50b}$$

on a

$$\forall n \geq 0, \quad |e_n| \leq \gamma \delta^{(p^n)}. \tag{4.51}$$

Démonstration du lemme 4.32. La preuve sera faite dans la correction de l'exercice de TD 4.3 page 45 ou dans l'annexe R page 243. □

Donnons un résultat proche de la proposition 4.21 :

Proposition 4.33 (Majoration de l'erreur pour une méthode d'ordre p.). *Si une méthode de point fixe est au moins d'ordre $p \in \mathbb{N}^*$ et :*

- *si $p = 1$ et si $C < 1$*
- *si $p > 1$ et si $|e_0|$ est assez petit de telle sorte que pour que*

$$|e_0| C^{\left(\frac{1}{p-1}\right)} < 1, \tag{4.52}$$

la suite x_n tend vers r. Sous ces hypothèses, si on cherche le point fixe r avec une erreur absolue ε (précision recherchée), il faut effectuer n itérations avec n défini par que

- *si $p = 1$*

$$n = \left\lceil -\frac{\ln\left(|e_0|/\varepsilon\right)}{\ln C} \right\rceil, \tag{4.53}$$

- *si $p > 1$*

$$n = \left\lceil \frac{1}{\ln p} \ln\left(\frac{\ln \frac{\varepsilon}{\gamma}}{\ln \delta} \right) \right\rceil, \tag{4.54}$$

avec γ et δ définis par (4.50).

On rappelle que $\lceil . \rceil$ est défini par (4.22).

Démonstration. La preuve sera faite dans la correction de l'exercice de TD 4.3 page 45 ou dans l'annexe R page 243. □

♢

Proposition 4.34. *Si g est de classe $\mathcal{C}^1$ sur un intervalle du type $[r - \varepsilon, r + \varepsilon]$, avec ε assez petit, si x_0 appartient à $[r - \varepsilon, r + \varepsilon]$, si $g(r) = r$ et si*

$$|g'(r)| < 1, \tag{4.55}$$

alors la méthode de point fixe est au moins linéaire, c'est-à-dire que la définition 4.30 est vérifiée pour $p = 1$. Si, de plus,

$$g'(r) \neq 0, \tag{4.56}$$

alors la méthode de point fixe est exactement linéaire.

Remarque 4.35. La convergence globale sera aussi assurée, si par exemple, les hypothèses de la proposition 4.19 sont satisfaites.

Rappelons la formule de Taylor-Lagrange (voir par exemple [Bas22a, Section : "1.4. Développements limités"] ou [BM03, Annexe A]), qui sera utilisée un certain nombre de fois :

Théorème 4.36 (Formule de Taylor-Lagrange). *Soient $a, b \in \mathbb{R}$ avec $a < b$ et f une fonction de $\mathbb{R}$ dans $\mathbb{R}$. Si f est de classe C^p sur $[a, b]$ et admet une dérivée d'ordre $p + 1$ en tout point de $]a, b[$, alors il existe $\xi \in]a, b[$ tel que*

$$f(b) = \sum_{k=0}^{p} \frac{f^{(k)}(a)}{k!} (b - a)^k + \frac{f^{(p+1)}(\xi)}{(p+1)!} (b - a)^{p+1}. \tag{4.57}$$

♢

Démonstration de la proposition 4.34.

Donnons deux preuves légèrement différentes :

(1) Notons

$$I = [r - \varepsilon, r + \varepsilon] \tag{4.58}$$

Montrons qu'il existe $\varepsilon > 0$ assez petit tel que

$$\forall n \in \mathbb{N}, \quad x_n \in I. \tag{4.59}$$

et que (4.47) et (4.48) ont lieu pour $p = 1$. D'après (4.55) et puisque g' est continue au voisinage de r, il existe $\varepsilon > 0$ assez petit tel que

$$C = \max_{x \in I} |g'(x)| < 1. \tag{4.60}$$

Démontrons (4.47) (pour $p = 1$) et (4.59) par récurrence sur n. Pour $n = 0$, (4.59) est vraie par hypothèse. Supposons maintenant (4.59) vraie pour un $n \geq 0$. Appliquons la formule (4.57) à g avec $a = r$, $b = x_n$ et $p = 0$, ce qui n'est autre que l'égalité des accroissements finis : il existe ξ_n tel que

$$\xi_n \in]r, x_n[\tag{4.61}$$

tel que

$$g(x_n) = g(r) + \frac{g'(\xi_n)}{(1)!}(x_n - r)^n = g(r) + g'(\xi_n)(x_n - r),$$

et donc

$$x_{n+1} - r = g(x_n) - g(r) = g'(\xi_n)(x_n - r), \tag{4.62}$$

soit, si $x_n \neq r$:

$$\frac{e_{n+1}}{e_n} = g'(\xi_n). \tag{4.63}$$

D'après (4.59) appliqué à n (hypothèse de récurrence) et (4.61), ξ_n appartient à I. Ainsi, d'après (4.60), on en déduit d'une part que (4.47) est vraie. Cela implique aussi

$$|e_{n+1}| \leq C\,|e_n|\,. \tag{4.64}$$

D'après (4.59) appliqué à n, on a $|e_n| \leq 1$ et d'après (4.60), on a donc $|e_{n+1}| \leq 1$ et (4.59) est vraie pour $n+1$. D'autre part d'après (4.64), (4.47) est vérifiée avec $p = 1$ en prenant la constante C définie par (4.60). La méthode est donc au moins linéaire. De plus, d'après (4.59) et (4.61) ξ_n tend vers r et, puisque g' est continue, $g'(\xi_n)$ tend vers $g'(r)$ et donc

$$\lim_{n \to +\infty} \frac{e_{n+1}}{e_n} = g'(r).$$

et donc la méthode est exactement linéaire d'après (4.48) et (4.56).

(2) La seconde est identique à la première, sauf que comme dans [BM03, Théorème 4.12], on remarque que (4.60) et l'inégalité des accroissements finis impliquent, que pour tout x, y dans $[r - \varepsilon, r + \varepsilon]$, on a

$$|g(x) - g(y)| \leq C|x - y|,$$

et en particulier, pour $y = r$, on a donc

$$|g(x) - r| \leq C|x - r| \leq |x - r|,$$

dont on déduit que $[r - \varepsilon, r + \varepsilon]$ est g-stable. Ainsi, pour tout x_0 de $[r - \varepsilon, r + \varepsilon]$, les x_n appartiennent à $[r - \varepsilon, r + \varepsilon]$. On finit comme précédemment en utilisant directement (4.63) qui implique l'ordre 1.

□

Remarque 4.37. Dans ce cas, la suite x_n converge pour tout x_0 dans $[r - \varepsilon, r + \varepsilon]$ et la proposition 4.33 s'applique (dans le cas $p = 1$).

En général, une méthode de point fixe est donc linéaire ..., sauf si son ordre est supérieur à 2, comme le montre la suite.

PROPOSITION 4.38. *Si g est de classe $\mathcal{C}^2$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$, avec ε assez petit, si x_0 appartient à $[r-\varepsilon, r+\varepsilon]$, si $g(r) = r$ et si*

$$g'(r) = 0. \tag{4.65}$$

alors la méthode de point fixe est au moins quadratique, c'est-à-dire que la définition 4.30 est vérifiée pour $p = 2$. Si, de plus,

$$g''(r) \neq 0, \tag{4.66}$$

alors la méthode de point fixe est exactement quadratique.

DÉMONSTRATION. Comme dans la proposition 4.34, donnons deux preuves légèrement différentes :

(1) La preuve est proche de la proposition 4.34 sauf que l'on applique la formule (4.57) à g avec $a = r$, $b = x_n$ et cette fois-ci $p = 1$: il existe ξ_n vérifiant (4.61) tel que

$$g(x_n) = g(r) + g'(r)(x_n - r) + \frac{1}{2}g''(\xi_n)(x_n - r)^2. \tag{4.67}$$

ce qui donne

$$x_{n+1} - r = g(x_n) - g(r) = g'(r)(x_n - r) + \frac{1}{2}g''(\xi_n)(x_n - r)^2, \tag{4.68}$$

soit, selon (4.65)

$$x_{n+1} - r = \frac{1}{2}g''(\xi_n)(x_n - r)^2, \tag{4.69}$$

soit, si $x_n \neq r$:

$$\frac{e_{n+1}}{e_n^2} = \frac{1}{2}g''(\xi_n). \tag{4.70}$$

Ensuite, le raisonnement est proche de la proposition 4.34 Dans toute cette preuve, on pose dorénavant

$$p = 2. \tag{4.71}$$

Notons I défini par (4.58). Montrons qu'il existe $\varepsilon > 0$ assez petit tel que (4.59) soit vérifiée et que (4.47) et (4.48) aient lieu. On pose, pour un $\varepsilon_0 > 0$ donné

$$C = \frac{1}{2}\max_{x\in[r-\varepsilon_0, r+\varepsilon_0]} |g''(x)|. \tag{4.72}$$

Choisissons ensuite $\varepsilon \in [0, \varepsilon_0]$ assez petit pour que

$$2\varepsilon C^{\left(\frac{1}{p-1}\right)} < 1, \tag{4.73}$$

ait lieu ce qui est loisible, il suffit pour cela de choisir

$$\varepsilon < \min\left\{\varepsilon_0, \frac{1}{2C^{\left(\frac{1}{p-1}\right)}}\right\}. \tag{4.74}$$

Démontrons (4.47) et (4.59) par récurrence sur n. Pour $n = 0$, (4.59) est vraie par hypothèse. Supposons maintenant (4.59) vraie pour un $n \geq 0$. D'après (4.59) appliqué à n (hypothèse de récurrence) et (4.61), ξ_n appartient à I. Cela implique d'une part d'après (4.72)

$$\frac{|e_{n+1}|}{|e_n|^p} \leq C, \tag{4.75}$$

et donc que (4.47) est vraie. D'autre part, on écrit :

$$\frac{|e_{n+1}|}{|e_n|} = \frac{|e_{n+1}|}{|e_n|^p}\frac{|e_n|^p}{|e_n|} = \frac{|e_{n+1}|}{|e_n|^p}|e_n|^{p-1}$$

et donc, d'après (4.75)

$$\frac{|e_{n+1}|}{|e_n|} \leq C|e_n|^{p-1},$$

et donc d'après (4.59) appliqué à n, on a

$$\frac{|e_{n+1}|}{|e_n|} \leq C(2\varepsilon)^{p-1},$$

ce que l'on écrit encore

$$\frac{|e_{n+1}|}{|e_n|} \leq \left(2\varepsilon C^{(\frac{1}{p-1})}\right)^{p-1} \tag{4.76}$$

et donc d'après le choix (4.73)

$$\frac{|e_{n+1}|}{|e_n|} \leq 1. \tag{4.77}$$

D'après (4.59) appliqué à n, on a $|e_n| \leq 1$ et d'après (4.77), on a donc $|e_{n+1}| \leq 1$ et (4.59) est vraie pour $n+1$. Ainsi, d'après (4.75), (4.47) est vraie avec constante C définie par (4.72). La méthode est donc au moins quadratique. De plus, d'après (4.59) et (4.61) ξ_n tend vers r et, puisque g'' est continue, $g''(\xi_n)$ tend vers $g''(r)$ et donc (4.70) implique

$$\lim_{n\to+\infty} \frac{e_{n+1}}{e_n^2} = \frac{1}{2}g''(r). \tag{4.78}$$

et donc la méthode est exactement quadratique d'après (4.66).

(2) Le raisonnement est identique. Puisque $g'(r) = 0$, et d'après, la convergence est au moins linéaire et d'après la proposition 4.34, pour ε assez petit, les x_n appartiennent tous à $[r-\varepsilon, r+\varepsilon]$. On considère alors

$$C = \frac{1}{2}\max_{x\in[r-\varepsilon,r+\varepsilon]} |g''(x)|, \tag{4.79}$$

et on montre (4.48) pour $p = 2$ directement à partir de (4.70).

□

REMARQUE 4.39. Dans ce cas, la suite x_n converge pour tout x_0 dans $[r-\varepsilon, r+\varepsilon]$ et la proposition 4.33 s'applique (dans le cas $p = 2$). De plus, on remarque que le choix (4.73) est conforme au choix (4.52).

Itérons une dernière fois !

PROPOSITION 4.40. *Si g est de classe $\mathcal{C}^3$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$, avec ε assez petit, si x_0 appartient à $[r-\varepsilon, r+\varepsilon]$, si $g(r) = r$ et si*

$$g'(r) = 0, \tag{4.80a}$$

$$g''(r) = 0, \tag{4.80b}$$

alors la méthode de point fixe est au moins cubique, c'est-à-dire que la définition 4.30 est vérifiée pour $p = 3$. Si, de plus,

$$g'''(r) \neq 0, \tag{4.81}$$

alors la méthode de point fixe est exactement cubique.

DÉMONSTRATION. La preuve est exactement identique à celle la proposition 4.38 (nous ne présentons que la première variante) sauf que l'on applique la formule (4.57) à g avec $a = r$, $b = x_n$ et cette fois-ci $p = 2$: il existe ξ_n vérifiant (4.61) tel que

$$g(x_n) = g(r) + g'(r)(x_n - r) + \frac{1}{2}g''(r)(x_n - r)^2 + \frac{1}{6}g'''(\xi_n)(x_n - r)^3,$$

et (4.68) devient

$$x_{n+1} - r = g'(r)(x_n - r) + \frac{1}{2}g''(r)(x_n - r)^2 + \frac{1}{6}g'''(\xi_n)(x_n - r)^3,$$

soit, selon (4.80)

$$x_{n+1} - r = \frac{1}{6}g'''(\xi_n)(x_n - r)^3, \tag{4.82}$$

et (4.70) devient

$$\frac{e_{n+1}}{e_n^3} = \frac{1}{6} g'''(\xi_n) \tag{4.83}$$

Ensuite, le raisonnement le même que celui de la proposition 4.38 où (4.71) est remplacé par

$$p = 3. \tag{4.84}$$

L'équation (4.72) est remplacée par

$$C = \frac{1}{6} \max_{x \in [r-\varepsilon_0, r+\varepsilon_0]} |g'''(x)| . \tag{4.85}$$

Enfin, puisque g''' est continue, $g'''(\xi_n)$ tend vers $g'''(r)$ et donc (4.70) implique

$$\lim_{n \to +\infty} \frac{e_{n+1}}{e_n^3} = \frac{1}{6} g'''(r). \tag{4.86}$$

□

REMARQUE 4.41. Dans ce cas, la suite x_n converge pour tout x_0 dans $[r-\varepsilon, r+\varepsilon]$ et la proposition 4.33 s'applique (dans le cas $p = 3$).

REMARQUE 4.42. Ces résultats sont dit locaux : ils n'impliquent pas la convergence sur un intervalle donné, mais il faut supposer que x_0 est assez proche de r. Voir remarques 4.31 et 4.35 ainsi que les remarques 4.37 et 4.39 et 4.41/

REMARQUE 4.43. En général, il est facile de vérifier la nullité des premières dérivées puisque les méthodes sont construites grâce à cette propriété. En revanche, la non nullité de la dernière dérivée n'est pas aisée à obtenir, puisque la racine n'est pas connue. En pratique, on se contentera de montrer que les méthodes sont au moins d'un ordre un, deux ou plus.

REMARQUE 4.44. Une méthode d'ordre 1, un point fixe général, est lente. On lui privilégiera une méthode quadratique. Cela sera revu en section 4.5. Les méthodes cubiques sont rares, car une quadratique suffit largement. Voir néanmoins l'exemple pédagogique 4.49 page 97.

REMARQUE 4.45. Bien que très rares en pratiques, on peut obtenir des méthodes d'ordre p quelconque.

PROPOSITION 4.46. *Soit $p \in \mathbb{N}^*$. Si g est de classe $\mathcal{C}^p$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$, avec ε assez petit, si x_0 appartient à $[r-\varepsilon, r+\varepsilon]$, si $g(r) = r$ et si*

$$\forall q \in \{1, ..., p-1\}, \quad g^{(q)}(r) = 0, \tag{4.87}$$

alors la méthode de point fixe est d'ordre au moins p, c'est-à-dire que la définition 4.30 est vérifiée pour p. Si, de plus,

$$g^{(p)}(r) \neq 0, \tag{4.88}$$

alors la méthode de point fixe est d'ordre exactement p.

DÉMONSTRATION. Cette proposition, proche de [BM03, Proposition D.5], se montre exactement comme la proposition 4.38 sauf que l'on applique la formule (4.57) à g avec $a = r$, $b = x_n$ et cette fois-ci $p-1$. Les équations (4.69) et (4.70) deviennent

$$x_{n+1} - r = \frac{1}{p!} g^{(p)}(\xi_n)(x_n - r)^p,$$

et (4.70) devient

$$\frac{e_{n+1}}{e_n^3 p} = \frac{1}{p!} g^{(p)}(\xi_n)$$

L'équation (4.72) est remplacée par

$$C = \frac{1}{p!} \max_{x \in [r-\varepsilon_0, r+\varepsilon_0]} \left| g^{(p)}((x) \right|,$$

et (4.78) devient

$$\lim_{n \to +\infty} \frac{e_{n+1}}{e_n^p} = \frac{1}{p} g^{(p)}(r).$$

□

♢

Remarque 4.47. On montre en annexe S que, à chaque itération, le nombre de décimales exactes est asymptotiquement multiplié par l'ordre de convergence de la méthode. Ainsi, pour une méthode quadratique, si on a 1 chiffre exact après la virgule (ce qui est aisé d'obtenir avec quelques itérations de la méthode dichotomie par exemple), on en obtient 2 à la deuxième itération, 4 à la troisième, 8 à la quatrième et 16 à la cinquième. On est dans ce cas sous la précision machine de matlab par exemple ! Voir les exemples 4.48, 4.49 et 4.50 qui mettent cela en évidence.

Exemple 4.48. Soit $A \in \mathbb{R}_+^*$ et la méthode de point fixe définie par

$$g(x) = \frac{1}{2}\left(x + \frac{A}{x}\right). \tag{4.89}$$

Un point fixe x de g sur $\mathbb{R}_+^*$ vérifie

$$2x = x + \frac{A}{x}$$

soit $x^2 = A$ et donc, puisque x est strictement positif, $x = \sqrt{A}$, qui est donc unique. On peut montrer que la méthode de point fixe converge pour tout $x_0 \in \mathbb{R}_+$ (c'est-à-dire qu'elle est globale). Voir annexe O page 212.

On vérifie que l'on a successivement

$$\begin{aligned} g\left(\sqrt{A}\right) &= \sqrt{A}, \\ g'\left(\sqrt{A}\right) &= 0, \\ g''\left(\sqrt{A}\right) &= \frac{1}{\sqrt{A}} \neq 0, \end{aligned}$$

et donc, d'après la proposition 4.38, la méthode de point fixe associée à g est d'ordre deux et converge vers $\sqrt{A}$.

Plus de détails dans [BM03, Exercice 4.4]. Notons aussi que cette méthode, connue des Babyloniens il y a 2500 ans est encore celle utilisée par les calculatrice pour déterminer la racine carrée d'un nombre ! Cet exemple sera repris dans l'exemple 4.58. Voir les simulations numériques dans l'exemple 4.50.

Exemple 4.49. Prenons maintenant, pour $A > 0$, g définies par

$$g(x) = \frac{1}{8}\left(3x + \frac{6A}{x} - \frac{A^3}{x^3}\right) \tag{4.90}$$

On vérifie que l'on a successivement

$$\begin{aligned} g\left(\sqrt{A}\right) &= \sqrt{A}, \\ g'\left(\sqrt{A}\right) &= 0, \\ g^{(2)}\left(\sqrt{A}\right) &= 0, \\ g^{(3)}\left(\sqrt{A}\right) &= 3\,A^{-1} \neq 0, \end{aligned}$$

et donc, d'après la proposition 4.40, la méthode de point fixe associée à g est d'ordre trois et converge localement vers $\sqrt{A}$. Cette méthode n'est pas globale. Si on choisit un x_0 trop éloigné de $\sqrt{A}$, la suite x_n ne converge plus nécessairement vers $\sqrt{A}$. Voir par exemple [BM03, exercice 4.6] et [BM03, TP 4.J], ce dernier étant disponible sur `https://www.dunod.com/sciences-techniques/introduction-analyse-numerique-applications-sous-matlab`. Cet exemple est surtout pédagogique. En pratique, la méthode quadratique de l'exemple 4.48 suffit déjà tout-à-fait ! Voir les simulations numériques dans l'exemple 4.50.

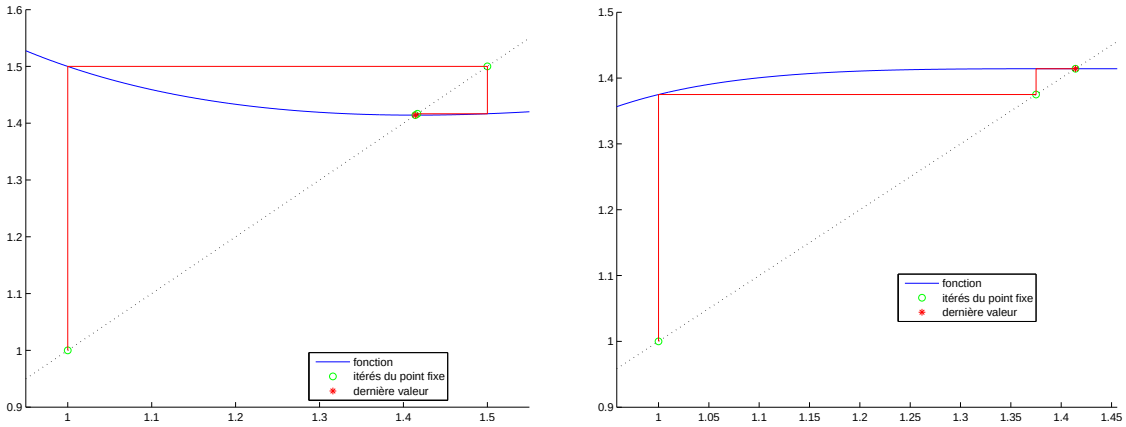
Exemple 4.50.

Présentons des simulations numériques pour les méthodes de points fixes présentées dans les exemples 4.48 et 4.49.

On choisit

$$A = 2, \quad x_0 = 1.$$

Voir la figure 4.10 et le tableau 4.5. Les deux méthodes fournissent $\sqrt{A} = 1.414213562373095$ très rapidement. Comme annoncé dans la remarque 4.47, le mombre de chiffres exacts est multiplié respectivement par 2 et 3 à chaque itération.



(a) Méthode d'ordre 2 (b) Méthode d'ordre 3

FIGURE 4.10. Graphiques associés aux deux méthodes de points fixes des fonctions g définies par (4.89) et (4.90).

n	x_n (ordre 2)	x_n (ordre 3)	$\lvert x_n - \sqrt{A}\rvert$ (ordre 2)	$\lvert x_n - \sqrt{A}\rvert$ (ordre 3)
0	1.000000000000000	1.000000000000000	0.414213562373095	0.414213562373095
1	1.500000000000000	1.375000000000000	0.085786437626905	0.039213562373095
2	1.416666666666667	1.414197501878287	0.002453104293571	0.000016060494808
3	1.414215686274510	1.414213562373094	0.000002123901415	0.000000000000001
4	1.414213562374690	1.414213562373095	0.000000000001595	0.000000000000000
5	1.414213562373095	1.414213562373095	0.000000000000000	0.000000000000000

TABLE 4.5. Valeurs et erreurs des itérés des fonctions g définies par (4.89) et (4.90).

REMARQUE 4.51. On pourra aussi construire une méthode d'odre 4, comme expliqué [BM03, TP 4.K], disponible sur `https://www.dunod.com/sciences-techniques/introduction-analyse-numerique-applications-sous-matlab`. Cet exemple est pédagogique, comme la méthode cubique de l'exemple 4.48.

♢

REMARQUE 4.52. Comme pour les méthodes d'intégration (voir exemple 3.28 page 66), un graphe log log pourra permettre de déterminer de façon numérique, l'ordre d'une méthode de point fixe. Voir l'exercice de TD 4.4.

4.5. Méthode de Newton

Elle constitue la plus utilisées des méthodes la résolution des équations non-linéaires, en raison de sa convergence quadratique (dont on ne rappelle qu'elle n'est que locale !).

4.5.1. Définition

On cherche de nouveau à résoudre (4.16). On suppose que f est dérivable.

DÉFINITION 4.53 (Méthode de Newton). La suite $(x_n)_{n\in\mathbb{N}}$ est définie par la donnée de $x_0 \in \mathbb{R}$ et

$$\forall n \in \mathbb{N}, \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{4.91}$$

Cette façon de procéder peut s'expliquer de deux façons :

(1) D'un point de vue géométrique d'abord. On suppose connu x_n.

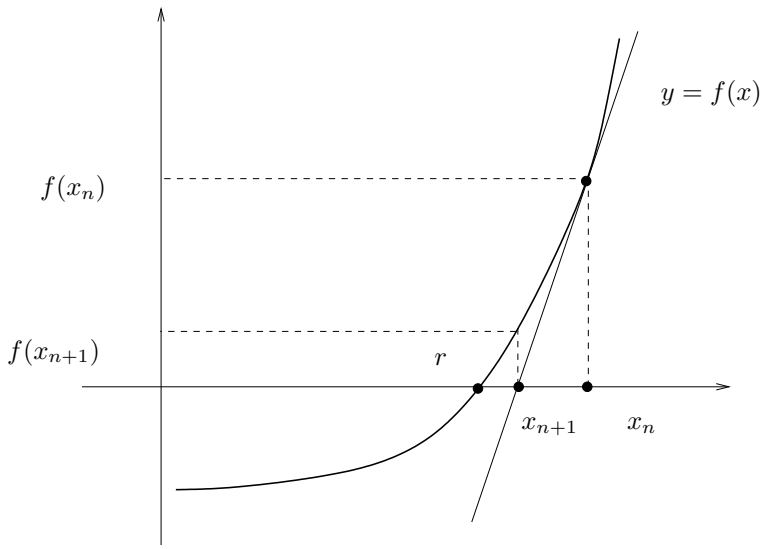


FIGURE 4.11. La méthode de Newton

La tangente Δ à la courbe de f au point d'abscisse x_n a pour équation

$$\frac{Y - f(x_n)}{X - x_n} = f'(x_n),$$

soit

$$Y = f(x_n) + f'(x_n)(X - x_n).$$

Elle coupe l'axe des x au point d'abscisse x_{n+1} qui vérifie donc

$$f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0,$$

et on retrouve (4.91) si $f'(x_n) \neq 0$.

(2) D'un point de vue calculatoire.

Un développement limité de f (voir par exemple [Bas22a, Chapitre]) au voisinage de x_n fournit

$$f(x_n + h) = f(x_n) + hf'(x_n) + o(h).$$

On remplace l'équation exacte $f(x_n + h) = 0$ par

$$f(x_n) + hf'(x_n) = 0$$

et donc

$$h = -\frac{f(x_n)}{f'(x_n)}.$$

On a aussi $h = x_{n+1} - x_n$ et on retrouve donc (4.91).

Ce qui est très important est que la méthode de Newton se ramène à une méthode de point fixe posant

$$g(x) = x - \frac{f(x)}{f'(x)}, \text{ si } f'(x) \neq 0. \tag{4.92}$$

On peut donc appliquer les résultats généraux de la section 4.4.3.

4.5.2. Convergence (locale)

PROPOSITION 4.54. *Si f est de classe $\mathcal{C}^3$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$, si $f(r)=0$, $f'(r)\neq 0$ et $f''(r)\neq 0$, alors la méthode de Newton est exactement quadratique.*

Remarquons que la convergence globale n'est pas assurée *a priori* par une condition suffisante, comme l'était celle du point fixe. On pourra cependant consulter les théorèmes Y.1 et Y.4 page 287.

DÉMONSTRATION DE LA PROPOSITION 4.54. Il suffit d'appliquer la proposition 4.38 à la fonction g définie par (4.92). On a, là où $f'\neq 0$,

$$g' = 1 - \left(\frac{f}{f'}\right)' = 1 - \frac{f'^2 - f''f}{f'^2} = \frac{1}{f'^2}\left(f'^2 - f'^2 + f''f\right),$$

et donc

$$g' = \frac{f''f}{f'^2}, \tag{4.93}$$

et en particulier, puisque $f'(r)\neq 0$,

$$g'(r) = \frac{f''(r)f(r)}{f'^2(r)}. \tag{4.94}$$

Ainsi, $g'(r)=0$ si et seulement si $f(r)=0$, ce qui est vrai. De plus, on peut montrer que

$$g'' = \frac{1}{(f')^4}\left((f')^3 f'' + ff'''(f')^2 - ff''\left((f')^2\right)'\right) \tag{4.95}$$

En particulier, en r, puisque $f'(r)$ est non nul et que $f(r)$ est nul, on a

$$g''(r) = \frac{1}{(f'(r))^4}\left((f'(r))^3 f''(r)\right) = \frac{f''(r)}{f'(r)},$$

et donc

$$g''(r) = \frac{f''(r)}{f'(r)}. \tag{4.96}$$

Ainsi, $g''(r)\neq 0$ si et seulement si $f''(r)\neq 0$, ce qui est vrai. □

REMARQUE 4.55. Remarquons que (4.95) s'écrit aussi

$$\begin{aligned} g'' &= \frac{1}{(f')^4}\left((f')^3 f'' + ff'''(f')^2 - ff''(2f'f'')\right), \\ &= \frac{1}{(f')^4}\left((f')^3 f'' + ff'''(f')^2 - 2ff'(f'')^2\right), \\ &= \frac{1}{(f')^3}\left((f')^2 f'' + ff'f''' - 2f(f'')^2\right). \end{aligned}$$

Dans la majoration On a donc, grâce à (4.72) :

$$C = \frac{1}{2}\max_{x\in[r-\varepsilon,r+\varepsilon]}\left|\frac{1}{(f'(x))^3}\left((f'(x))^2 f''(x) + f(x)f'(x)f'''(x) - 2f(x)(f''(x))^2\right)\right|. \tag{4.97}$$

♢

REMARQUE 4.56. Il sera assez facile de montrer que $f'(r)\neq 0$, en étudiant par exemple f. En revanche, il sera en pratique difficile de montrer que $f''(r)\neq 0$, puisque r n'est pas connu. En pratique, on se contentera du fait que la méthode de Newton est au moins quadratique. En général, elle ne sera pas cubique.

REMARQUE 4.57. On notera aussi que si $f'(r)=0$, la méthode de Newton redevient linéaire. Il est possible de modifier alors cette méthode pour qu'elle redevienne bien quadratique (voir annexe U).

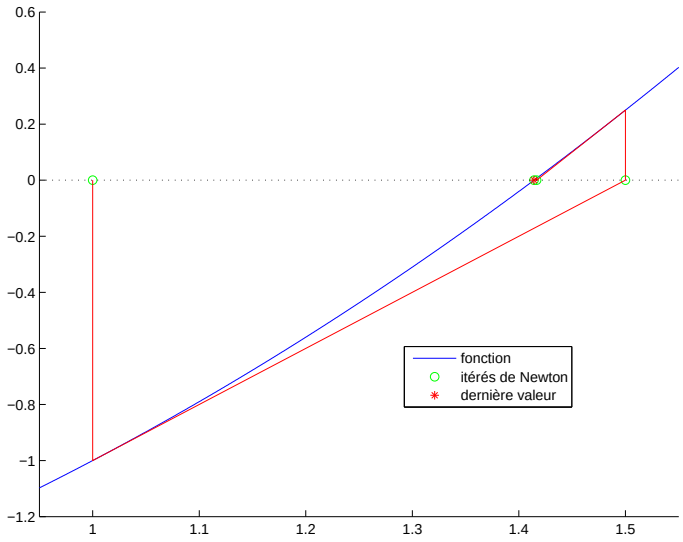


FIGURE 4.12. La méthode de Newton pour f définie par (4.98).

EXEMPLE 4.58.

On pourra de nouveau consulter l'exemple 4.48, qui correspond en fait à la méthode de newton appliquée à la recherche de l'unique zéro de

$$f(x) = x^2 - A, \tag{4.98}$$

sur $\mathbb{R}_+^*$, qui est $\sqrt{A}$, si $A > 0$. En effet,

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

Voir le graphique 4.12.

EXEMPLE 4.59. On pourra de nouveau consulter l'exemple de la section 4.1, point 3 page 77.

EXEMPLE 4.60. On pourra consulter l'annexe V page 267.

REMARQUE 4.61. On pourra consulter l'annexe W où est donné le cas d'une méthode de Newton divergente partout.

♢

On peut affaiblir les hypothèses de la proposition 4.54 :

PROPOSITION 4.62. *Si f est de classe $\mathcal{C}^2$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$ et si $f'(r) \neq 0$ et si $f''(r) \neq 0$, alors la méthode est quadratique.*

DÉMONSTRATION. On n'utilise plus la proposition 4.38 appliquée à la fonction g définie par (4.92), mais directement la définition de x_{n+1} qui donne :

$$x_{n+1} - r = \frac{f'(x_n)(x_n - r) - f(x_n)}{f'(x_n)} \tag{4.99}$$

On applique cette fois-ci la formule (4.57) à f avec $a = x_n$, $b = r$ et $p = 2$: il existe $\xi_n \in]x_n, r[$ tel que

$$f(r) = f(x_n) + f'(x_n)(r - x_n) + \frac{1}{2}f''(\xi_n)(r - x_n)^2,$$

ce qui donne, grâce à (4.99) :

$$x_{n+1} - r = \frac{-f(r) + \frac{1}{2}f''(\xi_n)(r - x_n)^2}{f'(x_n)}$$

et donc, puisque $f(r) = 0$:

$$x_{n+1} - r = \frac{1}{2}\frac{f''(\xi_n)(x_n - r)^2}{f'(x_n)}. \tag{4.100}$$

On conclue comme dans la fin de la proposition 4.38.

□

Remarque 4.63. D'après cette preuve, on peut remplacer la majoration de la remarque 4.55 par la majoration plus simple suivante :

$$C = \frac{1}{2} \max_{x\in[r-\varepsilon,r+\varepsilon]} \frac{|f''(x)|}{|f'(x)|}. \tag{4.101}$$

♢

4.5.3. Convergence (globale)

Donnons deux conditions suffisantes de convergence globale. On renvoie à l'annexe Y page 287. ♢

4.6. Méthode de la sécante (ou de Lagrange)

L'avantage de l'aspect quadratique de la méthode se paye par la connaissance et le calcul de f'. Une autre idée consiste à remplacer dans Newton le calcul de la pente $f'(x_n)$ de la tangente par son approximation

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

qui n'est autre que la sécante passant par $(x_n, f(x_n))$ et $(x_{n-1}, f(x_{n-1}))$.

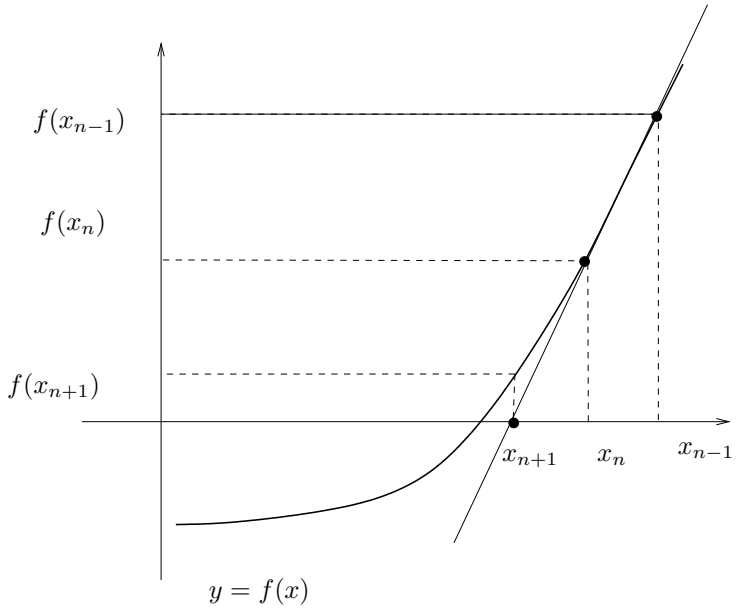


Figure 4.13. La méthode de la sécante

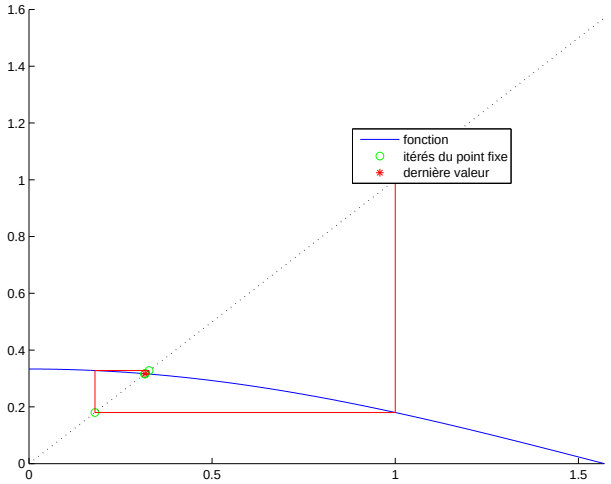
Voir figure 4.13.
Ainsi, la définition de la méthode de la sécante est

Définition 4.64 (Méthode de la Sécante). La suite $(x_n)_{n\in\mathbb{N}}$ est définie par les données de $x_0, x_1 \in \mathbb{R}$ et

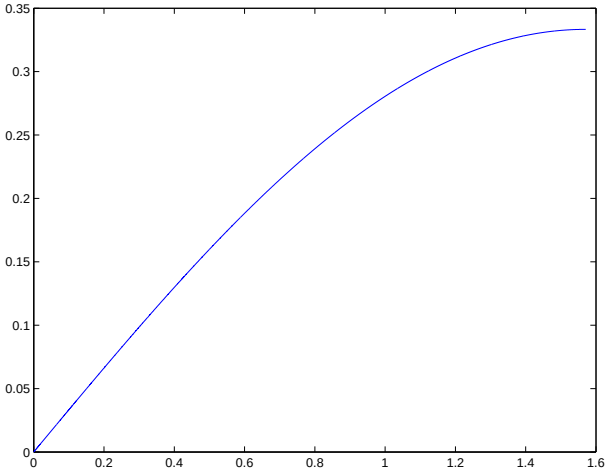
$$\forall n \in \mathbb{N}^*, \quad x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{4.102}$$

L'ordre méthode cette méthode est ni 1 ni 2, mais donné par la proposition suivante.

Proposition 4.65. *La méthode de la sécante est d'ordre* $p = (1+\sqrt{5})/2 \approx 1.61803$.



(a) Le graphe de la fonction g et les premières valeurs de la suite x_n.



(b) Le graphe de la fonction $|g'|$.

FIGURE 4.14. Les graphes des fonctions g et $|g'|$.

Cela signifie que la proposition 4.30 a lieu pour le réel p.

DÉMONSTRATION. Voir [Sch01]. □

♢

4.7. Méthode de la corde et de la fausse position

Ces méthodes, classiques, ne sont pas étudiées. On pourra consulter par exemple [BM03, Exercice 4.2].

4.8. Deux exercices types à savoir traiter parfaitement

Énoncé 1

On considère la fonction g définie par

$$\forall x \in \mathbb{R}, \quad g(x) = 1/3 \cos(x)\,. \tag{4.103}$$

et on pose

$$a = 0, \quad b = 1/2\,\pi. \tag{4.104}$$

(1) Montrer que g a un unique point fixe r sur $[a, b]$ et que la méthode du point fixe est convergente pour tout point de $x_0 \in [a, b]$ vers $r = 0.31675082877122117188679618061096$.

(2) Soit $(x_n)_{n\in\mathbb{N}}$, la suite associée à la méthode du point fixe. Déterminer l'entier n à partir duquel $|x_n - r| \leq \varepsilon$ où $\varepsilon = 10^{-3}$.

(3) Calculer les termes de la suite correspondant en choisissant $x_0 = 1$.

Corrigé 1

(1) (a) (i) On a

$$g'(x) = -1/3 \sin(x)\,. \tag{4.105}$$

(ii) Sur la figure 14(a), on constate que la fonction g semble avoir un point fixe, correspondant à la valeur
$$r = 0.31675082877122117188679618061096.$$
(iii) Sur la figure 14(b), on constate que les valeurs de la fonction $|g'|$ sont inférieures à 0.333333.
Démontrons cela rigoureusement. On majore la valeur absolue du sin par 1 et on majore donc la valeur absolue de la dérivée par 1/3.
(iv) Sur la figure 14(a), on constate que l'intervalle $[a, b]$ est g-stable.
Démontrons cela rigoureusement. La fonction g est monotone (car décroissante) ; ainsi, sur l'intervalle $[a, b]$, elle prend les valeurs comprises entre $g(a)$ et $g(b)$. On vérifie que $g(a) = 0.3333333333333$ et $g(b) = 0$ sont bien dans l'intervalle $[a, b]$.

(b) D'après les points 1(a)iii et 1(a)iv, les deux hypothèses de la proposition 4.19 sont vérifiées et donc g admet un point fixe unique r dans $I = [a, b]$ et, pour tout x_0 de I, la suite (x_n) est définie et converge vers r. Cette valeur est nécessairement celle donnée dans l'énoncé, par unicité de celle-ci !

(2) Appliquons le résutat de la proposition 4.21 ; on choisit n définit par (4.45), où la valeur de k a été donnée plus haut, ce qui donne numériquement
$$n = 7. \tag{4.106}$$

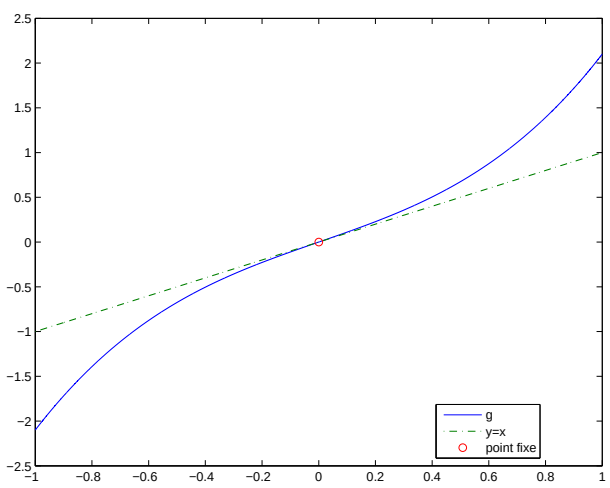
(3) On obtient alors progressivement :
$$\begin{aligned}
&x_0 = 1\,;\\
&x_1 = g(x_0) = 0.1801007686227\,;\\
&x_2 = g(x_1) = 0.3279418824098\,;\\
&x_3 = g(x_2) = 0.3155690860231\,;\\
&x_4 = g(x_3) = 0.3168733042460\,;\\
&x_5 = g(x_4) = 0.3167381101476\,;\\
&x_6 = g(x_5) = 0.3167521492808\,;\\
&x_7 = g(x_6) = 0.3167506916665.
\end{aligned}$$

Remarque 4.66. Si on calcule l'erreur réellement commise, en utilisant la valeur de x_n déterminée ci-dessous et la valeur de r donnée dans l'énoncé, on a
$$|x_n - r| = |0.3167506916665 - 0.31675082877122117188679618061096| = 0.0000001371047,$$
ce qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

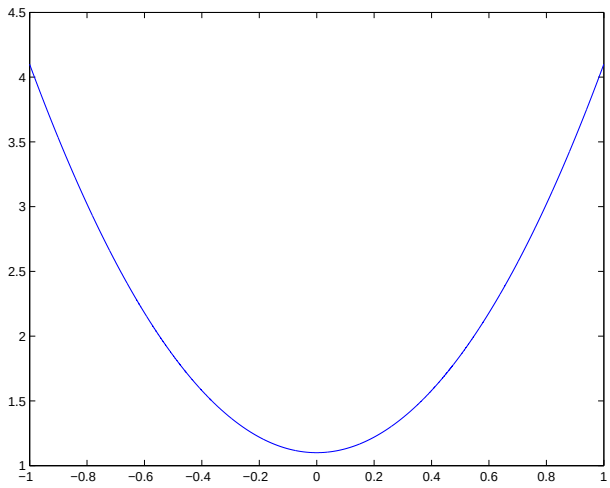
Remarque 4.67. Si on utilise la majoration donnée par (Q.13), on obtient
$$|x_n - r| \leq 0.0000007288071,$$
qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

Énoncé 2

On considère la fonction g définie par
$$\forall x \in \mathbb{R}, \quad g(x) = x^3 + \frac{11}{10}\,x. \tag{4.107}$$
et on pose
$$a = -1, \quad b = 1. \tag{4.108}$$
Montrer que la méthode du point fixe est divergente pour tout point x_0 de $[a, b] \setminus \{0\}$.



(a) Le graphe de la fonction g.



(b) Le graphe de la fonction $|g'|$.

FIGURE 4.15. Les graphes des fonctions g et $|g'|$.

Corrigé 2

(1) On a

$$g'(x) = 3\,x^2 + \frac{11}{10}. \tag{4.109}$$

Sur la figure 15(a), on constate que la fonction g semble avoir un point fixe, correspondant à la valeur

$$r = 0.$$

Sur la figure 15(b), on constate que les valeurs de la fonction $|g'|$ sont comprises entre 1.100000 et 4.100000. En particulier, en la racine r, on a

$$|g'(r)| > 1.$$

La proposition 4.11 ne peut s'appliquer ici car on ne sait pas montrer (4.25a) *a priori*.

(2) Utilisons donc la proposition 4.12 sur l'intervalle $[a, b]$.

(a) On a g définie sur I et pour tout $x \notin I$, si $g(x)$ est défini, il n'appartient pas à I Cela provient du fait suivant : Si $x \notin [-1, 1]$, $|x| \geq 1$ et $|g(x)| = |x^3 + 1.1x|$. Si par exemple, $x > 0$, on a $|g(x)| = x^3 + 1.1x \geq 1.1x > x > 1$. Il en est de même si $x < 0$. Donc, dans les deux cas, $g(x)$ est défini et n'appartient donc pas à I. Ainsi, l'hypothèse (4.32b) de la proposition 4.12 est établie.

(b) Montrons maintenant que

$$\forall x \in [a, b], \quad |g'(x)| \geq 1.1000000000000. \tag{4.110}$$

Cela provient du fait suivant : Il est évident que si $|x| \leq 1$, alors $|g'(x)| = 3x^2 + 11/10 \geq 11/10 > 1$. Ainsi, l'hypothèse (4.32c) de la proposition 4.12 est établie.

(3) On en déduit donc la divergence de la suite (x_n) pour tout $x_0 \in I \setminus \{0\}$.

Chapitre 5

Équations différentielles (ordinaires)

5.1. Motivations

5.1.1. Équations différentielles du premier ordre

La propagation d'une épidémie est modélisée, sous certaines conditions, par l'équation différentielle suivante

$$\forall t \geq 0, \quad y'(t) = ky(t)(L - y(t)), \text{ et } y(0) = y_0, \tag{5.1}$$

$y(t)$ étant le nombre d'individus infectés au temps t, L le nombre total d'individus considérés (infectés ou non) et k le coefficient de propagation lié à l'épidémie considérée. En partant de conditions de simulation de départ, L, k, y_0, on s'intéresse au nombre d'individus infectés $y(t)$, solution de l'équation (5.1), pour une gamme de temps donnée, et à la courbe donnant l'évolution du nombre d'individus infectés $y(t)$ en fonction du temps.

5.1.2. Système d'équations différentielles du premier ordre : évolution de populations (proie-prédateur)

Le modèle de Lotka-Volterra décrit les interactions entre deux espèces, une proie (les lapins) et un prédateur (les renards). Si $R(t)$ désigne le nombre de renards à un instant donné et $L(t)$ le nombre de lapins, l'évolution au cours du temps de ces populations est régie par le système différentiel

$$\frac{dL(t)}{dt} = k_L L(t) - AR(t)L(t), \tag{5.2a}$$

$$\frac{dR(t)}{dt} = -k_R R(t) + BR(t)L(t), \tag{5.2b}$$

$$L(0) = L_0, \tag{5.2c}$$

$$R(0) = R_0. \tag{5.2d}$$

où k_L et k_R sont des facteurs de croissance des deux populations et A, B des paramètres tenant compte de l'interaction entre les deux espèces. Il s'agit alors d'étudier l'évolution au cours du temps des populations en faisant varier les différents paramètres du problème.

5.1.3. Système d'équations différentielles du deuxième ordre : évolution de deux masselotes

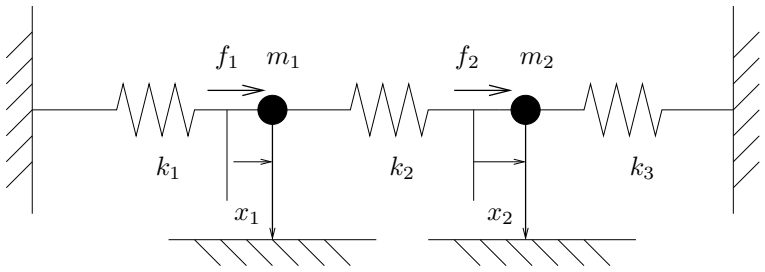


FIGURE 5.1. un système mécanique à deux degrés de liberté.

On considère le système mécanique représenté sur la figure 5.1, formé de deux points matériels de masses m_1 et m_2, d'abscisses par rapport à la position d'équilibre $x_1(t)$ et $x_2(t)$. Ces deux points matériels sont reliés à trois ressorts de raideur $k_1 > 0$, $k_2 > 0$ et $k_3 > 0$. On suppose de plus que chacun des points matériels est soumis à une force extérieure $f_i(t)$ et à une force de frottement égale à $-c_i\dot{x}_i(t) - d_i\dot{x}_i^3(t)$ où $c_i, d_i \geq 0$, pour $i = 1, 2$. Le principe fondamental de la dynamique conduit aux deux équations suivantes

$$m_1\ddot{x}_1(t) + (k_1 + k_2)\, x_1(t) - k_2 x_2(t) + c_1\dot{x}_1(t) + d_1\dot{x}_1^3(t) = f_1(t), \tag{5.3a}$$

$$m_2\ddot{x}_2(t) - k_2 x_1(t) + (k_2 + k_3)\, x_2(t) + c_2\dot{x}_2(t) + d_2\dot{x}_2^3(t) = f_2(t). \tag{5.3b}$$

On veut déterminer les abscisses et les vitesses des deux masselottes m_1 et m_2.

5.1.4. Modèles de propagation du coronavirus

La propagation du coronavirus a été, depuis peu, étudiée de très nombreuses fois, de façon beaucoup plus précise que dans la section 5.1.1. Des ouvrages généraux sur les propagations de maladies existent déjà. Le modèle le plus populaire est appelé le modèle dit SIR. La population est divisée en trois groupes S (en anglais ou en français "susceptible" d'être contaminés), I (les "infectés") et R (en anglais "recovered", c'est-à-dire ceux qui ont guéri). Le modèle consiste à décrire l'évolution de la maladie sous la forme de trois équations d'ordre 1 :

$$\frac{dS}{dt} = -\beta SI, \tag{5.4a}$$

$$\frac{dI}{dt} = \beta SI - \alpha I, \tag{5.4b}$$

$$\frac{dR}{dt} = \alpha I. \tag{5.4c}$$

Voir par exemple [BD65, page 161] ou [DHB13, page 17]. Ce modèle présente des versions plus élaborées (voir `https://interstices.info/modeliser-la-propagation-dune-epidemie/` ou `https://fr.wikipedia.org/wiki/Modèles_compartimentaux_en_épidémiologie`).

Plus récemment, divers travaux concernent l'évolution spécifique du coronavirus sur des données réelles. Voir par exemples les travaux de Dominique Sandri (voir `http://utbmjb.chez-alice.fr/Polytech/MNBif/modele_tout_coronavirus_d_sandri.pdf`) ou ceux de Nicolas Bacaër [Bac20].

5.2. Introduction et formalisme

Les problèmes de type (5.1), (5.2) et (5.3) (auquel on adjoint les valeurs des abscisses et des vitesses initiales de chacun des deux masselottes) dits problèmes aux conditions initiales. On parle aussi d'équations différentielles ordinaires [1].

Définition 5.1. Pour tout $T > 0$, pour tout $\xi_0 \in \mathbb{R}$, pour toute fonction f de $[0, T] \times \mathbb{R}$ dans $\mathbb{R}$, on considère l'équation différentielle ordinaire :

$$\forall t \in [0, T], \quad y'(t) = f(t, y(t)), \tag{5.5a}$$

avec la condition initiale

$$y(0) = \xi_0. \tag{5.5b}$$

Remarque 5.2.

(1) On a, dans (5.5a)

$$y'(t) = \frac{dy(t)}{dt}.$$

1. Par opposition aux équations aux dérivées partielles qui ne sont pas évoquées dans ce cours.

(2) Le problème (5.5) est aussi parfois appelé problème de Cauchy.

(3) t représente souvent le temps.

(4) f est (pour le moment) une fonction à 2 variables y et t supposée assez différentiable par la suite (voir plus loin).

(5) $y(0) = \xi_0$ est la condition initiale (état de la solution au moment où on commence à s'y intéresser).

(6) L'équation (5.5a) est d'ordre 1, car seule y' est présente.

(7) On peut prendre t_0 au lieu de 0, la condition initiale s'écrit alors $y(t_0) = \xi_0$ et on cherche $y(t)$ pour $t \in [t_0, t0 + T]$.

Ne connaissant pas nécessairement $y(t)$, pour $t \in [0, T]$, nous allons en déterminer une approximation.

Exemple 5.3. Le problème de Cauchy linéaire

$$\forall t \in \mathbb{R}_+, \quad y'(t) = 3y(t) - 3t, \tag{5.6a}$$

$$y(0) = 1, \tag{5.6b}$$

admet la solution $y(t) = (1 - 1/3)e^{3t} + t + 1/3$ et ce pour tout $t \geq 0$ (solution globale). On met (5.6a) sous la forme (5.5a) : pour tout t, on a

$$y'(t) = 3y(t) - 3t = f(t, y(t)),$$

et donc, pour tout t et pour tout y, on a :

$$f(t, y) = 3y - 3t. \tag{5.7}$$

Ne surtout pas écrire :

$$f(t, y) = 3y(t) - 3t.$$

Pour éviter cela, on remplace parfois (5.7) par

$$f(t, u) = 3u - 3t. \tag{5.8}$$

On a aussi

$$\xi_0 = 1. \tag{5.9}$$

Remarque 5.4. Il est fondamental de comprendre que, dans la définition 5.1, une équation différentielle est totalement définie par la donnée de f et de ξ_0. Sous matlab, par exemple, il faudrait définir f et ξ_0.

Exemple 5.5. Le problème de Cauchy non linéaire

$$\forall t \in \mathbb{R}_+, \quad y'(t) = 1 + y^2(t), \tag{5.10a}$$

$$y(0) = 0, \tag{5.10b}$$

où $f(t, y) = 1 + y^2$, admet la solution $y(t) = \tan(t)$ et ce pour tout $t \in [0, \pi/2[$. (solution locale).

Exemple 5.6. Le problème de Cauchy non linéaire

$$\forall t \in \mathbb{R}_+, \quad y'(t) = (y(t))^{1/3}, \tag{5.11a}$$

$$y(0) = 0, \tag{5.11b}$$

où $f(t, y) = (y)^{1/3}$, admet les trois solutions : $y(t) = 0$, $y(t) = \sqrt{8t^3/27}$ et $y(t) = -\sqrt{8t^3/27}$. On n'a donc pas d'unicité.

5.3. Un peu de théorie

De même que l'on s'intéresse à l'existence et l'unicité de la solution de l'équation du point fixe (4.23), on a aussi de tels résultats pour l'équation différentielle (5.5).

Pour les preuves des théorèmes, le lecteur pourra consulter [Bas78 ; CM84 ; RDO87 ; Sch01 ; QSS00]. ♢

Définition 5.7. On dit que f est lipschitzienne par rapport à sa deuxième variable, uniformément en la première, s'il existe $L \in \mathbb{R}_+$ tel que :

$$\forall t \in [0,T], \quad \forall y_1, y_2 \in \mathbb{R}, \quad |f(t,y_1) - f(t,y_2)| \leq L\,|y_1 - y_2|\,. \tag{5.12}$$

Théorème 5.8 (Cauchy-Lipschitz). *Si f est continue sur $[0,T]\times\mathbb{R}$ et vérifie* (5.12)*, alors il existe une unique solution de* (5.5) *de classe $\mathcal{C}^1$ sur $[0,T]$.*

Remarque 5.9. Parfois la condition (5.12) est trop forte ; la fonction f est seulement localement lipschitzienne, c'est-à-dire

$$\forall \tau \in [0,T], \quad \forall y_1 \in \mathbb{R}, \quad \exists (r,L) \in \mathbb{R}_+^* \times \mathbb{R}_+ : \quad \forall t \in [0,T], \quad \forall y_2 \in \mathbb{R},$$
$$\Big(|\tau - t| \leq r \text{ et } |y_1 - y_2| \leq r \Longrightarrow |f(t,y_1) - f(t,y_2)| \leq |Ly_1 - y_2|\Big). \tag{5.13}$$

Dans ce cas, seule l'existence locale de la solution de (5.5) est assurée :

Théorème 5.10. *Si f est continue sur $[0]\times\mathbb{R}$ et vérifie* (5.13)*, alors il existe un voisinage V de t_0 et une unique solution y de* (5.5) *de classe $\mathcal{C}^1$ sur V.*

♢

Remarque 5.11. Pour toute la suite, nous supposons assurées l'existence et l'unicité de la solution sur $[0,T]$.

Tous ces résultats s'étendent aussi au cas vectoriel (voir section 5.6). ♢

5.4. Schémas d'Euler progressif et rétrograde

Présentons la méthode de discrétisation la plus simple, celle d'Euler.

Pour $N \in \mathbb{N}^*$, on découpe l'intervalle $[0,T]$ en N sous-intervalles de taille h et on définis les points t_n équirépartis, comme dans la définition 2.26, qui devient donc ici :

Définition 5.12. Pour pour $T > 0$ et pour tout $N \in \mathbb{N}^*$, on pose

$$h = \frac{T}{N}, \tag{5.14a}$$
$$\forall n \in \{0,...,N\}, \quad t_n = hn. \tag{5.14b}$$

Nous allons déterminer les N réels , $(y_n)_{0\leq n\leq N}$, tels que, chaque réels y_n constitue une approximation de $y(t_n)$:

$$\forall n \in \{0,...,N\}, \quad y_n \approx y(t_n). \tag{5.15}$$

La première valeur de y étant donnée par (5.5b), on posera donc

$$y_0 = \xi_0. \tag{5.16}$$

Ensuite, on applique (5.5a) à t_n, pour $n \in \{0,...,N-1\}$:

$$\forall n \in \{0,...,N-1\}, \quad y'(t_n) = f(t_n, y(t_n)). \tag{5.17}$$

On approche la valeur de $y'(t_n)$ par le taux d'accroissement entre t_0 et t_{n+1} :

$$y'(t_n) \approx \frac{y(t_{n+1}) - y(t_n)}{t_{n+1} - t_n},$$

soit, compte tenu de (5.15) :

$$y'(t_n) \approx \frac{y_{n+1} - y_n}{h}. \tag{5.18}$$

Compte tenu de (5.15), $f(t_n, y(t_n))$ est approchée par

$$f(t_n, y(t_n)) \approx f(t_n, y_n). \tag{5.19}$$

Bref, compte tenu de (5.18) et de (5.19), on remplace (5.17) par

$$\forall n \in \{0, ..., N-1\}, \quad \frac{y_{n+1} - y_n}{h} = f(t_n, y_n), \tag{5.20}$$

ce qui est équivalent à

$$\forall n \in \{0, ..., N-1\}, \quad y_{n+1} = y_n + hf(t_n, y_n). \tag{5.21}$$

Cette équation est appelée le schéma d'Euler progressif (ou explicite). Ce schéma est dit à un pas : la connaissance de y_n, permet de déterminer y_{n+1}. Ainsi, y_0 étant défini par (5.16), on utilise (5.21) avec $n = 0$ ce qui fournit

$$y_1 = y_0 + hf(t_0, y_0), \tag{5.22}$$

et donc la valeur de y_1. Ensuite, connaissant cette valeur y_1, on utilise (5.21) avec $n = 1$ ce qui fournit

$$y_2 = y_1 + hf(t_1, y_1), \tag{5.23}$$

et donc la valeur de y_2. On calcule donc successivement y_3, ... jusqu'à y_N.

On donne donc la définition :

Définition 5.13 (Schéma numérique d'Euler explicite (ou progressif)). Avec les notations des définitions 5.1 et 5.12, les approximations du schéma numérique d'Euler explicite (ou progressif), $(y_n)_{0 \leq n \leq N}$ sont données par

$$y_0 = \xi_0, \tag{5.24a}$$

$$\forall n \in \{0, ..., N-1\}, \quad y_{n+1} = y_n + hf(t_n, y_n). \tag{5.24b}$$

Si on remplace (5.19) par $f(t_n, y(t_n)) \approx f(t_{n+1}, y_{n+1})$, on obtient :

Définition 5.14 (Schéma numérique d'Euler implicite (ou rétrograde)). Avec les notations des définitions 5.1 et 5.12, les approximations du schéma numérique d'Euler implicite (ou rétrograde), $(y_n)_{0 \leq n \leq N}$ sont données par

$$y_0 = \xi_0, \tag{5.25a}$$

et, pour tout $n \in \{0, ..., N-1\}$, y_{n+1} vérifie l'équation (non linéaire)

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}). \tag{5.25b}$$

L'ensemble de ces deux méthodes (Euler progressif et retrograde) est le plus simple. Ces deux méthodes sont aussi les moins précises de toutes.

Comme pour la méthode d'intégration des rectangles, on obtient la majoration d'erreur suivante :

Théorème 5.15. *Si f est assez régulière, alors, il existe une constante M telle que, si y est la solution de* (5.5) *et si les y_n sont donnés par la définition 5.13,*

$$\forall N \in \mathbb{N}^*, \quad \forall n \in \{0, ..., N\}, \quad |y(t_n) - y_n| \leq Mh. \tag{5.26}$$

Comme dans la définition 3.35, nous dirons que la méthode d'Euler explicite est d'ordre un.

Les hypothèses de régularité de f et la preuve de ce résultat figurent par exemple dans [BM03, Théorème 5.17]. ♢

On a un résultat analogue pour le schéma d'Euler implicite.

EXEMPLE 5.16. On considère l'équation différentielle (5.5), où

$$\forall t \in [0,T] \times \mathbb{R}, \quad f(t,y) = (-1/10\, t + \sin(t))\, y, \tag{5.27}$$

que l'on résout avec le schéma d'Euler progressif en prenant différentes valeurs de N. On choisit

$$T = 12, \tag{5.28a}$$

$$\xi_0 = 1. \tag{5.28b}$$

On peut aussi déterminer la solution exacte de (5.27)-(5.28b) :

$$\forall t \in [0,T], \quad y(t) = e \times e^{\left(-\frac{1}{20}t^2 - \cos(t)\right)}. \tag{5.29}$$

Voir la figure 5.2 sur laquelle on constate que, plus h est petit, plus la solution déterminée avec le schéma

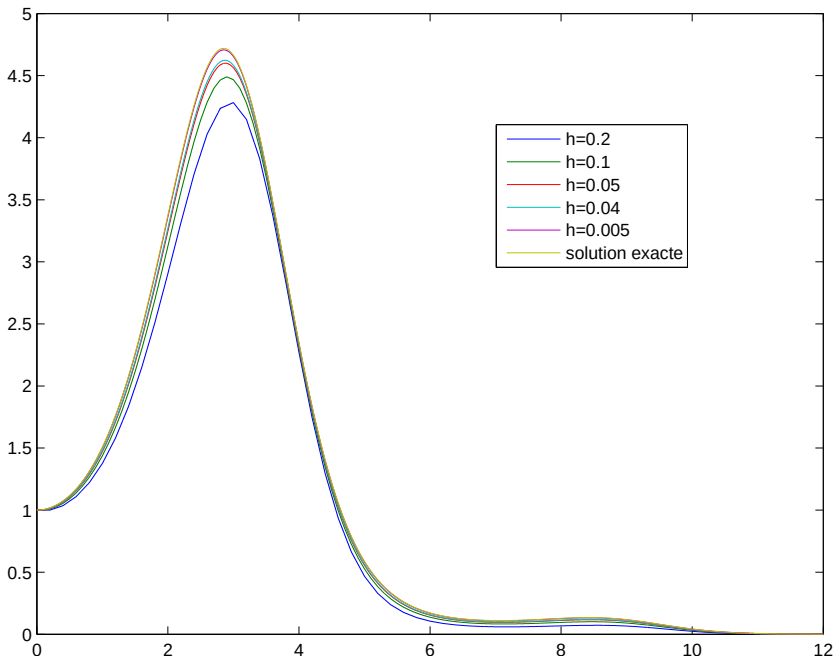


FIGURE 5.2. Solution exacte et quelques solutions approchées de l'équation différentielle (5.27) pour différentes valeurs de h.

d'Euler est proche de la solution exacte.

Le schéma d'Euler implicite se comporte parfois mieux que son homologue explicite comme le montre l'exemple suivant :

EXEMPLE 5.17. On considère l'équation différentielle (5.5), où

$$\forall t \in [0,T] \times \mathbb{R}, \quad f(t,y) = -\lambda t \tag{5.30}$$

On choisit

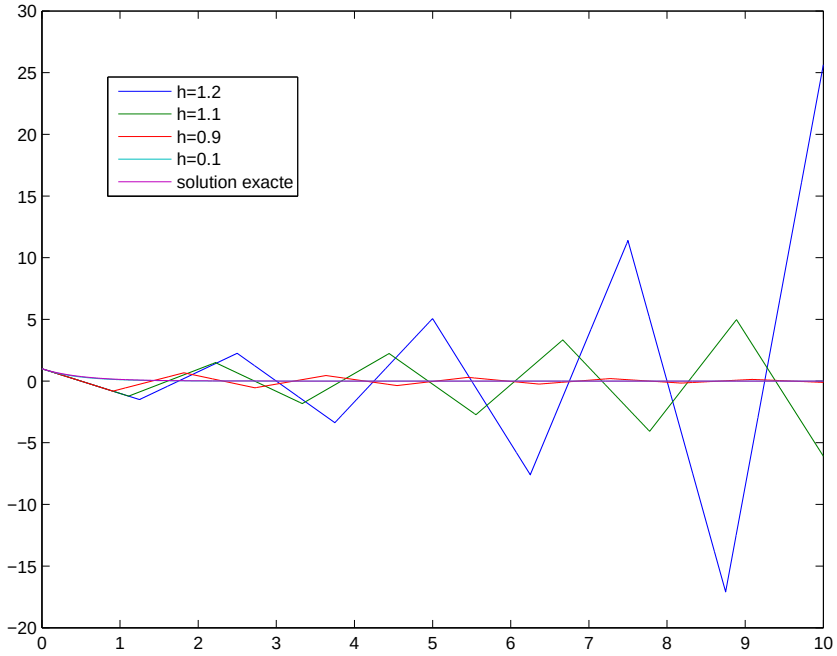
$$\lambda = 2, \tag{5.31a}$$

$$T = 10, \tag{5.31b}$$

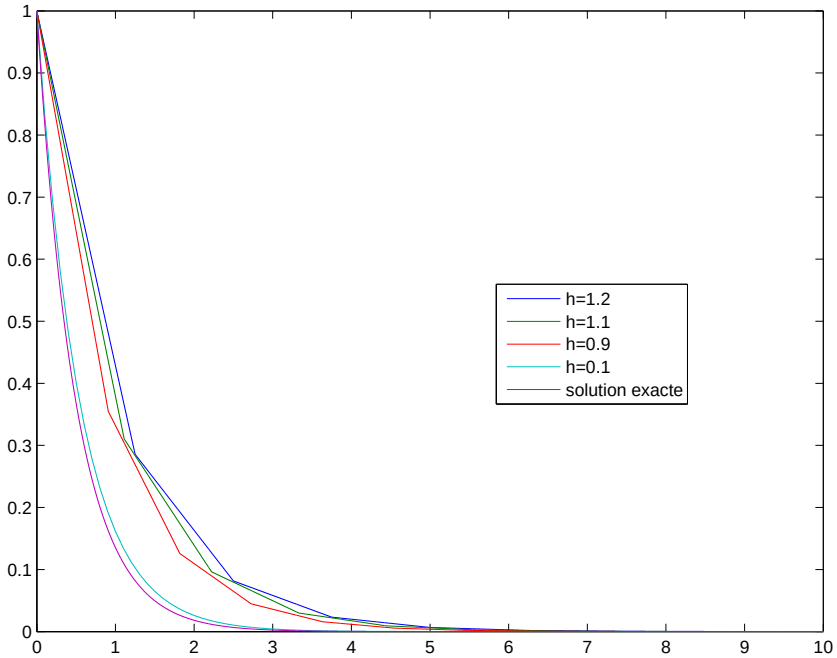
$$\xi_0 = 1. \tag{5.31c}$$

On peut aussi déterminer la solution exacte de (5.5)-(5.30) :

$$\forall t \in [0,T], \quad y(t) = \xi_0 e^{-\lambda t}. \tag{5.32}$$



(a) Explicite



(b) Implicite

FIGURE 5.3. Résolution de (5.5)-(5.30) avec la méthode d'Euler explicite et d'Euler implicite pour plusieurs valeurs de h.

Sur la figure 3(a), on constate que si h est trop grand, la solution "oscille" ce qui n'est pas le cas de la figure 3(b).

5.5. Schémas de Runge Kutta 2 et 4

De même que la méthode composite des rectangles, d'ordre 1, peut être améliorée en utilisant par exemple les méthodes des milieux ou de Simpson, d'ordre 2 et 4, le schéma d'Euler peut être améliorée. Citons, sans preuve, les schémas de Runge Kutta 2 et 4 :

Définition 5.18 (Schéma numérique de Runge-Kutta 2). Avec les notations des définitions 5.1 et 5.12, les approximations du schéma numérique de Runge-Kutta 2, $(y_n)_{0\leq n\leq N}$ sont données par

$$y_0 = \xi_0, \tag{5.33a}$$

$$\forall n \in \{0, ..., N-1\}, \quad \begin{cases} k_n^{(1)} = hf(t_n, y_n), \\ k_n^{(2)} = hf(t_n + h, y_n + k_n^{(1)}), \\ y_{n+1} = y_n + \frac{1}{2}(k_n^{(1)} + k_n^{(2)}). \end{cases} \tag{5.33b}$$

Définition 5.19 (Schéma numérique de Runge-Kutta 4). Avec les notations des définitions 5.1 et 5.12, les approximations du schéma numérique de Runge-Kutta 4, $(y_n)_{0\leq n\leq N}$ sont données par

$$y_0 = \xi_0, \tag{5.34a}$$

$$\forall n \in \{0, ..., N-1\}, \quad \begin{cases} k_n^{(1)} = hf(t_n, y_n), \\ k_n^{(2)} = hf\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}k_n^{(1)}\right), \\ k_n^{(3)} = hf\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}k_n^{(2)}\right), \\ k_n^{(4)} = hf\left(t_n + h, y_n + k_n^{(3)}\right), \\ y_{n+1} = y_n + \frac{1}{6}(k_n^{(1)} + 2k_n^{(2)} + 2k_n^{(3)} + k_n^{(4)}). \end{cases} \tag{5.34b}$$

On a alors

Théorème 5.20. *Si f est assez régulière, alors, il existe une constante M telle que, si y est la solution de (5.5) et si les y_n sont donnés par la définition 5.18,*

$$\forall N \in \mathbb{N}^*, \quad \forall n \in \{0, ..., N\}, \quad |y(t_n) - y_n| \leq Mh^2. \tag{5.35}$$

Nous dirons que le schéma Runge-Kutta 2 est d'ordre deux.

Théorème 5.21. *Si f est assez régulière, alors, il existe une constante M telle que, si y est la solution de (5.5) et si les y_n sont donnés par la définition 5.19,*

$$\forall N \in \mathbb{N}^*, \quad \forall n \in \{0, ..., N\}, \quad |y(t_n) - y_n| \leq Mh^4. \tag{5.36}$$

Nous dirons que le schéma Runge-Kutta 4 est d'ordre quatre.
Voir les preuves par exemple dans [BM03, section 5.4] ou [Sch01].

Exemple 5.22. On reprend l'exemple 5.16. Voir la figure 5.4 sur laquelle on a tracé les trois approximations : Euler, Runge-Kutta 2 et 4 pour $h = 0.20000$.

Remarque 5.23. Comme pour les méthodes d'intégration (voir remarque 3.38 page 70 et exemple 3.28), on peut aussi tracer un diagramme log-log qui met les ordre 1, 2 et 4 des trois méthodes vues en évidence. Voir l'exemple 5.24.

Exemple 5.24. On reprend l'exemple 5.16. Voir la figure 5.5 et le tableau 5.1.

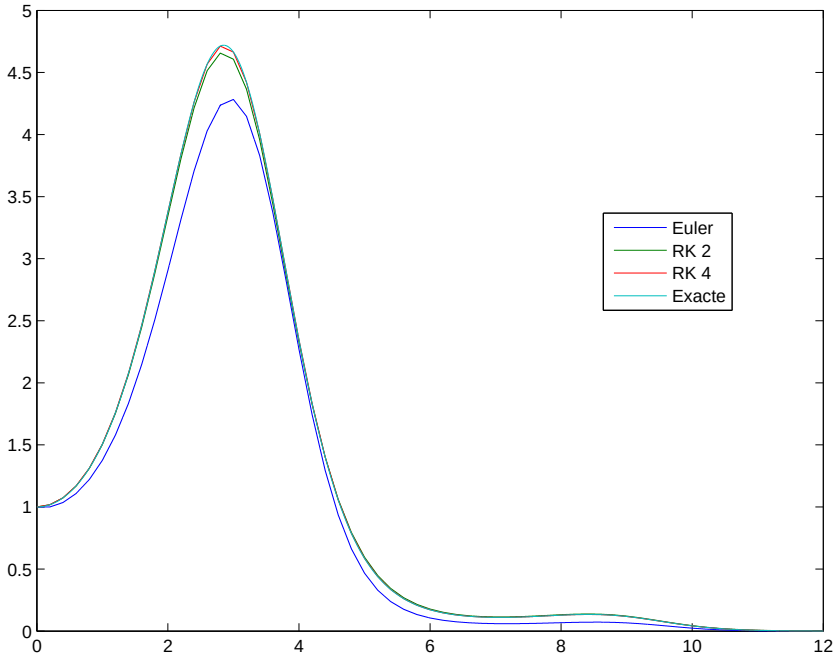


FIGURE 5.4. Solution exacte et trois approximations : Euler, Runge-Kutta 2 et 4.

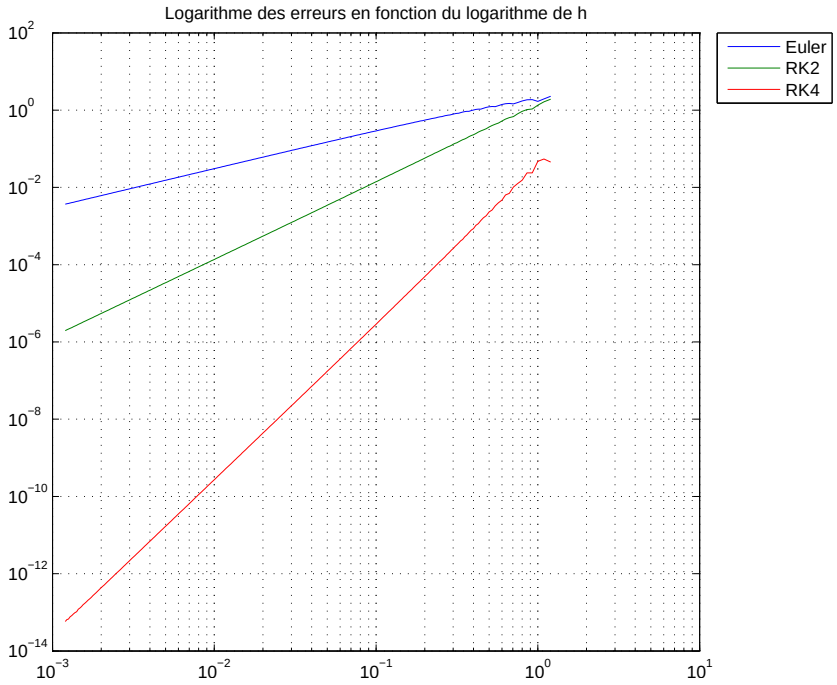


FIGURE 5.5. Le graphe log log de l'erreur pour les trois méthodes d'Euler, Runge-Kutta 2 et 4.

	exposant théorique	pente expérimentale
Euler	1	0.9605458
RK2	2	2.0094599
RK4	4	4.0397775

TABLE 5.1. Pentes théoriques pour les données de l'exemple 5.16

5.6. Équations différentielles d'ordres plus élevés

5.6.1. Forme générale

On remplace la définition 5.1 par

Définition 5.25. Soit $p \in \mathbb{N}^*$. Pour tout $T > 0$, pour tout $\Xi_0 \in \mathbb{R}^p$, pour toute fonction F de $[0,T] \times \mathbb{R}^p$ dans $\mathbb{R}^p$, on considère l'équation différentielle ordinaire :

$$\forall t \in [0,T], \quad Y'(t) = F(t, Y(t)), \tag{5.37a}$$

avec la condition initiale

$$Y(0) = \Xi_0. \tag{5.37b}$$

Ici, Y est un vecteur à p composantes.

Remarque 5.26. Comme dans la remarque 5.4, il est fondamental de comprendre que, dans la définition 5.25, une équation différentielle est totalement définie par la donnée de F et de Ξ_0.

Tous les résultats de la section 5.3 s'étendent au cas vectoriel, à condition de remplacer les valeurs absolues par des normes sur $\mathbb{R}^p$ (voir par exemple [BM03, Section A.3].).

Définition 5.27. On dit que F est lipschitzienne par rapport à sa deuxième variable, uniformément en la première, s'il existe $L \in \mathbb{R}_+$ tel que :

$$\forall t \in [0,T], \quad \forall Y_1, Y_2 \in \mathbb{R}^p, \quad \|F(t,Y_1) - F(t,Y_2)\| \leq L \|Y_1 - Y_2\| . \tag{5.38}$$

Théorème 5.28 (Cauchy-Lipschitz). *Si F est continue sur $[0,T] \times \mathbb{R}^p$ et vérifie (5.38), alors il existe une unique solution de (5.37) de classe $\mathcal{C}^1$ sur $[0,T]$, c'est-à-dire, que chacune des composantes de Y est de classe $\mathcal{C}^1$.*

Remarque 5.29. Parfois la condition (5.12) est trop forte ; la fonction F est seulement localement lipschitzienne, c'est-à-dire

$$\forall \tau \in [0,T], \quad \forall Y_1 \in \mathbb{R}^p, \quad \exists (r, L) \in \mathbb{R}_+^* \times \mathbb{R}_+ : \quad \forall t \in [0,T], \quad \forall Y_2 \in \mathbb{R}^n,$$
$$\Big(|\tau - t| \leq r \text{ et } \|Y_1 - Y_2\| \leq r \Longrightarrow \|F(t,Y_1) - F(t,Y_2)\| \leq \|LY_1 - Y_2\| \Big). \tag{5.39}$$

Dans ce cas, seule l'existence locale de la solution de (5.37) est assurée :

Théorème 5.30. *Si F est continue sur $[0] \times \mathbb{R}^p$ et vérifie (5.39), alors il existe un voisinage V de t_0 et une unique solution y de (5.37) de classe $\mathcal{C}^1$ sur V.*

♢

Les résultats de cette section nous permettent d'obtenir plusieurs types de résultat.

5.6.2. Systèmes différentiels

Commençons par traiter un exemple.

Exemple 5.31. On considère les deux équations différentielles d'ordre 1 couplées suivantes : pour tout $t \in [0,T]$:

$$y_1'(t) = y_1(t) + y_2^2(t) + \cos(t), \tag{5.40a}$$
$$y_2'(t) = y_1^2(t) + y_2^3(t) + \sin(t), \tag{5.40b}$$

vérifiant les deux conditions initiales

$$y_1(0) = 2, \tag{5.40c}$$
$$y_2(0) = -3, \tag{5.40d}$$

On cherche une fonction Y de $[0,T]$ dans $\mathbb{R}^2$:

$$\forall t \in [0,T], \quad Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} \tag{5.41}$$

et une équation différentielle de la forme (5.37) avec $p = 2$ vérifiée par la fonction Y. On a, d'après (5.40)

$$Y'(t) = \begin{pmatrix} y_1'(t) \\ y_2'(t) \end{pmatrix} = \begin{pmatrix} y_1(t) + y_2^2(t) + \cos(t) \\ y_1^2(t) + y_2^3(t) + \sin(t) \end{pmatrix}.$$

On a donc

$$Y'(t) = F(t, Y(t)),$$

en posant

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{5.42}$$

et

$$\forall t \in [0, T], \quad \forall Y \in \mathbb{R}^2, \quad F(t, Y) = \begin{pmatrix} y_1 + y_2^2 + \cos(t) \\ y_1^2 + y_2^3 + \sin(t) \end{pmatrix}. \tag{5.43}$$

Enfin, les conditions initiales (5.40c) et (5.40d) peuvent s'écrire

$$Y(0) = \Xi_0, \tag{5.44}$$

où

$$\Xi_0 = \begin{pmatrix} 2 \\ -3 \end{pmatrix} \tag{5.45}$$

On a donc mis le problème (5.40) sous la forme (5.37) avec $p = 2$.

De façon plus générale, on considère les p équations différentielles couplées : pour tout $t \in [0, T]$, on a

$$y_1'(t) = f_1(t, y_1(t), ..., y_p(t)), \tag{5.46a}$$
$$y_2'(t) = f_2(t, y_1(t), ..., y_p(t)), \tag{5.46b}$$
$$\vdots \tag{5.46c}$$
$$y_p'(t) = f_p(t, y_1(t), ..., y_p(t)), \tag{5.46d}$$

vérifiant les p conditions initiales

$$y_1(0) = y_{1,0}, \tag{5.47a}$$
$$y_2(0) = y_{2,0}, \tag{5.47b}$$
$$\vdots \tag{5.47c}$$
$$y_p(0) = y_{p,0}. \tag{5.47d}$$

On cherche une fonction Y de $[0, T]$ dans $\mathbb{R}^p$:

$$\forall t \in [0, T], \quad Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{pmatrix} \tag{5.48}$$

et une équation différentielle de la forme (5.37) avec $p \in \mathbb{N}^*$ vérifiée par la fonction Y. Il suffit d'écrire, comme dans l'exemple 5.31 :

$$Y'(t) = \begin{pmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_p'(t) \end{pmatrix}.$$

On a donc

$$Y'(t) = F(t, Y(t)),$$

en posant

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \tag{5.49}$$

et

$$\forall t \in [0,T], \quad \forall Y \in \mathbb{R}^p, \quad F(t,Y) = \begin{pmatrix} f_1(t,y_1,...,y_p) \\ f_2(t,y_1,...,y_p) \\ \vdots \\ f_p(t,y_1,...,y_p) \end{pmatrix}. \tag{5.50}$$

Enfin, les conditions initiales (5.47) peuvent s'écrire

$$Y(0) = \Xi_0, \tag{5.51}$$

où

$$\Xi_0 = \begin{pmatrix} y_{1,0} \\ y_{2,0} \\ \vdots \\ y_{p,0} \end{pmatrix} \tag{5.52}$$

On a donc mis le problème (5.46) sous la forme (5.37) avec $p \in \mathbb{N}^*$.

5.6.3. Équations différentielles d'ordre $p \in \mathbb{N}^*$

Commençons par traiter un exemple.

Exemple 5.32. On cherche y de $[0,T]$ dans $\mathbb{R}$ vérifiant l'équation différentielle d'ordre 3

$$\forall t \in [0,T], \quad y^{(3)}(t) = (y''(t))^2 + 2y'(t) + (y(t))^3 + t^4 + 1, \tag{5.53}$$

avec les conditions initiales

$$y(0) = 1, \quad y'(0) = 0, \quad y''(0) = 3. \tag{5.54}$$

Cette équation se ramène à un système de 3 équations différentielles d'ordre 1, en posant

$$y_1(t) = y(t) \tag{5.55a}$$
$$y_2(t) = y'(t) \tag{5.55b}$$
$$y_3(t) = y''(t) \tag{5.55c}$$

D'où le système,

$$y_1'(t) = y_2(t), \tag{5.56a}$$
$$y_2'(t) = y_3(t), \tag{5.56b}$$
$$y_3'(t) = (y_3(t))^2 + 2y_2(t) + (y_1(t))^3 + t^4 + 1, \tag{5.56c}$$
$$y_1(0) = 1, \tag{5.56d}$$
$$y_2(0) = 0, \tag{5.56e}$$
$$y_3(0) = 3. \tag{5.56f}$$

On peut alors utiliser les résultats de la section 5.6.2 pour transformer (5.56) sous la forme (5.37) avec $p = 3$. Il suffit, en effet, de poser

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \tag{5.57}$$

On a donc

$$\forall t \in [0,T], \quad \forall Y \in \mathbb{R}^3, \quad F(t,Y) = \begin{pmatrix} y_2 \\ y_3 \\ y_3^2 + 2y_2 + y_1^3 + t^4 + 1 \end{pmatrix}, \tag{5.58}$$

et

$$Y(0) = \Xi_0, \tag{5.59}$$

où

$$\Xi_0 = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} \tag{5.60}$$

Attention, ne pas confondre entre l'ordre du système (nombre des équations du système, p en général !) et l'ordre des équations différentielles du système (ordre 1 !) !

De façon plus générale, on considère l'équation différentielle d'ordre $p \in \mathbb{N}^*$

$$\forall t \in [0,T], \quad y^{(p)}(t) = f(t, y(t), y'(t), ..., y^{(p-1)}(t)), \tag{5.61}$$

vérifiant les p conditions initiales

$$y(0) = y_{1,0}, \tag{5.62a}$$
$$y'(0) = y_{2,0}, \tag{5.62b}$$
$$y''(0) = y_{3,0}, \tag{5.62c}$$
$$\vdots \tag{5.62d}$$
$$y^{(p-1)}(0) = y_{p-1,0}. \tag{5.62e}$$

On considère les p fonctions y_i de $[0,T]$ dans $\mathbb{R}$ définies par : pour tout $t \in [0,T]$

$$y_1(t) = y(t), \tag{5.63a}$$
$$y_2(t) = y'(t), \tag{5.63b}$$
$$y_3(t) = y''(t), \tag{5.63c}$$
$$\vdots \tag{5.63d}$$
$$y_{p-1}(t) = y^{(p-2)}(t) \tag{5.63e}$$
$$y_p(t) = y^{(p-1)}(t) \tag{5.63f}$$

de telle sorte que l'on a

$$\begin{aligned}
&y_1'(t) = y'(t) = y_2(t), \\
&y_2'(t) = y''(t) = y_3(t), \\
&y_3(t) = y'''(t) = y_4(t), \\
&\vdots \\
&y_{p-1}'(t) = y^{(p-1))}(t) = y_p(t), \\
&y_p'(t) = f(t, y(t), y'(t), ..., y^{(p-1)}(t)) = f(t, y_1(t), y_2(t), ..., y_p(t)),
\end{aligned}$$

et, en utilisant les résultats de la section 5.6.2, on obtient donc une équation différentielle sous la forme (5.37). Posons en effet

$$\forall t \in [0,T], \quad Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ \vdots \\ y_p(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ y'(t) \\ y''(t) \\ \vdots \\ y^{(p-1)}(t) \end{pmatrix} \tag{5.64}$$

On a donc une équation différentielle de la forme (5.37) avec $p \in \mathbb{N}^*$ vérifiée par la fonction Y. Il suffit d'écrire :

$$\forall t \in [0,T], \quad Y'(t) = \begin{pmatrix} y_2(t) \\ y_3(t) \\ y_4(t) \\ \vdots \\ y_p(t) \\ f(t, y_1(t), y_2(t), ..., y_p(t)), \end{pmatrix} = F(t, Y(t))$$

où

$$\forall t \in [0,T], \quad \forall Y \in \mathbb{R}^p, \quad F(t,Y) = \begin{pmatrix} y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_p \\ f(t, y_1, y_2, ..., y_p), \end{pmatrix} \tag{5.65}$$

Les conditions initiales sont $Y(0) = \Xi_0$ avec

$$\Xi_0 = \begin{pmatrix} y(0) \\ y'(0) \\ y''(0) \\ \vdots \\ y^{(p-1)}(0) \end{pmatrix} \tag{5.66}$$

5.6.4. Systèmes d'équations différentielles d'ordres $p_i \in \mathbb{N}^*$

On renvoie à l'exercice de TD 5.4.

5.6.5. Existence, Unicité et Schémas numériques

Grâce aux résultats des sections (5.6.2) (5.6.3) et (5.6.4), nous disposons du formalisme des équations (5.37) avec $p \in \mathbb{N}^*$. S'appliquent alors les résultats d'existence et d'unicité 5.28 et 5.30. Mais, surtout, on peut utiliser les schémas (5.24), (5.25), (5.33) et (5.34) à condition de considérer F à valeurs dans $\mathbb{R}^p$. De même, dans ces schémas, ξ_0 et y_n sont remplacés par leurs homologues vectoriels Ξ_0 et Y_n, à valeurs dans $\mathbb{R}^p$, ce qui fournirait les approximations

$$Y_n = \begin{pmatrix} y_{n,1} \\ y_{n,2} \\ \vdots \\ y_{n,p} \end{pmatrix} \approx \begin{pmatrix} y_1(t_n) \\ y_2(t_n) \\ \vdots \\ y_p(t_n) \end{pmatrix}$$

Plus précisément, on a :

Définition 5.33 (Schéma numérique d'Euler explicite (vectoriel)). Avec les notations des définitions 5.25 et 5.12, les approximations du schéma numérique d'Euler explicite (ou progressif), $(Y_n)_{0\leq n\leq N}$ sont données par

$$Y_0 = \Xi_0, \tag{5.67a}$$
$$\forall n \in \{0, ..., N-1\}, \quad Y_{n+1} = Y_n + hF(t_n, Y_n). \tag{5.67b}$$

Définition 5.34 (Schéma numérique de Runge-Kutta 2 (vectoriel)). Avec les notations des définitions 5.25 et 5.12, les approximations du schéma numérique de Runge-Kutta 2, $(Y_n)_{0\leq n\leq N}$ sont données par

$$Y_0 = \Xi_0, \tag{5.68a}$$
$$\forall n \in \{0, ..., N-1\}, \quad \begin{cases} K_n^{(1)} = hF(t_n, Y_n), \\ K_n^{(2)} = hF(t_n + h, Y_n + K_n^{(1)}), \\ Y_{n+1} = Y_n + \frac{1}{2}(K_n^{(1)} + K_n^{(2)}). \end{cases} \tag{5.68b}$$

Définition 5.35 (Schéma numérique de Runge-Kutta 4 (vectoriel)). Avec les notations des définitions 5.25 et 5.12, les approximations du schéma numérique de Runge-Kutta 4, $(Y_n)_{0\leq n\leq N}$ sont données par

$$Y_0 = \Xi_0, \tag{5.69a}$$
$$\forall n \in \{0, ..., N-1\}, \quad \begin{cases} K_n^{(1)} = hF(t_n, Y_n), \\ K_n^{(2)} = hF\left(t_n + \frac{1}{2}h, Y_n + \frac{1}{2}K_n^{(1)}\right), \\ K_n^{(3)} = hF\left(t_n + \frac{1}{2}h, Y_n + \frac{1}{2}K_n^{(2)}\right), \\ K_n^{(4)} = hF\left(t_n + h, Y_n + K_n^{(3)}\right), \\ Y_{n+1} = Y_n + \frac{1}{6}(K_n^{(1)} + 2K^{(2)} + 2K_n^{(3)} + K_n^{(4)}). \end{cases} \tag{5.69b}$$

♢

5.7. Retours sur les exemples introductifs

5.7.1. Exemple de la section 5.1.1

Simulations numériques en cours de rédaction

5.7.2. Exemple de la section 5.1.2

Le système de deux équations (5.2) se met sous la forme (5.37) avec $p = 2$, en posant

$$Y(t) = \begin{pmatrix} L(t) \\ R(t) \end{pmatrix}$$

et en considérant F définie par

$$\forall t \in [0, T], \quad \forall Y = \begin{pmatrix} L \\ R \end{pmatrix} \in \mathbb{R}^2, \quad F(t, Y) = \begin{pmatrix} k_L L - ARL, \\ -k_R R + BRL \end{pmatrix}$$

et

$$\Xi_0 = \begin{pmatrix} L_0 \\ R_0 \end{pmatrix}$$

Simulations numériques en cours de rédaction

5.7.3. Exemple de la section 5.1.3

On renvoie à l'exercice de TD 5.4.

5.8. Un exercice type à savoir traiter parfaitement

Énoncé

On étudie l'équation différentielle

$$\forall t \in [0,T], \quad y'(t) = \cos(t) + 1/10\, y(t)^2, \tag{5.70a}$$

$$y(0) = 1, \tag{5.70b}$$

avec $T = 1$. On pose, pour $N \in \mathbb{N}^*$, $h = T/N$ et, pour tout $n \in \{0, ..., N\}$, $t_n = hn$. On choisit $N = 5$. Déterminer y_n, les approximations de $y(t_n)$, pour $n \in \{0, ..., 5\}$, avec le schéma d'Euler progressif (dit aussi d'Euler explicite) , de Runge-Kutta d'ordre 2 et de Runge-Kutta d'ordre 4 pour le problème (5.70).

Corrigé

En posant $\xi_0 = 1$ et

$$f(t,y) = \cos(t) + 1/10\, y^2, \tag{5.71}$$

l'équation différentielle

$$\forall t \in [0,T], \quad y'(t) = \cos(t) + 1/10\, y(t)^2, \tag{5.72a}$$

$$y(0) = 1, \tag{5.72b}$$

est équivalente à

$$\forall t \in [0,T], \quad y'(t) = f(t, y(t)), \tag{5.73a}$$

$$y(0) = \xi_0. \tag{5.73b}$$

On calcule pour $n \in \{1, ..., 5\}$, les approximations $y_n \approx y(t_n)$.

n	y_n
0	1.00000000
1	1.22000000
2	1.44578132
3	1.67179919
4	1.89276456
5	2.10375706

TABLE 5.2. Solutions approchées avec Euler explicite

Les résultats sont donnés dans les tableaux 5.2, 5.3 et 5.4, obtenus en utilisant les définitions 5.13, 5.18 et 5.19.

n	y_n
0	1.00000000
1	1.22289066
2	1.44894863
3	1.67264408
4	1.88868540
5	2.09213115

TABLE 5.3. Solutions approchées avec RK2

n	y_n
0	1.00000000
1	1.22344942
2	1.45002001
3	1.67415723
4	1.89054577
5	2.09422177

TABLE 5.4. Solutions approchées avec RK4

Chapitre 6

Équations aux dérivées partielles

6.1. Dérivation numérique

En cours de rédaction

On rappellera (voir par exemple [BM03, Chapitre 3]) que si f est suffisamment régulière, on a

$$f''(t) \approx \frac{1}{h^2}\left(f(t-h) - 2f(t) + f(t+h)\right), \tag{6.1}$$

avec une erreur en $\mathcal{O}(h^2)$.

6.2. Applications : résolution numérique d'équation aux dérivées partielles

En cours de rédaction

Contenu provisoire :

On consultera le liens et les notes manuscrites provisoires (dont seuls quelques extraits seront traités) suivants

`https://fr.wikipedia.org/wiki/%C3%89quation_de_Schr%C3%B6dinger`
`https://fr.wikipedia.org/wiki/%C3%89quation_aux_d%C3%A9riv%C3%A9es_partielles`
`http://utbmjb.chez-alice.fr/Polytech/MNBmater/notes_manuscrites/discretisation_edp_chaleur.pdf`
`http://utbmjb.chez-alice.fr/Polytech/MNBmater/notes_manuscrites/mef.pdf`

Chapitre 7

Paradoxes

Si le temps le permet, nous essayerons de clore ce cours par une présentation de petits paradoxes. Vous constaterez que, peut-être, vous manipulez des choses par habitude, en en oubliant les raisons. La référence qui suit sera présentée de façon abrégée, lors du dernier CM.

Voir [Bas14b].

On pourra aussi consulter `http://utbmjb.chez-alice.fr/INSA/zetetique/paradoxes_enplus_papier.pdf`

Annexe A

Compléments sur les approximations polynômiales de $\ln(1+x)$ et de e^x

Nous donnons dans cette annexe des compléments facultatifs sur les résultats du chapitre 1 page 2.

A.1. Approximation de e^x par les séries

A.1.1. Principes théoriques

En utilisant quelques résultats classiques sur les séries, nous allons obtenir des résultats légèrement différents de la proposition 1.4 page 5.

Établissons maintenant le résultat principal. Nous rappelons que nous tenons pour vrai le résultat (1.1a) page 2.

Remarquons que la convergence de la série de terme général $x^n/n!$ résulte de [Bas22a, chapitre "Séries"].

Nous allons maintenant nous intéresser à l'étude du reste $R_n(x)$ de la série associée à l'exponentielle

$$\forall x \in]-1,1], \quad R_n(x) = \sum_{k=n+1}^{+\infty} \frac{x^k}{k!} \tag{A.1}$$

et qui vérifie donc aussi (1.23a) où p_n est défini par (1.20).

Proposition A.1. *On pose, pour tout $n \in \mathbb{N}^*$,*

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}. \tag{A.2}$$

On pose

(1) Si $x \geq 0$

$$\forall n \in \mathbb{N}, \quad a_n = 0, \tag{A.3a}$$

$$\forall n \geq E(x-1), \quad b_n = \frac{x^{n+1}}{(n+1)!}\left(\frac{1}{1-\frac{x}{n+2}}\right). \tag{A.3b}$$

(2) Si $x \leq 0$

$$\forall n \in \mathbb{N}, \quad a_n = \begin{cases} \dfrac{x^{n+1}}{(n+1)!} & \text{si } n \text{ est pair,} \\ 0 & \text{si } n \text{ est impair,} \end{cases} \tag{A.3c}$$

$$\forall n \in \mathbb{N}, \quad b_n = \begin{cases} 0 & \text{si } n \text{ est pair,} \\ \dfrac{x^{n+1}}{(n+1)!} & \text{si } n \text{ est impair,} \end{cases} \tag{A.3d}$$

On a alors

$$\forall n \geq N(x), \quad e^x - p_n(x) \in [a_n, b_n], \tag{A.4}$$

et en posant $\varepsilon_n = b_n - a_n \geq 0$, on a (1.23b). *En d'autres termes, pour tout $n \in \mathbb{N}^*$, pour tout $x \in \mathbb{R}$,*

(1) $p_n(x)$ constitue une approximation de e^x avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini,

(2) $p_n(x) + a_n$ et $p_n(x) + b_n$ constituent respectivement deux approximations par défaut et par excés de e^x avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini.

Démonstration.

(1) Premier cas : $x \in \mathbb{R}_+$.

La série définie par (1.1a) est une série à terme positif. On a donc

$$R_n(x) \geq 0. \tag{A.5}$$

On peut donc écrire successivement *a priori* dans $[0, +\infty]$ (voir par exemple [Bas22a, Chapitre "Séries"]) pour tout $x \in \mathbb{R}_+$ et $n \in \mathbb{N}$:

$$\begin{aligned}
\sum_{k=n+1}^{+\infty} \frac{x^k}{k!} &= \frac{x^{n+1}}{(n+1)!} + \sum_{k=n+2}^{+\infty} \frac{x^k}{k!}, \\
&= \frac{x^{n+1}}{(n+1)!} \left(1 + \sum_{k=n+2}^{+\infty} \frac{x^{k-n-1}(n+1)!}{k!} \right), \\
&= \frac{x^{n+1}}{(n+1)!} \left(1 + \sum_{k=n+2}^{+\infty} \frac{(n+1)!}{(n+1)!(n+2) \times ... \times k} x^{k-n-1} \right), \\
&= \frac{x^{n+1}}{(n+1)!} \left(1 + \sum_{k=n+2}^{+\infty} \frac{1}{(n+2) \times ... \times k} x^{k-n-1} \right),
\end{aligned}$$

chacun des $k - (n+2) + 1 = k - n - 1$ termes des dénominateurs $(n+2) \times ... \times k$ de la somme est minoré par $n+2$ de sorte que $\frac{1}{(n+2)\times ... \times k}$ est majoré par $1/((n+2)^{k-n-1}$:

$$= \frac{x^{n+1}}{(n+1)!} \left(1 + \sum_{k=n+2}^{+\infty} \frac{1}{(n+2)^{k-n-1}} x^{k-n-1} \right),$$

on pose $k' = k - n - 1$:

$$\begin{aligned}
&= \frac{x^{n+1}}{(n+1)!} \left(1 + \sum_{k=1}^{+\infty} \frac{1}{(n+2)^{k'}} x^{k'} \right), \\
&= \frac{x^{n+1}}{(n+1)!} \sum_{k=0}^{+\infty} \frac{1}{(n+2)^k} x^k
\end{aligned}$$

et donc,

$$R_n(x) = \frac{x^{n+1}}{(n+1)!} \sum_{k=0}^{+\infty} \left(\frac{x}{n+2} \right)^k \in [0, +\infty]. \tag{A.6}$$

Si on prend n tel que $n > x - 2$ en choisissant par exemple (ici $E(x)$ désigne la partie entière de x)

$$n \geq E(x-1) \tag{A.7}$$

alors $n > x - 2$ et $n + 2 > x$ et

$$0 \leq \frac{x}{n+2} < 1,$$

et la série géométrique de raison $\frac{x}{n+2}$ converge (voir par exemple [Bas22a, Chapitre "Séries"]) et on a

$$\begin{aligned} R_n(x) &= \frac{x^{n+1}}{(n+1)!}\left(\frac{1}{1-\frac{x}{n+2}}\right), \\ &< +\infty, \end{aligned}$$

et donc,

$$\forall x \in \mathbb{R}_+, \quad \forall n \geq E(x-1), \quad R_n(x) \leq \frac{x^{n+1}}{(n+1)!}\left(\frac{1}{1-\frac{x}{n+2}}\right). \tag{A.8}$$

Ainsi, grâce à (A.5) et (A.8), on considère a_n et b_n définis par (A.3a) et (A.3b). On retrouve donc l'équation (1.39) ainsi que des équations un peu différentes de celles du cas 2a page 7 de la preuve de la proposition 1.4 page 5. On conclue comme à la fin du point 2a page 7.

(2) Deuxième cas : $x \in \mathbb{R}_-$.

La série définie par (1.1a) est une série alternée (voir par exemple [Bas22a, Chapitre "Séries"]) . En effet :

- On pose
$$u_n(x) = \frac{x^n}{n!}, \tag{A.9}$$
qui est du signe de x^n, c'est à dire de $(-1)^n$.
- On a aussi
$$\lim_{n\to+\infty} |u_n(x)| = 0,$$
puisque la série associée converge.
- On a enfin
$$\frac{u_{n+1}(x)}{u_n(x)} = \frac{x^{n+1}}{(n+1)!}\frac{n!}{x^n} = \frac{x}{n+1} < 1,$$
et la suite $|u_n(x)|$ est décroissante, à partir d'un certain rang.

Ainsi, d'après le théorème sur les séries alternées (voir par exemple [Bas22a, Chapitre "Séries"]) , la série de terme général $u_n(x)$ converge. De plus le reste $R_n(x)$ a un signe égal à celui du premier terme négligé, qui est $(-1)^{n+1}$ et une valeur absolue inférieure à celle du premier premier terme négligé, qui est $\frac{x^{n+1}}{(n+1)!}$. On considère donc a_n et b_n définis par (A.3c) et (A.3d). On retrouve donc les équations (1.52) et (1.54) ainsi que des équations un peu moins générales que celles du cas 2b page 8 de la preuve de la proposition 1.4 page 5. On conclue comme à la fin du point 2a page 7.

□

REMARQUE A.2. On pourra consulter la fonction fournie sur le site habituel `approximation_exp.m` qui synthétise les différents calculs des propositions 1.4 page 5 et A.1 page 126. Cette fonction propose donc, pour $x \in \mathbb{R}$, les valeurs de $a_n + p_n(x)$ et de $b_n + p_n(x)$. Ces approximations sont aussi rationnelles si x l'est. Voir les simulations numériques de la section 1.4.3 page 9.

REMARQUE A.3. Il pourrait être intéressant de comparer ce que donnent les approximations fournies par e^x et e^{-x} censées être inverses l'un de l'autre.

A.1.2. Simulations numériques

Présentons quelques simulations numériques faites grâce à la fonction `approximation_exp.m`, qui utilise les résultats des propositions 1.4 page 5 et A.1 page 126. Elle envoie, pour n et x donnés, les valeurs

$$g_n = a_n + p_n(x), \tag{A.10a}$$
$$h_n = b_n + p_n(x); \tag{A.10b}$$

Une mesure de l'erreur peut être aussi donnée par

$$\eta_n = \max(e^x - g_n, n_n - e^x). \tag{A.11}$$

(1)

n	g_n	h_n	$h_n - g_n$	η_n
10	2.7182818	2.7182818	0.0000000022775	$2.26054\,10^{-9}$
11	2.7182818	2.7182818	0.00000000017397	$1.72876\,10^{-10}$
20	2.7182818	2.7182818	9.3204×10^{-22}	$9.30039\,10^{-22}$
21	2.7182818	2.7182818	4.0440×10^{-23}	$4.03605\,10^{-23}$
30	2.7182818	2.7182818	3.9230×10^{-36}	$3.91903\,10^{-36}$
31	2.7182818	2.7182818	1.1876×10^{-37}	$1.18650\,10^{-37}$
40	2.7182818	2.7182818	7.2910×10^{-52}	$7.28678\,10^{-52}$
41	2.7182818	2.7182818	1.6946×10^{-53}	$1.69368\,10^{-53}$
50	2.7182818	2.7182818	1.2641×10^{-68}	$1.26363\,10^{-68}$
51	2.7182818	2.7182818	2.3842×10^{-70}	$2.38337\,10^{-70}$
60	2.7182818	2.7182818	3.2297×10^{-86}	$3.22887\,10^{-86}$
61	2.7182818	2.7182818	5.1252×10^{-88}	$5.12389\,10^{-88}$
70	2.7182818	2.7182818	1.6561×10^{-104}	$1.65574\,10^{-104}$
71	2.7182818	2.7182818	2.2681×10^{-106}	$2.26772\,10^{-106}$
80	2.7182818	2.7182818	2.1296×10^{-123}	$2.12930\,10^{-123}$
81	2.7182818	2.7182818	2.5654×10^{-125}	$2.56504\,10^{-125}$

TABLE A.1. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.11) pour $x = 1$

Voir le tableau A.1 pour $x = 1$.

(2)

Voir le tableau A.2 page suivante pour $x = -1$.

A.2. Approximation de $\ln(1+x)$

A.2.1. Principes théoriques (par la formule de Taylor-Lagrange)

Montrons le résultat suivant :

PROPOSITION A.4. *On pose, pour tout $n \in \mathbb{N}^*$,*

$$p_n(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^k}{k}. \tag{A.12}$$

(1) On a si $x > 1$,

$$\lim_{n \to +\infty} |p_n(x)| = +\infty \tag{A.13}$$

n	g_n	h_n	$h_n - g_n$	η_n
10	0.36787944	0.36787946	0.000000015836	$1.38981\,10^{-8}$
11	0.36787944	0.36787944	0.0000000013197	$1.16982\,10^{-9}$
20	0.36787944	0.36787944	1.2372×10^{-20}	$1.15199\,10^{-20}$
21	0.36787944	0.36787944	5.6238×10^{-22}	$5.25251\,10^{-22}$
30	0.36787944	0.36787944	7.6874×10^{-35}	$7.31852\,10^{-35}$
31	0.36787944	0.36787944	2.4023×10^{-36}	$2.29043\,10^{-36}$
40	0.36787944	0.36787944	1.8896×10^{-50}	$1.82005\,10^{-50}$
41	0.36787944	0.36787944	4.4991×10^{-52}	$4.33722\,10^{-52}$
50	0.36787944	0.36787944	4.0753×10^{-67}	$3.95357\,10^{-67}$
51	0.36787944	0.36787944	7.8370×10^{-69}	$7.60735\,10^{-69}$
60	0.36787944	0.36787944	1.2454×10^{-84}	$1.21408\,10^{-84}$
61	0.36787944	0.36787944	2.0086×10^{-86}	$1.95898\,10^{-86}$
70	0.36787944	0.36787944	7.4325×10^{-103}	$7.27142\,10^{-103}$
71	0.36787944	0.36787944	1.0323×10^{-104}	$1.01021\,10^{-104}$
80	0.36787944	0.36787944	1.0904×10^{-121}	$1.06962\,10^{-121}$
81	0.36787944	0.36787944	1.3298×10^{-123}	$1.30471\,10^{-123}$

TABLE A.2. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.11) pour $x = -1$

(2) Si au contraire, $x \in]-1, 1]$, en posant

$$a_n = \begin{cases} \begin{cases} 0, & \text{si } n \text{ pair,} \\ -\frac{x^{n+1}}{n+1}, & \text{si } n \text{ impair,} \end{cases} & \text{si } x \in [0,1], \\ \max\left(-\left(\frac{-x}{x+1}\right)^{n+1}, -\frac{1}{1+x}\frac{(-x)^{n+1}}{n+1}\right) & \text{si } \in]-1,0], \end{cases} \tag{A.14a}$$

$$b_n = \begin{cases} \begin{cases} \frac{x^{n+1}}{n+1}, & \text{si } n \text{ pair,} \\ 0, & \text{si } n \text{ impair.} \end{cases} & \text{si } x \in [0,1], \\ 0, & \text{si } \in]-1,0], \end{cases} \tag{A.14b}$$

on a alors

$$\forall n \in \mathbb{N}, \quad \ln(1+x) - p_n(x) \in [a_n, b_n], \tag{A.15a}$$

et en posant $\varepsilon_n = b_n - a_n \geq 0$, on a

$$\lim_{n \to +\infty} \varepsilon_n = 0. \tag{A.15b}$$

En d'autres termes, pout tout $n \in \mathbb{N}^*$, pour tout $x \in]-1, 1]$,

(1) $p_n(x)$ constitue une approximation de $\ln(1+x)$ avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini,

(2) $p_n(x) + a_n$ et $p_n(x) + b_n$ constituent respectivement deux approximations par défaut et par excès de $\ln(1+x)$ avec une erreur inférieure à ε_n, qui tend vers zéro quand n tend vers l'infini.

Au contraire, si $x > 1$, cette approximation est inutilisable.

On retrouve donc d'une part le développement usuel de $\ln(1+x)$ par exemple de [Bas22b, Annexe "Quelques développements limités usuels"] et d'autre part le résultat habituel sur les séries (voir par exemple [Bas22a, Chapitre "Séries"]) .

REMARQUE A.5. Notons que si x est rationnel, la proposition A.4 propose une approximation rationnelle de $\ln(1+x)$.

REMARQUE A.6. Le cas $x=-1$ n'est pas pertinent, puisque $\ln(1+x)$ n'est pas défini. De plus, on a alors d'après (A.12)

$$p_n(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{(-1)^k}{k} = -\sum_{k=1}^{n} \frac{1}{k},$$

pour laquelle on sait, [Bas22a, Chapitre "Séries"], que

$$\lim_{n\to+\infty} p_n(-1) = -\infty. \tag{A.16}$$

Cela peut nous permettre néanmoins d'écrire formellement que

$$\lim_{n\to+\infty} p_n(-1) = \ln(0).$$

DÉMONSTRATION DE LA PROPOSITION A.4. On s'appuiera pour montrer ce résultat sur la formule de Taylor-Lagrange, en complétant par la formule de la somme de la suite géométrique pour combler deux cas lacunaires.

(1) Considérons la fonction f définie par

$$\forall x \in]-1,+\infty], \quad f(x) = \ln(1+x). \tag{A.17}$$

On a aisément

$$\begin{aligned} f'(x) &= \frac{1}{x+1} = (x+1)^{-1}, \\ f''(x) &= -(x+1)^{-2}, \\ f'''(x) &= 2(x+1)^{-3}, \\ f^{(4)}(x) &= -2\times 3(x+1)^{-4}, \end{aligned}$$

et on vérifie par récurrence sur n que

$$\forall n \in \mathbb{N}^*, \quad f^{(n)}(x) = (-1)^{n-1}(n-1)!(x+1)^{-n},$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!}{(x+1)^n}. \tag{A.18}$$

On a alors

$$f(1) = 0, \tag{A.19}$$

et

$$\forall n \in \mathbb{N}^*, \quad f^{(n)}(0) = (-1)^{n-1}(n-1)! \tag{A.20}$$

La formule de Taylor-Lagrange (voir [Bas22b, Chapitre "Dérivée, différentiation"]), à l'ordre n appliquée à la fonction f sur l'intervalle $[0,x]$ (ou $[x,0]$) fournit

$$\forall x \in]-1,+\infty], \quad \ln(1+x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0) x^k + \frac{1}{(n+1)!} f^{(n+1)}(\xi) x^{n+1},$$

où

$$\xi \in \begin{cases}]0,x[, & \text{si } x \geq 0, \\ ]x,0[, & \text{si } x \leq 0. \end{cases} \tag{A.21}$$

et donc, pour $n \in \mathbb{N}^*$, d'après (A.19) et (A.20),

$$\forall x \in]-1,+\infty], \quad \ln(1+x) = \sum_{k=1}^{n} \frac{1}{k!}(-1)^{k-1}(k-1)!x^k + \frac{1}{(n+1)!}\frac{(-1)^n n!}{(\xi+1)^{n+1}}x^{n+1},$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad \ln(1+x) = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k + \frac{(-1)^n x^{n+1}}{(n+1)(\xi+1)^{n+1}}. \tag{A.22}$$

Pour toute la suite, on pose

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad p_n(x) = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k, \tag{A.23a}$$

$$R_n(x) = \frac{(-1)^n x^{n+1}}{(n+1)(\xi+1)^{n+1}}, \tag{A.23b}$$

de sorte que (A.22) s'écrit

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad \ln(1+x) = p_n(x) + R_n(x). \tag{A.24}$$

(2) Proposons une expression alternative de $R_n(x)$ qui permettra de combler deux cas lacunaires. Cette petite astuce est issue de `https://les-mathematiques.net/vanilla/index.php?p=discussion/881075#Comment_881075`.

Remarquons que l'on a classiquement (somme de la suite géométrique de raison $-t \neq 1$)

$$\forall n \in \mathbb{N}^*, \quad \forall t \in]-1,+\infty], \quad \frac{1-(-t)^n}{1+t} = \sum_{k=0}^{n-1}(-1)^k t^k,$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall t \in]-1,+\infty], \quad \frac{1}{1+t} = \sum_{k=0}^{n-1}(-1)^k t^k + \frac{1}{1+t},$$

ce qui donne par intégration par rapport à t entre 0 et $x \in]-1,+\infty]$:

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad \int_0^x \frac{1}{1+t}dt = \sum_{k=0}^{n-1}(-1)^k \int_0^x t^k dt + (-1)^n \int_0^x \frac{t^n}{1+t}dt,$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad \ln(1+x) = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k + (-1)^n \int_0^x \frac{t^n}{1+t}dt,$$

ce qui donne, en comparant avec (A.23) et (A.24) sont toujours valables, à condition de remplacer (A.23b) par

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad R_n(x) = (-1)^n \int_0^x \frac{t^n}{1+t}dt. \tag{A.25}$$

Il ne reste plus qu'à étudier la suite $R_n(x)$, selon les différentes valeurs de x.

(3) (a) Premier cas : $x \in]1,+\infty[$. On laisse au lecteur le soin de vérifier que l'expression (A.23b) est dans ce cas inutile. Si on utilise plutôt (A.25), puisque $x > 1$ et $t \leq x$, on a $t+1 \leq x+x = 2x$ et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad |R_n(x)| = \left|\int_0^x \frac{t^n}{1+t}dt\right|, \tag{A.26}$$

$$= \int_0^x \frac{t^n}{1+t}dt, \tag{A.27}$$

$$\geq \int_0^x \frac{t^n}{2x}dt, \tag{A.28}$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad |R_n(x)| \geq \frac{1}{2x}\frac{x^{n+1}}{n+1},$$

dont on déduit (voir [Bas22a, Chapitre "Suites"]), puisque $x > 1$, que

$$\forall x \in]-1,+\infty], \quad \lim_{n\to+\infty} |R_n(x)| = +\infty. \tag{A.29}$$

On écrit d'après (A.24)

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad |p_n(x)| = |R_n(x) - \ln(1+x)|.$$

et donc d'après l'inégalité triangulaire

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad |p_n(x)| \geq |R_n(x)| - |\ln(1+x)|. \tag{A.30}$$

On déduit alors (A.13) de (A.29) et de (A.30).

Remarque A.7. De (A.25), on déduit que le signe de $R_n(x)$ est celui de $(-1)^n$. On a donc d'après (A.29)

$$\forall x \in]1,+\infty[, \quad \lim_{n+\to+\infty} R_{2n}(x) = +\infty, \tag{A.31a}$$

$$\lim_{n+\to+\infty} R_{2n+1}(x) = -\infty. \tag{A.31b}$$

D'après (A.24), on a donc

$$p_{2n}(x) = \ln(1+x) - R_{2n}(x),$$
$$p_{2n+1}(x) = \ln(1+x) - R_{2n+1}(x),$$

et d'après (A.31), on a

$$\forall x \in]1,+\infty[, \quad \lim_{n+\to+\infty} p_{2n}(x) = -\infty, \tag{A.32a}$$

$$\lim_{n+\to+\infty} p_{2n+1}(x) = +\infty. \tag{A.32b}$$

(b) Deuxième cas : $x \in [-1,1]$.

On utilise *a priori* l'expression (A.23b) de $R_n(x)$.

(i) Supposons $x \in [0,1]$.

D'après (A.21), on a $\frac{x^{n+1}}{(n+1)(\xi+1)^{n+1}} \geq 0$ et donc $R_n(x)$ est du signe de $(-1)^n$. On a aussi, d'après (A.21), $\xi + 1 \geq 1$ et donc

$$|R_n(x)| \leq \left|\frac{x^{n+1}}{(n+1)(\xi+1)^{n+1}}\right| = \frac{x^{n+1}}{(n+1)(\xi+1)^{n+1}} \leq \frac{x^{n+1}}{n+1},$$

et donc

$$\forall x \in [0,1], \quad \forall n \in \mathbb{N}^*, \quad R_n(x) \in [\widetilde{a}_n, \widetilde{b}_n], \tag{A.33a}$$

avec

$$\widetilde{a}_n = \begin{cases} 0, & \text{si } n \text{ pair}, \\ -\frac{x^{n+1}}{n+1}, & \text{si } n \text{ impair}, \end{cases} \tag{A.33b}$$

$$\widetilde{b}_n = \begin{cases} \frac{x^{n+1}}{n+1}, & \text{si } n \text{ pair}, \\ 0, & \text{si } n \text{ impair}. \end{cases} \tag{A.33c}$$

Remarquons aussi grâce au fait que la puissance l'emporte puisque $x \in [0,1]$ que

$$|R_n(x)| = \frac{x^{n+1}}{n+1} \to 0 \text{ quand } n \text{ tend vers l'infini.} \tag{A.33d}$$

Puis on conclut de la même façon que dans le cas 2a page 7.

(ii) Supposons $x \in [-1,0]$. D'après (A.21), on a $x < \xi < 0$ et donc

$$0 < x+1 < \xi+1 < 1$$

et donc

$$\frac{1}{(x+1)^n} > \frac{1}{(\xi+1)^n} > 1 \tag{A.34}$$

On a aussi d'après (A.23b)

$$R_n(x) = -\frac{(-x)^{n+1}}{(n+1)(\xi+1)^{n+1}}$$

et donc

$$R_n(x) \leq 0, \tag{A.35}$$

et aussi, d'après (A.34)

$$\begin{aligned} |R_n(x)| &= \frac{(-x)^{n+1}}{(n+1)(\xi+1)^{n+1}}, \\ &\leq \frac{(-x)^{n+1}}{(n+1)(x+1)^{n+1}}, \end{aligned}$$

et donc

$$|R_n(x)| = \left(\frac{-x}{x+1}\right)^{n+1}. \tag{A.36}$$

(A) On a enfin

$$\begin{aligned} 0 < \frac{-x}{x+1} < 1 &\iff -x < x+1, \\ &\iff 2x > -1, \\ &\iff x > -1/2, \end{aligned}$$

et donc, d'après (A.35) et (A.36), il vient

$$\forall x \in]-1/2, 0], \quad \forall n \in \mathbb{N}^*, \quad R_n(x) \in [\widetilde{a}_n, \widetilde{b}_n], \tag{A.37a}$$

avec

$$\widetilde{a}_n = -\left(\frac{-x}{x+1}\right)^{n+1}, \tag{A.37b}$$

$$\widetilde{b}_n = 0, \tag{A.37c}$$

$$|R_n(x)| = \left(\frac{-x}{x+1}\right)^{n+1} \to 0 \text{ quand } n \text{ tend vers l'infini.} \tag{A.37d}$$

Puis on conclut de la même façon que dans le cas 2a page 7.

(B) Il reste donc à traiter le cas $x \in]-1,-1/2]$. On utilise alors, dans ce cas, l'expression (A.25) qui s'écrit

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad R_n(x) = -(-1)^n \int_x^0 \frac{t^n}{1+t} dt. \tag{A.38}$$

On retrouve alors, d'une part, (A.35), puisque t^n est du signe de $(-1)^n$. D'autre part, on a d'après (A.38) :

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,+\infty], \quad |R_n(x)| = \left| \int_x^0 \frac{t^n}{1+t} dt \right|, \\ \leq \int_x^0 \frac{|t^n|}{1+t} dt.$$

si on choisit $x \in [-1,0]$, on a $t \geq x$ et $1+t \geq 1+x > 0$, $0 < 1/(t+1) < 1/(1+x)$ et

$$\leq \frac{1}{1+x} \int_x^0 |t^n| dt,$$

si on fait le changement de variable $u = -t \geq 0$

$$\leq -\frac{1}{1+x} \int_{-x}^0 |(-u)^n| du, \\ \leq \frac{1}{1+x} \int_0^{-x} u^n du,$$

et donc

$$\forall n \in \mathbb{N}^*, \quad \forall x \in]-1,0], \quad |R_n(x)| \leq \frac{1}{1+x} \frac{(-x)^{n+1}}{n+1} \to 0 \text{ quand } n \text{ tend vers l'infini.} \tag{A.39}$$

Si on compare les expressions des majorants de $|R_n(x)|$ données par (A.37d) et (A.39), on constate que, dans le cas où $x \in]-1/2,0[$, celle fournie par (A.39) est théoriquement meilleure puisque l'on a successivement

$$\frac{1}{1+x} \frac{(-x)^{n+1}}{n+1} < \left(\frac{-x}{x+1} \right)^{n+1} \iff \frac{1}{1+x} \frac{(-x)^{n+1}}{n+1} < \frac{(-x)^{n+1}}{(x+1)^{n+1}}, \\ \iff \frac{1}{1+x} \frac{1}{n+1} < \frac{1}{(x+1)^{n+1}}, \\ \iff \frac{1}{n+1} < \frac{1}{(x+1)^n},$$

ce qui est vrai en théorie pour n tendant vers l'infini puisque $\frac{1}{n+1}$ tend vers zéro et $\frac{1}{(x+1)^n}$ vers l'infini. Néanmoins, on poura, par sécurité conserver les deux expressions et unifier les cas $x \in]-1/2,0]$ et $x \in [-1,0]$ en écrivant finalement à la place de (A.37), l'expression suivante qui reste valable pour toutes les valeurs $x \in [-1,0]$:

$$\forall x \in]-1,0], \quad \forall n \in \mathbb{N}^*, \quad R_n(x) \in [\widetilde{a}_n, \widetilde{b}_n], \tag{A.40a}$$

avec

$$\widetilde{a}_n = \max \left(-\left(\frac{-x}{x+1} \right)^{n+1}, -\frac{1}{1+x} \frac{(-x)^{n+1}}{n+1} \right), \tag{A.40b}$$

$$\widetilde{b}_n = 0, \tag{A.40c}$$

$$|R_n(x)| \to 0 \text{ quand } n \text{ tend vers l'infini.} \tag{A.40d}$$

Puis on conclut de la même façon que dans le cas 2a page 7.

REMARQUE A.8. Si on utilise l'expression de $R_n(x)$ donnée par (A.25) dans le cas où $x \in [0,1]$, on obtient que $R_n(x)$ est du signe de $(-1)^n$ et que l'on a puisque $1+t \geq 1$

$$\begin{aligned} |R_n(x)| &\leq \int_0^x \frac{t^n}{1+t} dt, \\ &\leq \int_0^x t^n dt, \\ &= \frac{x^{n+1}}{n+1}, \end{aligned}$$

et on obtient donc exactement (A.33).

On unifie, pour conclure, les deux cas donnés par les expressions (A.33) et (A.40). □

REMARQUE A.9. On peut grâce à la proposition A.4, proposer une ou deux approximations par défaut et par excès, g_n et h_n, de $\ln(X)$ pour tout $X \in \mathbb{R}_+^*$ sous la forme

$$\ln(X) \in [g_n, h_n]. \tag{A.41}$$

(1) Premier cas : $X > 2$.

On pose

$$x = \frac{1}{X} - 1. \tag{A.42}$$

Puisque $X \in]0, 1/2[$, on a

$$x = \in]-1, -1/2[, \tag{A.43}$$

On a

$$\ln(1+x) = \ln\left(\frac{1}{X}\right) = -\ln(X),$$

et donc

$$\ln(X) = -\ln(1+x). \tag{A.44}$$

D'après (A.15a), on a aussi

$$\ln(1+x) \in [a_n + p_n(x), b_n + p_n(x)]$$

et donc, d'après (A.44),

$$\ln(X) \in [-b_n - p_n(x), -a_n - b_n(x)]. \tag{A.45}$$

(2) Deuxième cas : $X \in [1,2]$.

Nous avons deux approximations possibles.

(a) La première est donnée de nouveau par les formules (A.42), (A.44) et (A.45) du cas (1). En effet, $1/X \in [1/2, 1]$ et (A.43) est remplacée par

$$x = \in]-1/2, 0]. \tag{A.46}$$

(b) Une autre façon de procéder et de poser

$$x = X - 1 \in [0,1], \tag{A.47}$$

puis

$$\ln(X) = \ln(1+x), \tag{A.48}$$

et donc d'après (A.15a),

$$\ln(X) \in [a_n + p_n(x), b_n + p_n(x)]. \tag{A.49}$$

REMARQUE A.10. Dans ce cas, les approximations fournies du cas (2a) et du cas (2b) ne sont pas nécessairement de la même qualité. Par exemple, pour $X = 2$, l'approximation fournie par (A.45) correspond d'après la proposition A.4 page 129 à une erreur correspondant à $x = 1$, c'est à dire, d'après (A.14) à a_n et b_n donnés par

$$a_n = \begin{cases} 0, & \text{si } n \text{ pair,} \\ -\frac{1}{n+1}, & \text{si } n \text{ impair,} \end{cases}$$

$$b_n = \begin{cases} \frac{1}{n+1}, & \text{si } n \text{ pair,} \\ 0, & \text{si } n \text{ impair.} \end{cases}$$

c'est-à-dire à un majorant de l'erreur donné par

$$\varepsilon_n = \frac{1}{n+1}. \tag{A.50}$$

Si, au contraire, on utilise l'approximation fournie par (A.49), d'après la proposition A.4 page 129 utilisée avec x donné par (A.42), c'est-à-dire $x = -1/2$ les expressions de a_n et b_n fournies par (A.14) s'écrivent

$$a_n = -\frac{1}{2^n(n+1)},$$
$$b_n = 0,$$

c'est-à-dire à un majorant de l'erreur donné par

$$\varepsilon_n = \frac{1}{2^n(n+1)}, \tag{A.51}$$

expression qui tend vers zéro plus rapidement que celle donnée par (A.50) du fait du facteur supplémentaire 2^n. Voir les simulations numériques de la section 1.4.3 page 9.

(3) Troisième cas : $X \in]0, 1]$.

On utilise de nouveau les formules du cas 2b : on pose

$$x = X - 1 \in]-1, 0], \tag{A.52}$$

puis (A.48). On a de nouveau d'après (A.15a),

$$\ln(X) \in [a_n + p_n(x), b_n + p_n(x)]. \tag{A.53}$$

On pourra consulter la fonction fournie sur le site habituel `approximation_ln.m` qui synthètise les différents calculs de cette remarque et propose donc, pour $X > 0$, les valeurs de $[g_n, h_n]$, approximations par défaut et par excès de $\ln(X)$ d'après (A.45), (A.49), et (A.53). Voir les simulations numériques de la section 1.4.3 page 9.

‘

REMARQUE A.11. D'après la remarque A.5 page 131, la remarque A.9 propose aussi une approximation rationnelle de $\ln(X)$ si X est rationnel.

A.2.2. Principes théoriques (par les séries)

En utilisant quelques résultats classiques sur les séries, nous allons retrouver (partiellement) les résultats de la proposition A.4 page 129.

Comme rappelé par exemple dans [Bas22c, Chapitre "Séries entières et fonctions usuelles sur $\mathbb{C}$"], le logarithme peut être introduit de différentes façons.

Établissons mainteant le résultat principal. Nous rappelons que nous tenons pour vrai le résultat (1.1b) page 2.

Remarquons que la convergence de la série de terme général $(-1)^{n-1}\frac{x^n}{n}$ résulte aussi

— dans le cas où $x \in [0,1]$ du théorème des séries alternées. Voir [Bas22a, chapitre "Séries"]
— dans le cas où $x \in]-1,0]$ du fait que par exemple la valeur absolue de $(-1)^{n-1}\frac{x^n}{n}$ est un petit o de $(-x)^n$, série géométrique convergente.

Remarque A.12. Pour ceux qui connaissent la notion de séries entières, la série entière de terme général $\frac{(-1)^{n-1}}{n}$ est de rayon de convergence égal à un. Comme dans [Bas22c, Chapitre intitulé "Séries entières et fonctions usuelles sur $\mathbb{C}$"], on peut par dérivation de la serie terme à terme montrer que cette série entière correspond à $\ln(1+x)$. Le problème de cette méthode est que l'on montre que le rayon de convergence de cette série vaut 1 et que (1.1b) n'est vrai que sur $]-1,1[$. Le passage par continuité en $x=1$ peut-être aussi montré. Voir [Bas22c, section 2 de l'annexe intitulée "Quelques calculs explicites de sommes de Séries"].

Nous allons maintenant nous intéresser à l'étude du reste $R_n(x)$ de la série associée au logarithme et défini par

$$\forall x \in]-1,1], \quad R_n(x) = \sum_{k=n+1}^{+\infty} (-1)^{k-1}\frac{x^k}{k}, \tag{A.54}$$

et qui vérifie donc aussi (A.24) où p_n est défini par (A.12).

(1) Premier cas : $x \in [0,1]$.

La série définie par (1.1b) est une série alternée (voir par exemple [Bas22a, Chapitre "Séries"]) . En effet :

- En posant
$$u_n(x) = (-1)^{n-1}\frac{x^n}{n}, \tag{A.55}$$
On a clairement
$$u_n(x) = (-1)^{n-1}|u_n(x)|,$$
en prenant garde au fait que c'est en fait l'opposé de la série qui est alternée.
- On a aussi
$$\lim_{n\to+\infty} |u_n(x)| = 0.$$
- On a enfin
$$\frac{u_{n+1}(x)}{u_n(x)} = \frac{x^{n+1}}{n+1}\frac{n}{x^n} = \frac{n}{n+1}x \leq \frac{n}{n+1} < 1,$$
et la suite $|u_n(x)|$ est décroissante.

Ainsi, d'après le théorème des séries alternées (voir par exemple [Bas22a, Chapitre "Séries"]) la série de terme général $u_n(x)$ converge. De plus le reste $R_n(x)$ a un signe égal à celui du premier terme négligé, qui est $(-1)^n$ et une valeur absolue inférieure à celle du premier terme négligé, qui est $\frac{x^{n+1}}{n+1}$, ce qui est exactement le résultat de la proposition A.4 page 129. On retrouve donc le cas 3(b)i page 133 de la preuve de la proposition A.4 page 129.

(2) Deuxième cas : $x \in]-1,0]$.

La série définie par (1.1b) n'est plus une série alternée, puisque son terme général vaut

$$u_n(x) = (-1)^{n-1}\frac{x^n}{n},$$

et donc

$$u_n(x) = -\frac{(-x)^n}{n} \leq 0. \tag{A.56}$$

On a donc une série à termes négatifs. Le reste $R_n(x)$ défini par (A.54) est donc négatif et on écrit en utilisant la convention habituelle que l'on travaille dans $]-\infty,+\infty]$:

$$\begin{aligned}\forall x \in]-1,1], \quad |R_n(x)| &= \sum_{k=n+1}^{+\infty} \frac{(-x)^k}{k}, \\ &\leq \sum_{k=n+1}^{+\infty} \frac{(-x)^k}{n+1}, \\ &\leq \frac{1}{n+1} \sum_{k=n+1}^{+\infty} (-x)^k, \\ &\leq \frac{(-x)^{n+1}}{n+1} \sum_{k=0}^{+\infty} (-x)^k,\end{aligned}$$

et d'après la formule de la série géométrique puisque $(-x) \in [0,1[$

$$\leq \frac{(-x)^{n+1}}{n+1} \frac{1}{1+x}.$$

On retrouve donc exactement (A.35) et (A.39). On retrouve donc le cas 3(b)iiB page 134 de la preuve de la proposition A.4 page 129, valable en fait pour $x \in]-1,0]$.

Remarque A.13. Par cette méthode, on ne retrouve plus la majoration du cas 3(b)iiA page 134 de la preuve de la proposition A.4 page 129.

(3) Troisième cas : $x = -1$.

On a, en utilisant la notation (A.55)

$$u_n(x) = (-1)^{n-1} \frac{(-1)^n}{n} = -\frac{1}{n},$$

qui est le terme général (au signe près) d'une série de Riemann divergente (voir par exemple [Bas22a, Chapitre "Séries"]) Puisqu'elle est à termes négatifs, d'après le [Bas22a, Chapitre "Séries"] on a

$$\lim_{n\to+\infty} \sum_{k=1}^{n} (-1)^{n-1} \frac{(-1)^n}{n} = -\infty.$$

On retrouve donc la remarque A.6 page 131.

(4) Quatrième cas : $x > 1$.

On a, en utilisant la notation (A.55)

$$|u_n(x)| = \frac{x^{n-1}}{n},$$

et donc

$$\lim_{n\to+\infty} |u_n(x)| = +\infty. \tag{A.57}$$

La suite $u_n(x)$ ne tend donc pas vers zéro et la série associée ne peut converger. On retrouve donc (partiellement) le cas 3a page 132 de la preuve de la proposition A.4 page 129.

Plus précisément, utilisons le théorème suivant

Théorème A.14 (Série alternée divergente). *Soit une série $\sum u_n$ alternée (voir par exemple [Bas22a, Chapitre "Séries"]) pour laquelle la suite $(|u_n|)$ est croissante et contient au moins une*

valeur strictement positive. Alors la série de terme général u_n est divergente et

$$\lim_{n+\to+\infty} \sum_{k=0}^{2n} u_k = +\infty, \tag{A.58a}$$

$$\lim_{n+\to+\infty} \sum_{k=0}^{2n+1} u_k = -\infty. \tag{A.58b}$$

Démonstration. La preuve est très proche de celle du théorème des séries alternées ([Bas22a, Chapitre "Séries"]) dont la preuve est donnée par exemple dans [RDO87, section 1.3.3.2)]. Il suffit de l'adapter. Notons que l'hypothèse sur la suite $|u_n|$ implique qu'à partir d'un certain rang, tous les termes de la suite $|u_n|$ sont strictement positifs et que d'après le [Bas22a, Chapitre "Suites"]

$$\exists l \in]0, +\infty], \quad \lim_{n+\to+\infty} |u_n| = l. \tag{A.59}$$

Ainsi, en notant comme habituellement

$$\forall n \in \mathbb{N}, \quad S_n = \sum_{k=0}^{n} u_k, \tag{A.60}$$

il vient

$$\text{les suites } (S_{2n}) \text{ et } (S_{2n+1}) \text{ sont divergentes.} \tag{A.61}$$

En effet, si l'une d'entre elle convergeait, par exemple S_{2n}, alors puisque, d'après (A.60)

$$\forall n \in \mathbb{N}, \quad S_{2n} = \sum_{k=0}^{2n} u_k,$$

on aurait $\lim_{n\to+\infty} u_n = 0$, ce qui contredit (A.59). Remarquons enfin que

$$\begin{aligned}\forall n \in \mathbb{N}, \quad S_{2n+2} - S_{2n} &= u_{2n+2} + u_{2n+1} = (-1)^{2n+2}|u_{2n+2}| + (-1)^{2n+1}|u_{2n+1}| = |u_{2n+2}| - |u_{2n+1}| \geq 0,\\ S_{2n+3} - S_{2n+1} &= u_{2n+3} + u_{2n+2} = (-1)^{2n+3}|u_{2n+3}| + (-1)^{2n+2}|u_{2n+2}| = -|u_{2n+3}| + |u_{2n+2}| \leq 0.\end{aligned}$$

ce qui implique que les suites (S_{2n}) et (S_{2n+1}) sont respectivement croissantes et décroissantes. D'après le théorème de la limite monotone (voir [Bas22a, Chapitre "Suites"]) et (A.61), on obtient alors (A.58). □

Appliquons ce théorème à la suite définie par (A.55). On a

$$\begin{aligned}\frac{|u_{n+1}(x)|}{|u_n(x)|} &= \frac{x^{n+1}}{n+1}\frac{n}{x^n},\\ &= \frac{x^{n+1}}{x^n}\frac{n}{n+1},\\ &= x\frac{n}{n+1}\end{aligned}$$

Ainsi

$$\begin{aligned}\frac{|u_{n+1}(x)|}{|u_n(x)|} > 1 &\iff x\frac{n}{n+1} > 1,\\ &\iff x > \frac{n+1}{n},\\ &\iff x > 1 + \frac{1}{n},\end{aligned}$$

ce qui est vrai à partir d'un certain rang, puisque $x > 1$ et que $1 + 1/n \to 1$ quand n tend vers l'infini. L'opposé de la suite (u_n) obéit donc aux hypothèses du théorème A.14 page 139 dont on déduit

$$\lim_{n+\to+\infty} \sum_{k=1}^{2n} u_k(x) = -\infty,$$
$$\lim_{n+\to+\infty} \sum_{k=1}^{2n+1} u_k(x) = +\infty,$$

et on retrouve donc et (A.13) et (A.32).

A.2.3. Simulations numériques

Présentons quelques simulations numériques faites grâce à la fonction `approximation_ln.m`, qui utilise les résultats de la proposition A.4 page 129 et de la remarque A.9 page 136. Elle envoie, pour n et x donnés, les valeurs données par (A.10).

(1)

n	$p_n(-1)$
100	-5.18738
1000	-7.48547
10000	-9.78761
100000	-12.09015
1000000	-14.39273
10000000	-16.69531
100000000	-18.99790

TABLE A.3. Quelques valeurs de $p_n(-1)$

Pour illustrer (A.16), nous présentons dans le tableau A.3 quelques valeurs de $p_n(-1)$ pour différentes valeurs de n, dont on constate qu'elles semblent bien tendre numériquement vers $-\infty$, mais "lentement" (en $\ln(x)$ en fait !).

(2)

Pour illustrer (A.13) et la remarque A.7 page 133, nous présentons dans le tableau A.4 page suivante quelques valeurs de $p_n(2)$ pour différentes valeurs de n, dont on constate qu'elles semblent bien tendre numériquement vers $\pm\infty$.

(3)

Présentons maintenant quelques simulations pertinentes pour illustrer le cas 2 page 129 de la proposition A.4 page 129 en utilisant la remarque A.9 page 136 et l'équation (A.41). Une mesure de l'erreur peut être aussi donnée par

$$\eta_n = \max(\ln(X) - g_n, h_n - \ln(X)). \tag{A.62}$$

(a)

Voir le tableau A.5 page suivante pour $X = 3$.

(b)

Voir le tableau A.6 page 143 pour $X = 1/3$.

n	$p_n(x)$
100	$-8.4227\,10^{27}$
101	$1.6678\,10^{28}$
1000	$-7.1410\,10^{297}$
1001	$1.4268\,10^{298}$
10000	$-\infty$
10001	$+\infty$
100000	$-\infty$
100001	$+\infty$
1000000	$-\infty$
1000001	$+\infty$

TABLE A.4. Quelques valeurs de $p_n(x)$ pour $x = 2$

n	g_n	h_n	$h_n - g_n$	η_n
10	1.0958692	1.0990222	$3.15301\,10^{-3}$	$2.74308\,10^{-3}$
11	1.0969202	1.0988470	$1.92684\,10^{-3}$	$1.69208\,10^{-3}$
20	1.0985859	1.0986145	$2.86408\,10^{-5}$	$2.64032\,10^{-5}$
21	1.0985954	1.0986137	$1.82260\,10^{-5}$	$1.68563\,10^{-5}$
30	1.0986120	1.0986123	$3.36458\,10^{-7}$	$3.17608\,10^{-7}$
31	1.0986121	1.0986123	$2.17295\,10^{-7}$	$2.05457\,10^{-7}$
40	1.0986123	1.0986123	$4.41160\,10^{-9}$	$4.21879\,10^{-9}$
41	1.0986123	1.0986123	$2.87103\,10^{-9}$	$2.74826\,10^{-9}$
50	1.0986123	1.0986123	$6.15029\,10^{-11}$	$5.92987\,10^{-11}$
51	1.0986123	1.0986123	$4.02135\,10^{-11}$	$3.87978\,10^{-11}$
60	1.0986123	1.0986123	$8.91731\,10^{-13}$	$8.64642\,10^{-13}$
61	1.0986123	1.0986123	$5.84865\,10^{-13}$	$5.67546\,10^{-13}$
70	1.0986123	1.0986123	$1.33227\,10^{-14}$	$1.31006\,10^{-14}$
71	1.0986123	1.0986123	$8.65973\,10^{-15}$	$8.65973\,10^{-15}$
80	1.0986123	1.0986123	$2.22045\,10^{-16}$	$2.22045\,10^{-16}$
81	1.0986123	1.0986123	$2.22045\,10^{-16}$	$2.22045\,10^{-16}$

TABLE A.5. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.62) pour $X = 3$

(c)

Enfin, voir les tableaux A.7 page suivante et A.8 page 144 pour $X = 2$. Les calculs ont été fait de façon symbolique, grâce à la fonction `approximation_ln.m` et convertis ensuite en numérique, pour plus de précision dans les calculs. Comme le prévoit la remarque A.10 page 137, les simulations utilisant la méthode 2b page 136 (voir tableau A.7) convergent beaucoup plus vite que celles utilisant la méthode 2a page 136 (voir tableau A.8).

n	g_n	h_n	$h_n - g_n$	η_n
10	-1.0990222	-1.0958692	$3.15301\,10^{-3}$	$2.74308\,10^{-3}$
11	-1.0988470	-1.0969202	$1.92684\,10^{-3}$	$1.69208\,10^{-3}$
20	-1.0986145	-1.0985859	$2.86408\,10^{-5}$	$2.64032\,10^{-5}$
21	-1.0986137	-1.0985954	$1.82260\,10^{-5}$	$1.68563\,10^{-5}$
30	-1.0986123	-1.0986120	$3.36458\,10^{-7}$	$3.17608\,10^{-7}$
31	-1.0986123	-1.0986121	$2.17295\,10^{-7}$	$2.05457\,10^{-7}$
40	-1.0986123	-1.0986123	$4.41160\,10^{-9}$	$4.21879\,10^{-9}$
41	-1.0986123	-1.0986123	$2.87103\,10^{-9}$	$2.74826\,10^{-9}$
50	-1.0986123	-1.0986123	$6.15029\,10^{-11}$	$5.92987\,10^{-11}$
51	-1.0986123	-1.0986123	$4.02135\,10^{-11}$	$3.87978\,10^{-11}$
60	-1.0986123	-1.0986123	$8.91731\,10^{-13}$	$8.64642\,10^{-13}$
61	-1.0986123	-1.0986123	$5.84865\,10^{-13}$	$5.67546\,10^{-13}$
70	-1.0986123	-1.0986123	$1.33227\,10^{-14}$	$1.31006\,10^{-14}$
71	-1.0986123	-1.0986123	$8.65973\,10^{-15}$	$8.65973\,10^{-15}$
80	-1.0986123	-1.0986123	$2.22045\,10^{-16}$	$2.22045\,10^{-16}$
81	-1.0986123	-1.0986123	$2.22045\,10^{-16}$	$2.22045\,10^{-16}$

TABLE A.6. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.62) pour $X = 1/3$.

n	g_n	h_n	$h_n - g_n$	η_n
10	0.69306486	0.69315363	0.000088778	0.00008232440915165862358132780739
11	0.69310925	0.69314994	0.000040690	0.00003793520460620407812678235284
20	0.69314714	0.69314718	0.000000045413	0.00000004350891637314459361875002
21	0.69314716	0.69314718	0.000000021674	0.00000002080238503013864123779764
30	0.69314718	0.69314718	3.0043×10^{-11}	$2.915621396521253272124 \times 10^{-11}$
31	0.69314718	0.69314718	1.4552×10^{-11}	$1.413488211657578246923 \times 10^{-11}$
40	0.69314718	0.69314718	2.2183×10^{-14}	$2.167766099828288803 \times 10^{-14}$
41	0.69314718	0.69314718	1.0827×10^{-14}	$1.058626219617400708 \times 10^{-14}$
50	0.69314718	0.69314718	1.7415×10^{-17}	$1.709233711439502 \times 10^{-17}$
51	0.69314718	0.69314718	8.5402×10^{-18}	$8.38470554870752 \times 10^{-18}$
60	0.69314718	0.69314718	1.4219×10^{-20}	$1.399666432951 \times 10^{-20}$
61	0.69314718	0.69314718	6.9949×10^{-21}	$6.88714188698 \times 10^{-21}$
70	0.69314718	0.69314718	1.1930×10^{-23}	$1.176871126 \times 10^{-23}$
71	0.69314718	0.69314718	5.8822×10^{-24}	$5.80369051 \times 10^{-24}$
80	0.69314718	0.69314718	1.0212×10^{-26}	1.009047×10^{-26}
81	0.69314718	0.69314718	5.0438×10^{-27}	4.98442×10^{-27}

TABLE A.7. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.62) pour $X = 2$ pour la méthode 2b page 136.

n	g_n	h_n	$h_n - g_n$	η_n
10	0.64563492	0.73654401	0.090909	0.047512
11	0.65321068	0.73654401	0.083333	0.043397
20	0.66877140	0.71639045	0.047619	0.024376
21	0.67093591	0.71639045	0.045455	0.023243
30	0.67675814	0.70901620	0.032258	0.016389
31	0.67776620	0.70901620	0.031250	0.015869
40	0.68080338	0.70519363	0.024390	0.012344
41	0.68138410	0.70519363	0.023810	0.012047
50	0.68324716	0.70285500	0.019608	0.009900
51	0.68362423	0.70285500	0.019231	0.009708
60	0.68488328	0.70127672	0.016393	0.008264
61	0.68514769	0.70127672	0.016129	0.008130
70	0.68605534	0.70013985	0.014084	0.007092
71	0.68625096	0.70013985	0.013889	0.006993
80	0.68693624	0.69928192	0.012346	0.006211
81	0.68708680	0.69928192	0.012195	0.006135

TABLE A.8. Quelques valeurs de g_n, h_n, $h_n - g_n$ et η_n donné par (A.62) pour $X = 2$ pour la méthode 2a page 136.

Annexe B

Majoration de l'erreur relative

B.1. Principe théorique

LEMME B.1. *Soient un réel ou un complexe l et une suite réelle ou complexe (u_n), pour $n \geq n_0$, vérifiant* (1.2) *page 2 avec $l \neq 0$. On considère le nombre η_n défini par* (1.3) *et un nombre ε_n, supposé connu, vérifiant* (1.4) *page 2 et* (1.5) *page 2. Alors,*

$$\forall n \geq n_0, \quad \forall \varepsilon > 0, \quad \varepsilon_n \leq \frac{\varepsilon}{1+\varepsilon}|u_n| \Longrightarrow \left|\frac{u_n - l}{l}\right| \leq \varepsilon. \tag{B.1}$$

DÉMONSTRATION. On écrit d'après l'inégalité triangulaire et d'après (1.4),

$$|l| = |(-u_n) - (l - u_n)| \geq |u_n| - |u_n - l| = |u_n| - \eta_n \geq |u_n| - \varepsilon_n$$

D'après (1.5), la quantité $|u_n| - \varepsilon_n$ tend vers $|l| \neq 0$ quand n tend vers l'infini. Ainsi, il existe $N \geq n_0$ tel que pour tout $n \geq N$,

$$|u_n| - \varepsilon_n > 0 \tag{B.2}$$

et donc d'après (1.4),

$$\begin{aligned} \forall n \geq N, \quad \left|\frac{u_n - l}{l}\right| &\leq \frac{|u_n - l|}{|l|}, \\ &= \frac{\varepsilon_n}{|u_n| - \varepsilon_n}. \end{aligned}$$

Ainsi, on aura

$$\left|\frac{u_n - l}{l}\right| \leq \varepsilon, \tag{B.3}$$

dès que $n \geq N$ et

$$\frac{\varepsilon_n}{|u_n| - \varepsilon_n} \leq \varepsilon,$$

ce qui est équivalent à

$$\varepsilon_n \leq \varepsilon|u_n| - \varepsilon\varepsilon_n,$$

soit encore à

$$(1+\varepsilon)\varepsilon_n \leq \varepsilon|u_n|,$$

et donc à

$$\varepsilon_n \leq \frac{\varepsilon}{1+\varepsilon}|u_n|. \tag{B.4}$$

Notons que, puisque $\frac{\varepsilon}{1+\varepsilon} < 1$, si (B.4) est vérifié alors

$$\varepsilon_n < |u_n|, \tag{B.5}$$

et (B.2) est vérifié. Ainsi, (B.4) implique (B.3).

□

On sait que

$$\lim_{n\to+\infty} \varepsilon_n = 0,$$

et que

$$\lim_{n\to+\infty} \frac{\varepsilon}{1+\varepsilon}|u_n| = \frac{\varepsilon}{1+\varepsilon}|l| > 0.$$

et donc

$$\exists N \geq n_0, \quad \forall n \geq N, \quad \varepsilon_n \leq \frac{\varepsilon}{1+\varepsilon}|u_n|.$$

et en particulier

$$\exists N \geq n_0, \quad \varepsilon_N \leq \frac{\varepsilon}{1+\varepsilon}|u_N|. \tag{B.6}$$

Selon (B.1), (B.6) assure que

$$\left|\frac{u_N - l}{l}\right| \leq \varepsilon, \tag{B.7}$$

et donc que l'erreur relative est plus petite que ε.

Algorithme B.1 Algorithme de détermination de l'entier $N \leq n_{\max}$ pour lequel l'erreur relative au rang N est plus petite que ε : *rang_ erreur*($\varepsilon > 0$, $n_0 \in \mathbb{N}$, $n_{\max} \geq n_0$, $(u_n)_{n\geq n_0}$, $(\varepsilon_n)_{n\geq n_0}$) $\longrightarrow N$

Entrée :

ε, réel strictement positif.
n_0 entier naturel.
$n_{\max}$ entier naturel, supérieur à n_0.
$(u_n)_{n\geq n_0}$, suite approchant le nombre l recherché.
$(\varepsilon_n)_{n\geq n_0}$, suite de majorants de l'erreur $|u_n - l|$.

Sortie :

N, entier naturel, supérieur à n_0 tel que l'erreur relative au rang N est plus petite que ε.

test $\longleftarrow$ *vrai*
$n \longleftarrow n_0 - 1$
tant que test **faire**
  $n \longleftarrow n + 1$
  test $\longleftarrow \left(\left(\varepsilon_n > \frac{\varepsilon}{1+\varepsilon}|u_n|\right) \text{ et } (n < n_{\max})\right)$
fin tant que
$N \longleftarrow n$

L'intérêt (en analyse numérique) est que l'on peut déterminer le rang N pour lequel l'erreur relative commise est inférieur à ε (voir inégalité (B.7)) et, ce de façon algorithmique, en n'utilisant que les valeurs connues de u_n et de ε_n (voir inégalité (B.6)), comme le montre l'algorithme B.1. Dans cet algorithme, le paramètre $n_{\max}$ est le nombre maximal d'itérations, ce qui évite que l'algorithme ne tourne à l'infini.

Remarque B.2. Dans l'algorithme B.1, dans l'instruction

$$\text{test} \longleftarrow \left(\left(\varepsilon_n > \frac{\varepsilon}{1+\varepsilon}|u_n|\right) \text{ et } (n < n_{\max})\right)$$

les expressions $\left(\varepsilon_n > \frac{\varepsilon}{1+\varepsilon}|u_n|\right)$ et $(n < n_{\max})$ sont des booléens qui valent *faux* ou *vrai* selon que l'inégalité en question soit vraie ou fausse.

Notons aussi que cet algorithme détermine le plus entier N vérifiant (B.6). Mais l'erreur relative peut être strictement plus petite que ε.

REMARQUE B.3. On peut aussi assurer que l'erreur absolue $|u_n - l|$ soit inférieure à ε en déterminant N tel que

$$\varepsilon_N \leq \varepsilon$$

et donc, en remplaçant le test de l'algorithme B.1, par

$$\text{test} \longleftarrow ((\varepsilon_n > \varepsilon) \text{ et } (n < n_{\max}))$$

On peut enfin assurer que les erreurs relatives et absolues sont toutes les deux inférieures à $\varepsilon > 0$ en déterminant N tel que

$$\varepsilon_N \leq \min\left(\varepsilon, \frac{\varepsilon}{1+\varepsilon}|u_N|\right),$$

et donc, en remplaçant le test de l'algorithme B.1, par

$$\text{test} \longleftarrow \left(\left(\left(\varepsilon_n > \frac{\varepsilon}{1+\varepsilon}|u_n|\right) \text{ ou } \varepsilon_n > \varepsilon\right) \text{ et } (n < n_{\max})\right)$$

B.2. Simulations numériques

Présentons quelques simulations numériques utilisant les résulats du chapitre 1 page 2 et de cette annexe.

Nous reprenons les calculs du chapitre 1 pour présenter quelques calculs de l'entier N déterminé par l'algorithme B.1 page précédente. Nous calculerons aussi P le plus petit entier pour lequel l'erreur réelle est elle-même inférieure à ε. Nous choisissons, pour tout $x \in \mathbb{R}$, pour tout $n \in \mathbb{N}$:

$$u_n = p_n(x), \tag{B.8a}$$

$$l = e^x, \tag{B.8b}$$

et conformément à l'équation (1.14) page 4 :

$$\varepsilon_n = b_n - a_n, \tag{B.8c}$$

où a_n et b_n sont les approximations par défaut et par excés de e^x définis dans la remarque A.2 page 128.

Voir les figures B.1 page suivante et B.2 sur lesquelles on a tracé les indices N et P en fonction de ε correspondant d'abord au calcul de l'erreur relative, puis absolue pour deux valeurs différentes de x. Ce graphe est semi-logarithmique en abscisse. On constate bien sur ces graphiques que les entiers augmentent quand ε diminue.

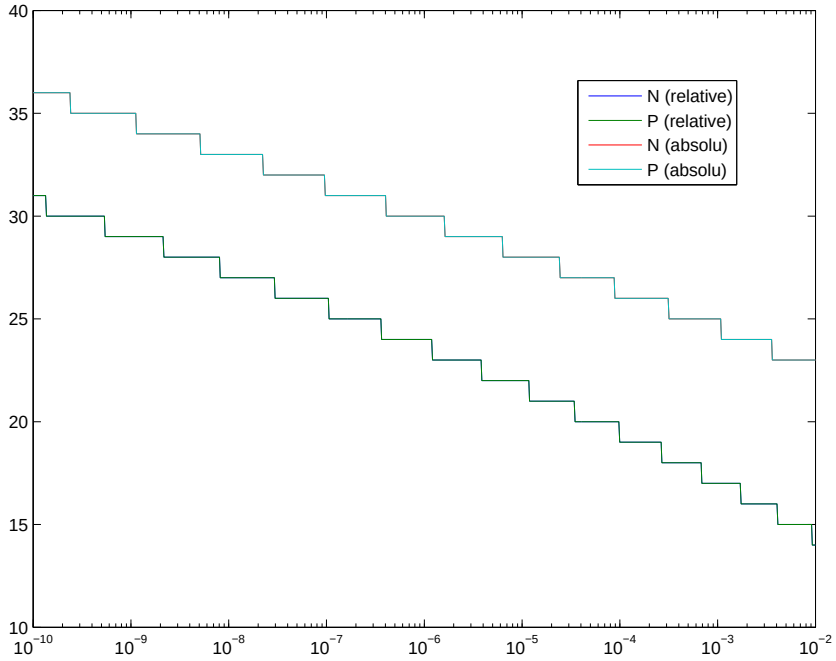


FIGURE B.1. Le graphe semi-logarithmique avec ε en abscisses et les indices N et P en ordonnées pour $x = 8$.

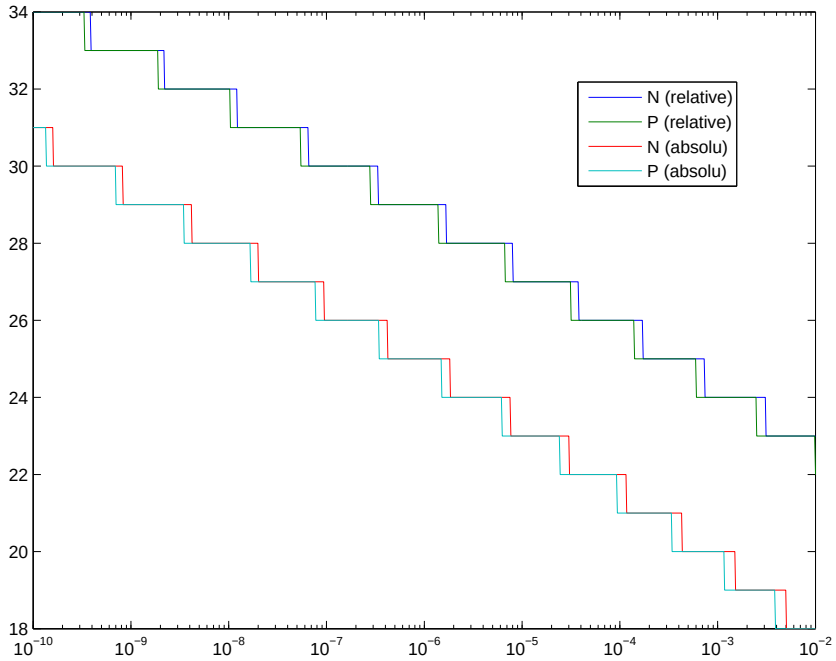


FIGURE B.2. Le graphe semi-logarithmique avec ε en abscisses et les indices N et P en ordonnées pour $x = -6$.

Annexe C

Différents calculs de l'exemple 2.21 page 29 du cours

(1) Si on utilise la méthode de la section 2.2.2.1, l'équation de la droite est donnée par

$$y = \alpha x + \beta$$

où, d'après l'équation (2.18)

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}1 & 1\\ x_A & x_B\end{pmatrix}^{-1}\begin{pmatrix}y_A\\ y_B\end{pmatrix}.$$

On a aisément

$$\begin{pmatrix}1 & x_A\\ 1 & x_B\end{pmatrix}^{-1} = \frac{1}{x_B - x_A}\begin{pmatrix}x_B & -1\\ -xA & 1\end{pmatrix}$$

et donc

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \frac{1}{x_B - x_A}\begin{pmatrix}x_B & -1\\ -xA & 1\end{pmatrix}\begin{pmatrix}y_A\\ y_B\end{pmatrix} = \frac{1}{x_B - x_A}\begin{pmatrix}x_B y_A - x_A y_B\\ y_B - yA\end{pmatrix},$$

et donc l'équation de la droite est

$$y = \frac{y_B - y_A}{x_B - x_A}x + \frac{x_B y_A - x_A y_B}{x_B - x_A}, \tag{C.1}$$

ce qui est bien l'équation (2.9) équivalente à l'équation (2.10).

(2) Si on utilise la méthode de la section 2.2.2.2, on utilise l'expression des polynômes de Lagranges l_0 et l_1 donnés par les équations (2.25b) et (2.25c), puis l'équation (2.27) qui donne

$$y = y_A l_0(x) + y_B l_1(x)$$

ce qui donne encore

$$y = y_A\frac{x - x_B}{x_A - x_B} + y_B\frac{x - x_A}{x_B - x_A},$$

ce qui bien équivalent à l'équation (C.1) et donc à l'équation (2.10).

(3) Enfin, si on utilise la méthode de la section 2.2.2.3, d'après les calculs de l'exemple 2.16, on a

$$y = y_A + \frac{y_B - y_A}{x_B - x_A}(x - x_A), \tag{C.2}$$

ce qui bien équivalent à l'équation (2.10).

(4) Lors d'un examen, une méthode originale a été proposée par Monsieur Mehdi Rosine à l'automne 2023 : Il s'agit de remarquer que la droite en question passe par le point $(x_A, f(x_A))$ et a pour coefficient directeur le taux d'acroissement de f entre x_A et x_B (qui n'est autre que la différence divisée $f[x_A, x_B] = \frac{y_B - y_A}{x_B - x_A}$). Si $y = g(x)$ désigne l'équation de la droite, on a donc

$$g(x_A) = y_A,$$
$$g'(x_A) = \frac{y_B - y_A}{x_B - x_A},$$

et la formule de Taylor appliquée à g donne, entre x et x_A, puisque $g'' = 0$:

$$\begin{aligned} y &= g(x), \\ &= g(x_A) + g'(x_A)(x - x_A) + \frac{1}{2}g''(\xi)(x - x_A), \\ &= y_A + \frac{y_B - y_A}{x_B - x_A}(x - x_A), \end{aligned}$$

et on retrouve donc bien l'équation (2.10).

REMARQUE C.1. C'est exactement l'idée de l'interpolation de Hermite ! Voir le début de la section 2.7 page 40 du cours.

Annexe D

Étude théorique d'un problème de moindres carrés

D.1. Rappels sur la régression linéaire

Les coordonnées $(x_i, y_i)_{1\leq i\leq n}$ étant connues (de façon expérimentale ou par mesure), on cherche donc à résoudre le problème suivant :

$$\text{trouver } (a,b) \text{ qui minimise } S = \sum_{i=1}^{n} (ax_i + b - y_i)^2. \tag{D.1}$$

La quantité S est appelé l'écart entre les données et la droite d'équation $Y = aX + b$. Ce problème s'écrit aussi : trouver le couple (a_0, b_0) tel que

$$\forall (a,b) \in \mathbb{R}^2, \quad \sum_{i=1}^{n} (a_0 x_i + b_0 - y_i)^2 \leq \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{D.2}$$

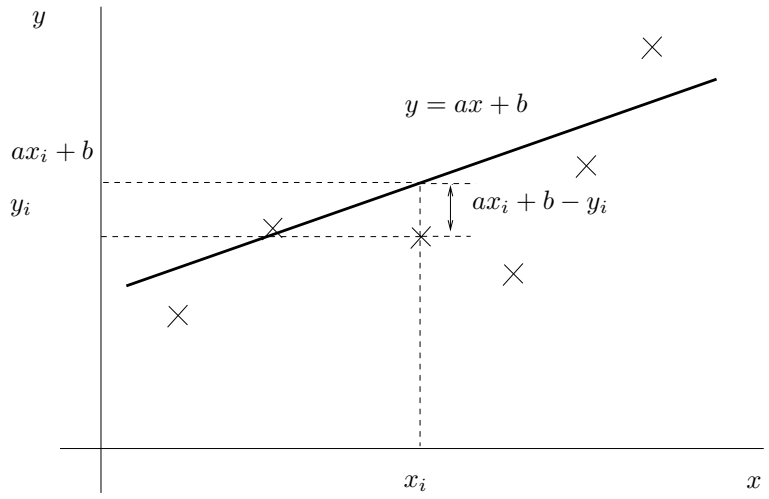


FIGURE D.1. le principe de la droite de régression linéaire

Voir la figure D.1.

On peut expliciter les coefficients a_0 et b_0 en fonction de $(x_i, y_i)_{1\leq i\leq n}$; voir par exemple les rubriques "régression linéaire" et "Ajustement affine" de Wikipédia (`https://fr.wikipedia.org/wiki/Régression_linéaire` et `https://fr.wikipedia.org/wiki/Ajustement_affine`).

REMARQUE D.1. Voir le calcul exact et la preuve dans l'exemple D.5.

Voir par exemple la figure D.2 page suivante, où sont tracés les points expérimentaux $(x_i, y_i)_{1\leq i\leq n}$, deux droites différentes correspondant à deux couples (a, b) avec les écart associés et la "meilleure droite". Sur cette figure,

— les points de coordonnées $(x_i, y_i)_{1\leq i\leq n}$ sont représentés par des carrés noirs ;

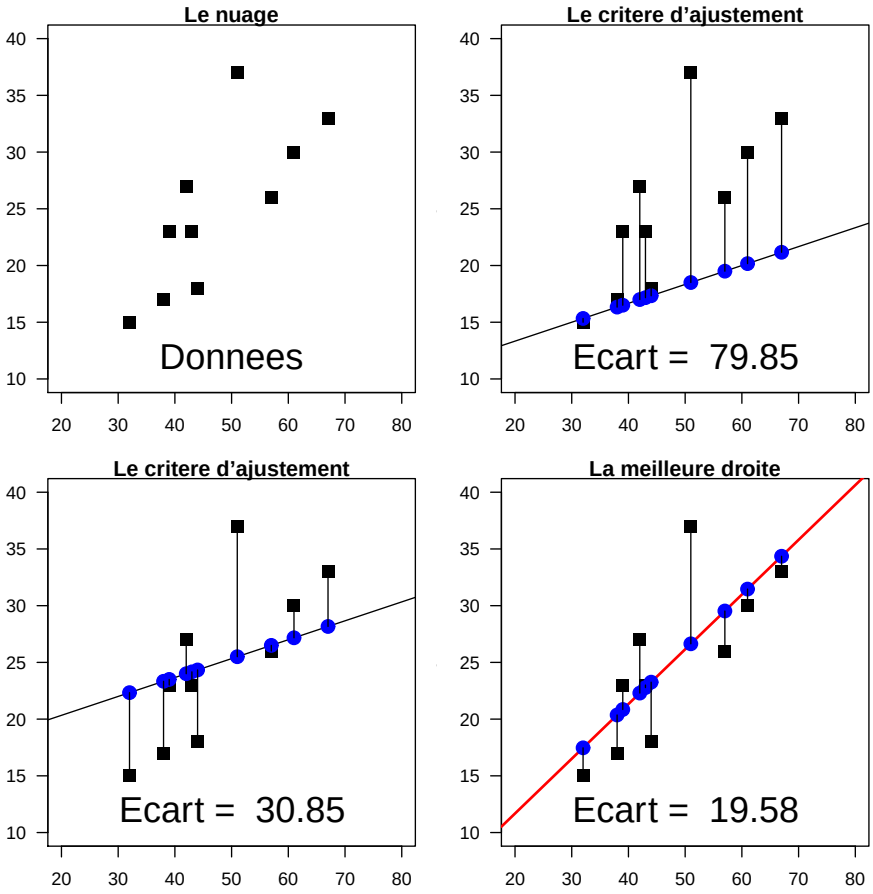


FIGURE D.2. la droite de régression linéaire

— les points de coordonnées $(x_i, ax_i + b)_{1\leq i\leq n}$ sont représentés par des ronds bleu ;
— deux droites sont tracées en noir et la "meilleure" en rouge. Cette droite a une pente a positive.
Cette figure, faite sous R, est extraite de [Bas12, chapitre 4].

D.2. Rappels sur la norme Euclidienne et le produit scalaire associé

On rappelle que $\|\|_2$ définit la norme Euclidienne usuelle sur $\mathbb{R}^p$:

$$\forall x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}, \quad \|x\|_2 = \sqrt{\sum_{k=1}^{p} x_k^2} = \sqrt{{}^t x x}, \tag{D.3}$$

où ${}^tC$ désigne la matrice transposée de la matrice C. Cette norme est associée au produit scalaire $< .,. >$ défini par

$$\forall x, y \in \mathbb{R}^p, \quad < x, y >= {}^t x y. \tag{D.4}$$

Pour un rappel sur les normes, on pourra consulter l'annexe A de [BM03] ou [Sch01].

D.3. Théorie

Pour plus d'information sur les systèmes au sens des moindres carrés, on pourra consulter [Cia82].

De façon plus générale, on se donne $n \geq p$ deux entiers non nuls, b un vecteur de $\mathbb{R}^n$, $A = (a_{i,j})_{\substack{1\leq i\leq n,\\ 1\leq j\leq p}}$ une matrice de $\mathcal{M}_{n,p}(\mathbb{R})$, on cherche à trouver le vecteur X qui vérifie le système sur-déterminé $AX = b$

$$\forall i \in \{1, ..., n\}, \quad \sum_{j=1}^{p} a_{i,j}x_j = b_i, \tag{D.5}$$

mais au sens des moindres carrés : on veut trouver X tel que la quantité

$$\sum_{j=i}^{n} \left(\sum_{j=1}^{p} a_{i,j}x_j - b_j \right)^2 \tag{D.6}$$

soit minimale. Par exemple, le cas de la section D.1, correspond à

$$p = 2, \tag{D.7a}$$

et

$$\forall i \in \{1, ..., n\}, \quad a_{i,1} = 1, \quad , a_{i,2} = x_i. \tag{D.7b}$$

La section D.1 traite le cas particulier de la de régression linéaire (c'est-à-dire, trouver la droite qui passe le plus proche d'un nuage de points donnés), avec (D.7).

De façon plus générale, on cherche donc à résoudre

$$\|Ax_0 - b\|_2^2 = \inf_{x\in\mathbb{R}^p} \|Ax - b\|_2^2, \tag{D.8}$$

où $n \geq p$ sont deux entiers non nuls, A est une matrice donnée de $\mathcal{M}_{n,p}(\mathbb{R})$, b est un vecteur donné de $\mathbb{R}^n$ et x_0 est l'inconnue dans $\mathbb{R}^p$. Voir la section D.2 pour les rappels sur la norme $\|.\|_2$. On écrit parfois que l'on résout le système «rectangulaire»

$$Ax = b, \tag{D.9}$$

au sens des moindres carrés. En effet, on a

$$Ax \approx b, \tag{D.10}$$

et c'est l'éccart entre Ax et b que l'on cherche à minimiser.

Lemme D.2. *Si la matrice A est une matrice de $\mathcal{M}_{n,p}(\mathbb{R})$ de rang p, alors ${}^tAA$ est inversible.*

Démonstration. Il suffit de vérifier que, pour tout vecteur $x \in \mathbb{R}^p$,

$${}^tAAx = 0 \Longrightarrow x = 0.$$

Si le vecteur de $\mathbb{R}^p$, ${}^tAAx$ est nul alors ${}^tx{}^tAAx$ est nul et donc, par définition

$$\|Ax\|_2 = {}^t(Ax)(Ax) = {}^tx{}^tAAx = 0,$$

et donc $Ax = 0$. Ainsi x appartient au noyau de A. D'après la relation rang-noyau, on a $\dim(\mathrm{Ker}A) = p - \mathrm{rg}(A) = 0$, par hypothèse. Ainsi, le noyau de A est réduit à $\{0\}$ et x est donc nul. □

Remarque D.3. Si les x_i ont au moins deux valeurs différentes, alors la matrice A donnée par (D.7), c'est-à-dire le cas de la section D.1, est de rang 2. En effet, on peut extraire de A, une sous-matrice de type

$$\begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix}$$

avec $x_i \neq x_j$ et l'on sait que cette matrice est inversible (voir par exemple la section 2.2.2.1).

On peut donc déduire du lemme D.2 le résultat suivant :

PROPOSITION D.4. *On suppose que la matrice A est de rang p. On note x_0 défini par*

$$x_0 = \left({}^tAA\right)^{-1}\left({}^tAb\right). \tag{D.11}$$

La solution de (D.8) *est unique et est égale à x_0.*

DÉMONSTRATION. Voir [Cia82].

Donnons une preuve directe de ce résultat. On renvoie par exemple à [Cia82, section 8.1]. On considère les espaces Euclidiens $E = \mathbb{R}^p$ et $F = \mathbb{R}^n$, munis de leur structure Euclidienne usuelle (voir section D.2). A définit donc une application linéaire de E dans F. Il suffit de remarquer que quand x décrit $E = \mathbb{R}^p$, $y = Ax$ décrit

$$H = \mathrm{Im}(A), \tag{D.12}$$

sous-espace vectoriel de F, l'image de A. Ainsi, en notant

$$z = Ax_0, \tag{D.13}$$

(D.8) est équivalent à

$$\|z - b\|_2^2 = \inf_{y \in H} \|y - b\|_2^2, \tag{D.14}$$

Il s'agit donc de minimiser la distance entre $b \in F$ et un vecteur y de H. C'est exactement le théorème de projection sur un

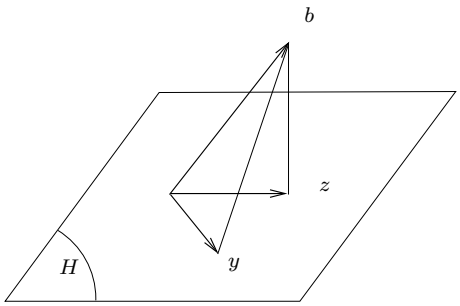


FIGURE D.3. La projection sur un sous-espace vectoriel.

sous-espace vectoriel (de dimension finie), qui est un cas particulier de [Cia82, Théorème 8.1]. Voir la figure D.3. On sait que la solution de (D.14) est unique et est donnée par $z = p_H(b)$, projection orthogonale de b sur H. (D.14) est donc équivalente à

$$z \in H,$$
$$b - z \in H^{\perp},$$

autrement dit, par définition de H ((D.12)), on a (D.13) et, $b - Ax_0$ est perpendiculaire à pour tout vecteur de H. Puisque H est donné par (D.12), il suffit d'écrire que $b - Ax_0$ est perpendiculaire à tous les vecteurs colonnes de A, soit encore, compte tenu de la section D.2 (le produit scalaire est donné par (D.4)), on a

$$\forall i \in \{1, ..., p\}, \quad {}^tA_i\,(b - Ax_0) = 0,$$

où A_i est le i-ième vecteur colonne de A, soit encore

$${}^tA\,(b - Ax_0) = \begin{pmatrix} {}^tA_1\,(b - Ax_0) \\ {}^tA_2\,(b - Ax_0) \\ \vdots \\ {}^tA_p\,(b - Ax_0) \end{pmatrix} = 0,$$

dans $\mathbb{R}^p$. On a donc établi les équations dites normales

$${}^tAAx_0 = {}^tAb. \tag{D.15}$$

Le vecteur x_0 existe puisqu'il est défini par $Ax_0 = z$ où z est unique. De plus, d'après le lemme D.2, il est donc unique et donné par (D.11). ♢ □

Exemple D.5. Utilisons ces résultats pour déterminer la droite aux moindres carrés donnés dans la section D.1 :

La résolution des n équations données dans (D.9) au sens des moindres carrés :

$$\forall i \in \{1, ..., n\}, \quad ax_i + b = y_i,$$

est donnée par (D.11) avec

$$x_0 = \begin{pmatrix} b \\ a \end{pmatrix}, \tag{D.16a}$$

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \tag{D.16b}$$

$$b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{D.16c}$$

On a

$${}^tAA = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}, \tag{D.17a}$$

$${}^tAb = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{D.17b}$$

En utilisant le fait que

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{-1} = \frac{1}{\alpha\delta - \beta\gamma} \begin{pmatrix} \delta & -\beta \\ -\gamma & \gamma \end{pmatrix}, \tag{D.18}$$

les équations (D.11), (D.16) et (D.17) sont équivalentes à

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{D.19a}$$

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}. \tag{D.19b}$$

Utilisons les notations classiques : $\overline{x}$ est la moyenne des abscisses, $\overline{y}$ est la moyenne des ordonnées :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{D.20a}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{D.20b}$$

$V(x)$ est la variance des abscisses :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2, \tag{D.20c}$$

$\mathrm{Cov}(x, y)$ est la covariance des couples de coordonnées :

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{D.20d}$$

Montrons que (D.19) est équivalent à

$$a = \frac{\mathrm{Cov}(x, y)}{V(x)}, \tag{D.21a}$$

$$b = \overline{y} - a\overline{x}. \tag{D.21b}$$

On remarque alors que si les x_i sont différents (au moins deux valeurs distinctes), alors

$$V(x) \neq 0, \tag{D.22}$$

ce qui montre que la matrice ${}^tAA$ est inversible (puisque A est de rang 2 dans ce cas-là). Pour montrer (D.21), il est plus aisé de reprendre le système (D.11), où l'on exprime les différentes sommes à partir des grandeurs introduites dans (D.20). Remarquons que

$$\sum_{i=1}^{n} x_i = n\overline{x}, \tag{D.23a}$$

$$\sum_{i=1}^{n} y_i = n\overline{y}. \tag{D.23b}$$

On a aussi successivement

$$\begin{aligned}
V(x) &= \frac{1}{n}\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right), \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x} + \overline{x}^2\right), \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2\right), \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2\right), \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right),
\end{aligned}$$

dont on déduit que

$$\sum_{i=1}^{n} x_i^2 = nV(x) + n\overline{x}^2. \tag{D.24}$$

On a, de même,

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right),\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - \overline{x}\sum_{i=1}^{n}y_i - \overline{y}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\overline{xy}\right),\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - 2n\overline{xy} + n\overline{xy}\right),\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_iy_i - n\overline{xy}\right),
\end{aligned}$$

dont on déduit que

$$\sum_{i=1}^{n} x_iy_i = n\mathrm{Cov}(x,y) + n\overline{xy}. \tag{D.25}$$

Compte tenu de ces résultats, (D.17) se réécrit

$${}^tAA = n\begin{pmatrix}1 & \overline{x}\\ \overline{x} & V(x)+\overline{x}^2\end{pmatrix}, \tag{D.26a}$$

$${}^tAb = n\begin{pmatrix}\overline{y}\\ \mathrm{Cov}(x,y)+\overline{xy}\end{pmatrix}. \tag{D.26b}$$

Ainsi, grâce à (D.18), les équations (D.11) et (D.16) sont équivalentes à

$$\begin{pmatrix}b\\a\end{pmatrix} = \frac{1}{V(x)}\begin{pmatrix}V(x)+\overline{x}^2 & -\overline{x}\\ -\overline{x} & 1\end{pmatrix}\begin{pmatrix}\overline{y}\\ \mathrm{Cov}(x,y)+\overline{xy}\end{pmatrix}$$

ce qui donne, après calculs, (D.21).

Moins généralement, mais beaucoup plus rapidement, comme c'est fait dans `https://fr.wikipedia.org/wiki/Ajustement_affine` et dans [Mer10] on peut écrire : Pour a donné, on considère f_a définie par

$$f_a(b) = \sum_{i=1}^{n}(ax_i + b - y_i)^2 \tag{D.27}$$

comme une fonction du second degré en b dont on peut déterminer le minimum, puis, ce b étant exprimé par sa valeur en fonction de f, trouver le minimum de la fonction du second degré en a. On écrit donc successivement

$$\begin{aligned}
f_a(b) &= \sum_{i=1}^{n}(b + (ax_i - y_i))^2,\\
&= \sum_{i=1}^{n} b^2 + 2b(ax_i - y_i) + (ax_i - y_i)^2.
\end{aligned}$$

On a donc

$$f_a(b) = nb^2 + 2b\left(a\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}y_i\right) + n(ax_i - y_i)^2. \tag{D.28}$$

Grâce à (D.23), (D.28) s'écrit

$$f_a(b) = n\left(b^2 + 2b\left(a\overline{x} - \overline{y}\right) + n(ax_i - y_i)^2\right), \tag{D.29}$$

proportionnel à un polynôme en b de la forme Ab^2+Bb+C dont la valeur minimale est atteinte en $b=-B/(2A)$ (racine de dérivée de ce polynôme). On a donc

$$b = -\frac{1}{2}\left(2\left(a\overline{x}-\overline{y}\right)\right),$$

soit

$$b = \overline{y} - a\overline{x}. \tag{D.30}$$

Si maintenant on remplace, dans $f_a(b)$ (donnée par (D.27)), b par sa valeur en fonction de a, on obtient une fonction du second degré en a :

$$\begin{aligned} f_a(b) &= \sum_{i=1}^{n}\left(ax_i+\overline{y}-a\overline{x}-y_i\right)^2, \\ &= \sum_{i=1}^{n}\left(\left(y_i-\overline{y}\right)-a\left(x_i-\overline{x}\right)\right)^2, \\ &= \sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 + a^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2 - 2a\sum_{i=1}^{n}\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right), \end{aligned}$$

Grâce à (D.20c) et (D.20d), c'est égal à

$$= \sum_{i=1}^{n} V(y) + na^2(V(x))^2 - 2an\mathrm{Cov}(x,y),$$

soit

$$g(a) = n\left(V(x)a^2 - 2\mathrm{Cov}(x,y)a + V(y)\right), \tag{D.31}$$

donc proportionnel à un polynôme du second degré de la forme Ua^2+Vb+c. D'après (D.22), $U=V(x)$ est non nul. Ce polynôme atteint son minimum lorsque a est égal à $-V/(2U)$ donc lorsque

$$a = \frac{\mathrm{Cov}(x,y)}{V(x)}. \tag{D.32}$$

Grâce à (D.30) et (D.32), on a donc retrouvé (D.21).

La qualité de l'ajustement affine est alors mesurée par le coefficient de corrélation linéaire r donné par

$$r = \frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)V(y)}} = \frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{D.33}$$

où σ_x et σ_y sont les écarts types des deux variables statistiques. Il intervient dans l'évaluation de la somme des carrés des résidus :

$$S = nV(Y)(1-r^2). \tag{D.34}$$

En effet, si on reprend le calcul fait plus haut et que l'on réinjecte la valeur de a donnée par (D.32) dans l'expression donnée par (D.31), on obtient la valeur minimale de la somme donnée par

$$\begin{aligned} S &= n\left(V(x)\left(\frac{\mathrm{Cov}(x,y)}{V(x)}\right)^2 - 2\mathrm{Cov}(x,y)\left(\frac{\mathrm{Cov}(x,y)}{V(x)}\right) + V(y)\right), \\ &= n\left(-\frac{(\mathrm{Cov}(x,y))^2}{V(x)} + V(y)\right), \\ &= nV(y)\left(1 - \left(\frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)V(y)}}\right)^2\right), \end{aligned}$$

et on obtient donc (D.34). D'après (D.34), plus le coefficient de corrélation est proche de ± 1, et plus la somme des carrés des résidus est voisine de 0. Le coefficient de corrélation est donc un bon indicateur de la validité de l'ajustement affine. Enfin, l'Inégalité de Cauchy-Schwarz permet d'affirmer que

$$\left|\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right| \leq \left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^{1/2}\left(\sum_{i=1}^n (y_i - \overline{y})^2\right)^{1/2} \iff |\mathrm{Cov}(x,y)| \leq \sqrt{V(x)}\sqrt{V(y)} \iff |r| \leq 1$$

avec égalité seulement si les $y_i - \bar{y}$ sont proportionnels aux $x_i - \bar{x}_i$. Donc $|r| \leq 1$ avec égalité seulement s'il existe un réel a tel que,

$$\forall i \in \{1, ..., n\}, \quad y_i - \bar{y} = a\,(x_i - \bar{x}), \tag{D.35}$$

ce qui traduit l'alignement exact des points. Ainsi, plus $|r|$ est proche de 1 plus l'ajustement affine semble légitime.

Concluons par quelques exemples numériques.

Remarque D.6. On pourra refaire ces simulations grâce à la fonction `polyfit.m` de matlab, comme cela sera proposé dans le TP 2.2.2.

(1)

x_i	0	1	2	3
y_i	$-9/2$	-1	3	4

Table D.1. valeurs des couples (x_i, y_i)

On choisit $n = 4$, les couples (x_i, y_i) sont donnés dans le tableau D.1. Les matrices ${}^tAA$ et ${}^tAb$ données par (D.17a) et (D.17b) sont données par

$${}^tAA = \begin{pmatrix} 4 & 6 \\ 6 & 14 \end{pmatrix},$$

$${}^tAb = \begin{pmatrix} 3/2 \\ 17 \end{pmatrix},$$

et la solution du système (D.11) est donnée par

$$X = \begin{pmatrix} -\frac{81}{20} \\ \frac{59}{20} \end{pmatrix},$$

et donc, compte tenu de (D.16a), l'équation de la droite de régression linéaire est donnée par

$$y = \frac{59}{20}\,x - \frac{81}{20}.$$

La somme des carrés des écarts est donnée par

$$S = 2.1750. \tag{D.36}$$

La valeur du coefficient de corrélation linéaire donnés par (D.33) est ici égale à

$$r = 0.9759, \tag{D.37}$$

ce qui donne des points modérémment alignés. Enfin, sur la figure D.4 sont représentés les couples

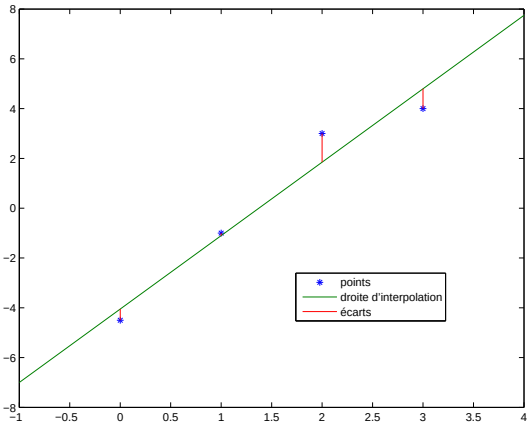


FIGURE D.4. Les couples (x_i, y_i) et la droite de correlation linéaire pour l'exemple 1.

(x_i, y_i) et la droite aux moindres carrés. On a aussi tracé les écarts (dont la somme constitue la somme S donnée par (D.36)) entre la droite et les points.

(2)

x_i	0	1	2	3
y_i	-4	-1	2	5

TABLE D.2. valeurs des couples (x_i, y_i)

On choisit $n = 4$, les couples (x_i, y_i) sont donnés dans le tableau D.2. Les matrices ${}^tAA$ et ${}^tAb$ données par (D.17a) et (D.17b) sont données par

$${}^tAA = \begin{pmatrix} 4 & 6 \\ 6 & 14 \end{pmatrix},$$

$${}^tAb = \begin{pmatrix} 2 \\ 18 \end{pmatrix},$$

et la solution du système (D.11) est donnée par

$$X = \begin{pmatrix} -4 \\ 3 \end{pmatrix},$$

et donc, compte tenu de (D.16a), l'équation de la droite de régression linéaire est donnée par

$$y = 3\,x - 4.$$

La somme des carrés des écarts est donnée par

$$S = 0, \tag{D.38}$$

ce qui est normal puisqu'on va voir que les point sont alinés. La valeur du coefficient de corrélation linéaire donnés par (D.33) est ici égale à

$$r = 1, \tag{D.39}$$

ce qui donne des points exactement alignés. Enfin, sur la figure D.5 sont représentés les couples (x_i, y_i)

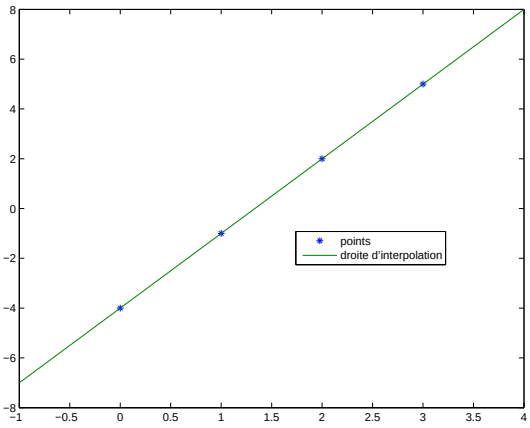


FIGURE D.5. Les couples (x_i, y_i) et la droite de corrélation linéaire pour l'exemple 2.

et la droite aux moindres carrés.

(3)

On choisit cette fois-ci $N = 100$, les couples (x_i, y_i) sont donnés par la formule suivante :

$$\forall i \in \{1, ..., N\}, \quad x_i = \frac{i-1}{N-1}, \tag{D.40a}$$

$$y_i = a x_i + b + \varepsilon_i \mathcal{N}_i, \tag{D.40b}$$

où les $\mathcal{N}_i$, pour $1 \leq i \leq q$, avec $q = 6$, sont des nombres aléatoires issus d'une distribution normale de moyenne 0 et d'éccart-type 1. Les nombres ε_i sont donnés dans le tableau D.3. On choisit

i	1	2	3	4	5	6
ε_i	0	$1.0\,10^{-6}$	$1.0\,10^{-1}$	2.0	$1.0\,10^{3}$	$1.0\,10^{7}$

TABLE D.3. valeurs des nombres ε_i

$$a = 2, \tag{D.41a}$$

$$b = 10. \tag{D.41b}$$

Les valeurs absolues des coefficients de corrélation linéaire $|r_i|$, pour $1 \leq i \leq q$, donnés par (D.33) sont

i	1	2	3	4	5	6
r_i	1.0	$10.0\,10^{-1}$	$9.879876773\,10^{-1}$	$2.963927220\,10^{-1}$	$5.506795276\,10^{-2}$	$5.561757702\,10^{-2}$

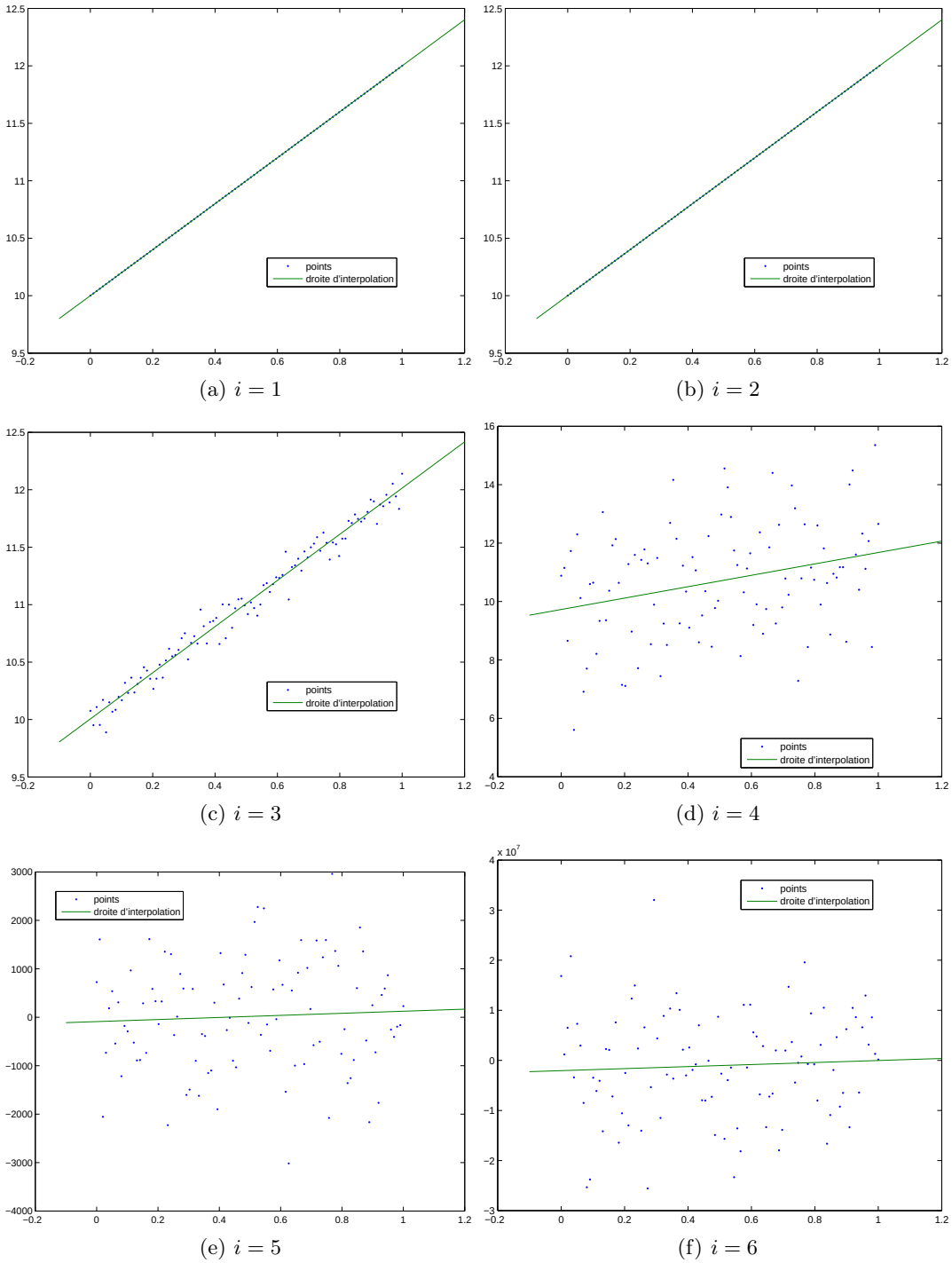
TABLE D.4. valeurs des nombres $|r_i|$

donnés dans le tableau D.4. Les valeurs des pentes a_i et des ordonnées à l'origne b_i, pour $1 \leq i \leq q$, sont

i	1	2	3	4	5	6
a_i	2.0	1.999999753	2.9181815	1.950935255	$2.136121422\,10^{2}$	$2.13910010\,10^{6}$
b_i	$1.0\,10^{1}$	$1.6\,10^{1}$	$1.521218\,10^{1}$	9.725791486	$-8.941009407\,10^{1}$	$-2.44242058\,10^{6}$

TABLE D.5. valeurs des nombres a_i et b_i

donnés dans le tableau D.5. Enfin, sur les figures D.6 sont représentés les couples (x_i, y_i) et la droite aux moindres carrés. On constate que, $\varepsilon_i = 0$, pour $i = 1$ et les points sont alignés ($r_1 = 1$) et que les coefficients a_1 et b_1 sont égaux aux coefficients a et b. Quand $i \geq 2$ augmente, ε_i augmente, d'une petite quantité à une grande quantité et plus ε_i est grand, moins les points sont alignés, moins les coefficients a_i et b_i sont proches de a et b (voir tableau D.5) et le nuage de points semble désordonné. De plus, les valeurs absolues des coefficients de corrélation $|r_i|$, proche de 1 quand ε_i est petit se rapproche de zéro quand ε_i augmente (voir tableau D.4). Cela est normal, puisque avec i augmentant, ε_i augmente et l'effet du hasard (donné par la formule (D.40b)) est de plus en plus important.



(a) $i = 1$ (b) $i = 2$

(c) $i = 3$ (d) $i = 4$

(e) $i = 5$ (f) $i = 6$

FIGURE D.6. Les couples (x_i, y_i) et la droite aux moindres carrés, pour $1 \leq i \leq q$.

Annexe E

Définition et utilisation de la fonction W de Lambert

On pourra consulter par exemple `https://fr.wikipedia.org/wiki/Fonction_W_de_Lambert`

E.1. Définition de la fonction W de Lambert

On cherche à résoudre l'équation suivante : pour $z \in \mathbb{R}$ donné, on cherche $w \in \mathbb{R}$ tel que

$$we^w = z. \tag{E.1}$$

On a alors le résultat suivant :

PROPOSITION E.1 (les deux branches W_0 et W_{-1} de la "fonction"de Lambert).

- *Si $z < -\frac{1}{e}$, il n'existe aucun $w \in \mathbb{R}$ solution de* (E.1) *;*
- *Il existe une unique fonction W_0 de $[-\frac{1}{e}, +\infty[$ dans $[-1, +\infty[$ telle que* (E.1) *est équivalent à $w = W_0(z)$;*
- *Il existe une unique fonction W_{-1} de $[-\frac{1}{e}, 0[$ dans $]-\infty, -1]$ telle que* (E.1) *est équivalent à $w = W_{-1}(z)$.*

Cela est équivalent à dire : Pour tout $z \in \mathbb{R}$:

- *Si $z < -\frac{1}{e}$, l'équation* (E.1) *n'a pas de solution ;*
- *Si $z = -\frac{1}{e}$, la seule solution de* (E.1) *est donnée par*

$$w = W_{-1}\left(-\frac{1}{e}\right) = W_0\left(-\frac{1}{e}\right) = -1; \tag{E.2}$$

- *Si $z \in]-\frac{1}{e}, 0[$, les deux solutions distinctes de* (E.1) *sont données par*

$$w = W_{-1}(z), \tag{E.3a}$$

$$w = W_0(z); \tag{E.3b}$$

- *Si $z \in [0, +\infty[$, l'unique solution de* (E.1) *est donnée par*

$$w = W_0(z). \tag{E.4}$$

DÉMONSTRATION. Il suffit d'étudier la fonction f définie sur $\mathbb{R}$ par

$$\forall w \in \mathbb{R}, \quad f(w) = we^w, \tag{E.5}$$

dont la dérivée est donné par

$$\forall w \in \mathbb{R}, \quad f'(w) = (1 + w)e^w, \tag{E.6}$$

strictement positive sur $]-1, +\infty[$ et strictement négative sur $]-\infty, -1[$. On en déduit le tableau de variation de f, que l'on complète grâce à

$$\begin{aligned}
&\lim_{w \to -\infty} f(w) = 0,\\
&\lim_{w \to +\infty} f(w) = +\infty,\\
&f(-1) = -\frac{1}{e},\\
&f(0) = 0.
\end{aligned}$$

x	$-\infty$		-1		0		$+\infty$
signe de $v'(x)$		$-$	0	$+$			
variations de v	0	↘	$-\frac{1}{e}$	↗	0	↗	$+\infty$

TABLE E.1. Tableau de variation de f

Voir le tableau E.1 et les figure 1(a) et 1(b).

Compte tenu du tableau E.1, on constate que

— l'image de $\mathbb{R}$ par f est égale à $[-1/e, +\infty[$;

— f est strictement décroissante sur $]-\infty, -1]$ et définit donc une bijection de $]-\infty, -1]$ sur $f(]-\infty, -1]) = [-1/e, 0[$;

— f est strictement croissante sur $[-1, +\infty[$ et définit donc une bijection de $[-1, +\infty[$ sur $f([-1, +\infty[) = [-1/e, +\infty[$.

Voir les figures 1(c) et 1(d). On peut donc conclure à partir de ces éléments. □

E.2. Utilisation de la fonction W de Lambert : résolution de l'équation $ae^x + bx + c = 0$

LEMME E.2. *Soient a, b et c trois réels, a et b étant non nuls. On pose*

$$\Delta = \frac{a}{b} e^{-\frac{c}{b}}. \tag{E.7}$$

L'équation

$$ae^x + bx + c = 0, \quad x \in \mathbb{R}, \tag{E.8}$$

admet

- *aucune solution si $\Delta < -\frac{1}{e}$;*
- *une unique solution x si $\Delta = -\frac{1}{e}$, donnée par*

$$x = -\frac{c}{b} + 1, \tag{E.9}$$

- *deux solutions deux à deux distinctes x_0 et x_{-1} si Δ appartient à $]-\frac{1}{e}, 0[$, données par*

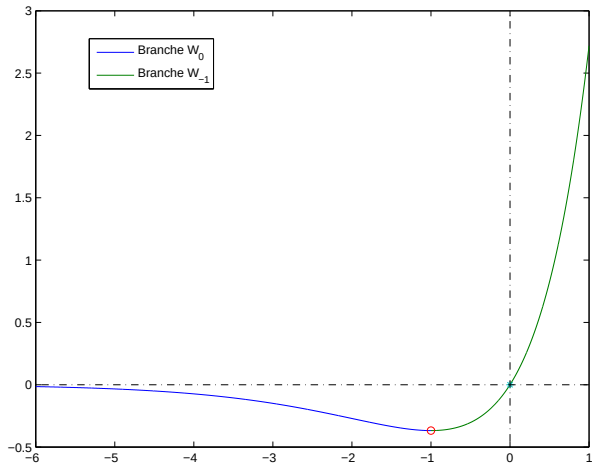
$$x_k = -\frac{c}{b} - W_k(\Delta), \quad \forall k \in \{0, -1\}, \tag{E.10}$$

- *une unique solution x si Δ appartient à $[0, +\infty[$, donnée par*

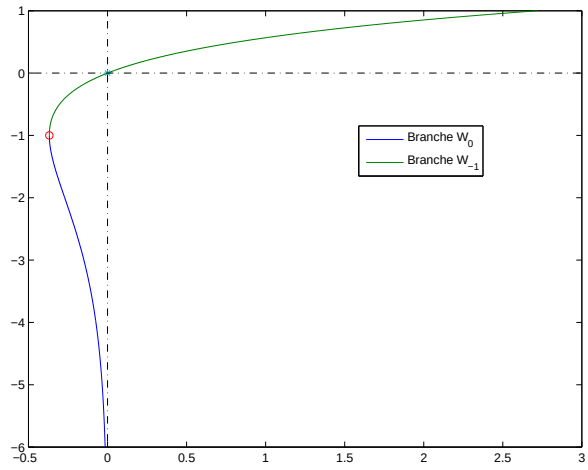
$$x = -\frac{c}{b} - W_0(\Delta). \tag{E.11}$$

DÉMONSTRATION. L'équation (E.8) est successivement équivalente à (puisque a et b sont non nuls)

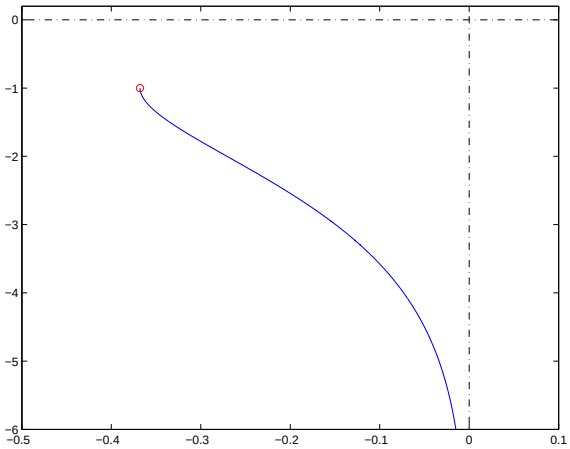
$$\begin{aligned} \frac{a}{b} e^x = -x - \frac{c}{b} &\iff \left(-x - \frac{c}{b}\right) \frac{b}{a} e^{-x} = -1, \\ &\iff \left(-x - \frac{c}{b}\right) e^{-x} e^{-\frac{c}{b}} = \frac{a}{b} e^{-\frac{c}{b}}, \end{aligned}$$



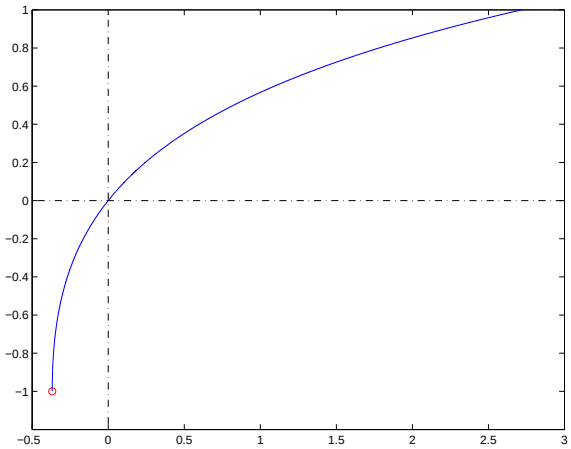
(a) La fonction f et les deux branches W_0 (en vert) et W_{-1} (en bleu).



(b) La fonction f^{-1} et les deux branches W_0 (en vert) et W_{-1} (en bleu).



(c) la branche W_{-1}.



(d) la branche W_0.

FIGURE E.1. La fonction f et les deux branches W_0 et W_{-1}.

cela est équivalent, en posant

$$X = -x - \frac{c}{b},$$
$$\Delta = \frac{a}{b} e^{-\frac{c}{b}},$$

à

$$Xe^X = \Delta. \tag{E.12}$$

Il ne reste plus qu'à appliquer les résultats de la proposition E.1, qui donne X en fonction de $W_k(\Delta)$ et donc x de la forme (selon les valeurs de Δ) :

$$x = -\frac{c}{b} - W_k(\Delta).$$

□

E.3. Utilisation de la fonction W de Lambert : résolution de l'équation $x^x = z$

On cherche à résoudre l'équation suivante : pour $z \in \mathbb{R}$ donné, on cherche $x \in \mathbb{R}$ tel que

$$x^x = z. \tag{E.13}$$

Si x existe, on a nécessairement

$$z > 0. \tag{E.14}$$

L'équation (E.13) est équivalente à

$$e^{x \ln x} = z,$$

soit encore

$$x \ln x = \ln z,$$

et donc

$$\ln x e^{\ln x} = \ln z,$$

et donc encore

$$X e^X = \ln z. \tag{E.15}$$

où

$$X = \ln x. \tag{E.16}$$

D'après la section E.1, cette équation a une solution supérieur à -1 unique, si $\ln(z) \geq -1/e$ et donc si

$$z \geq e^{-\frac{1}{e}}. \tag{E.17}$$

Cette solution est donnée par

$$X = W(\ln z),$$

et, d'après (E.16)

$$x = e^{W(\ln z)},$$

et d'après l'équation (E.1), on a finalement

$$x = \frac{\ln z}{W(\ln z)} \geq \frac{1}{e}. \tag{E.18}$$

Annexe F

La primitive est l'opération inverse de la dérivation (sous forme d'exercice)

Nous proposons dans cette annexe de montrer de façon un peu rigoureuse, que l'opération inverse de l'intégration est bien la dérivée., c'est-à-dire que la dérivée de la primitive et une primitive de la dérivée coïncident avec l'identité.

Énoncé

Soit v une fonction continue sur un intervalle $[t_0, T]$.
On pose

$$\forall t \in [t_0, T], \quad x(t) = x_0 + \int_{t_0}^{t} v(s)ds.$$

(1) Montrer que

$$\frac{x(t+h) - x(t)}{h} = \frac{1}{h}\int_{t}^{t+h} v(s)ds.$$

(2) On suppose que la fonction v est continue en t, ce qui s'écrit

$$v(s) = v(t) + \varepsilon(s),$$

avec

$$\lim_{s \to t} \varepsilon(s) = 0.$$

En déduire, en passant à la limite quand h tend vers zéro que

$$x'(t) = v(t)$$

(3) En déduire aussi que si on pose

$$\forall t \in [t_0, T], \quad V(t) = \int_{t_0}^{t} v'(s)ds,$$

alors

$$V(t) = v(t) - v(t_0).$$

Corrigé

(1) Par définition

$$\begin{aligned}
\frac{x(t+h) - x(t)}{h} &= \frac{1}{h}\left(x_0 + \int_{t_0}^{t+h} v(s)ds - x_0 - \int_{t_0}^{t} v(s)ds\right), \\
&= \frac{1}{h}\left(\int_{t_0}^{t+h} v(s)ds + \int_{t}^{t_0} v(s)ds\right), \\
&= \frac{1}{h}\int_{t}^{t+h} v(s)ds.
\end{aligned}$$

(2) On en déduit que

$$\begin{aligned}\frac{x(t+h)-x(t)}{h} &= \frac{1}{h}\int_t^{t+h} v(t)+\varepsilon(s)ds,\\ &= \frac{v(t)}{h}\int_t^{t+h} ds + \frac{1}{h}\int_t^{t+h}\varepsilon(s)ds,\\ &= v(t)+\frac{1}{h}\int_t^{t+h}\varepsilon(s)ds,\end{aligned}$$

et donc que

$$\left|\frac{x(t+h)-x(t)}{h}-v(t)\right| = \left|\frac{1}{h}\int_t^{t+h}\varepsilon(s)ds\right| \qquad \leq \frac{1}{h}\max_{s\in[t,t+h]}|\varepsilon(s)|\int_t^{t+h} ds \leq \max_{s\in[t,t+h]}|\varepsilon(s)|$$

Quand h tend vers zéro, cette dernière quantité tend vers zéro, ce qui nous permet de conclure.

(3) D'après ce qui précède, on a

$$V't) = v'(t),$$

donc $V - v$ est constant et donc

$$V(t) = v(t) + C.$$

On conclue en évaluant cela en t_0 :

$$V(t_0) - v(t_0) = C,$$

et donc

$$V(t) = v(t) + V(t_0) - v(t_0) = v(t) - v(t_0).$$

Annexe G

Calcul d'une intégrale "impropre" par la méthode de newton (sous forme d'exercice corrigé)

Cet exercice a été donné à l'examen de MNB à l'Automne 2023. Il constitue l'exercice 3.3 de [BM03], légèrement adapté pour ce cours.

Énoncé de l'exercice

On se propose, dans cet exercice, de montrer que l'intégrale

$$I = \int_0^{+\infty} e^{-x^2} dx \tag{G.1}$$

est convergente et d'en calculer une valeur approchée par la méthode de Simpson à ε près ($\varepsilon \in \mathbb{R}_+^*$ fixé).

(1) Peut-on calculer directement I par la méthode de Simpson ? Pourquoi ?

(2) Soit A appartenant à $[1, +\infty[$. On veut évaluer

$$I_A = \int_0^{A} e^{-x^2} dx$$

à $\varepsilon/2$ près.

(a) Soit la fonction g donnée par

$$g(x) = e^{-x^2} p(x), \tag{G.2}$$

où p est un polynôme quelconque.

Montrer par récurrence sur $n \geq 0$ la propriété suivante : il existe un polynôme p_n tel que la dérivée n-ième de g vérifie

$$\forall x \in \mathbb{R}, \quad g^{(n)}(x) = e^{-x^2} p_n(x), \tag{G.3}$$

où les polynômes $(p_n)_{n\in\mathbb{N}}$ vérifient

$$p_0 = p, \tag{G.4}$$

$$\forall n \in \mathbb{N}, \quad p_{n+1}(x) = p_n'(x) - 2x p_n(x). \tag{G.5}$$

(b) En déduire la dérivée quatrième $f^{(4)}$ de la fonction f définie par $f(x) = e^{-x^2}$.

(c) Déterminer un majorant M de $\left|f^{(4)}\right|$ sur $[0, A]$.
On pourra admettre que, pour tout A, $M = 12$.

(d) En déduire, en fonction de A, M et de ε, le pas maximal $h_{\max}$ autorisé en méthode de Simpson pour que l'erreur d'intégration commise soit de valeur absolue inférieure à $\varepsilon/2$.

(3) Convergence de I et choix de A.

Soit X un réel vérifiant $X \geq A$ et A un élément de $[1, +\infty[$.

(a) On pose

$$R(X) = \int_A^X e^{-x^2} dx.$$

Montrer que $R(X)$ est défini et que

$$R(X) \leq \int_A^X e^{-x} dx. \tag{G.6}$$

(b) Calculer

$$b(X) = \int_A^X e^{-x} dx \text{ et } l = \lim_{X \to +\infty} b(X) \text{ en fonction de } A.$$

(c) Montrer que

$$\forall X \geq A, \quad R(X) \leq e^{-A}. \tag{G.7}$$

(d) En déduire que I est convergente et que

$$\int_A^{+\infty} e^{-x^2} dx \leq e^{-A}. \tag{G.8}$$

(e) Déterminer en fonction de ε une valeur de A telle que

$$\int_A^{+\infty} e^{-x^2} dx \leq \frac{\varepsilon}{2}. \tag{G.9}$$

(4) Bilan

On donne $\varepsilon = 10^{-4}$.

(a) Déterminer A issu de la question 3e.

(b) Déterminer le pas maximal h_{max} issu des questions 2c et 2d.

(c) Conclure sur le nombre de sous-intervalles à considérer pour évaluer I à 10^{-4} près.

Corrigé de l'exercice

Attention, à l'erreur présente [BM03, p. 263] : remplacer les équations (6.81) et la précédente par : ... il vient

$$p_4(x) = 16x^4 - 48x^2 + 12.$$

On en déduit donc

$$f^{(4)}(x) = 4\left(4x^4 - 12x^2 + 3\right) e^{-x^2}. \tag{G.10}$$

(1) L'intégrale met en œuvre l'intervalle $[0, +\infty[$, non borné et, par conséquence, la méthode de Simpson n'est pas applicable.

(2) (a) On note, pour $n \in \mathbb{N}$, $\mathcal{P}_n$ la propriété suivante (pour g définie par (G.2)) : g est n fois dérivable sur $\mathbb{R}$ et il existe une fonction polynôme p_n vérifiant (G.3). Démontrons par récurrence $\mathcal{P}_n$ pour $n \in \mathbb{N}$. Nous démontrerons en même temps les propriétés (G.4) et (G.5). La propriété $\mathcal{P}_0$ est évidente (ainsi que (G.4))

Supposons $\mathcal{P}_n$ vraie et démontrons $\mathcal{P}_{n+1}$. On sait donc que g est de classe C^n. Puisque p_n est un polynôme, la fonction $x \mapsto e^{-x^2} p_n(x)$ est dérivable et, d'après (G.3), $g^{(n)}$ est donc dérivable ; ainsi g est de classe C^{n+1} et

$$\forall x \in \mathbb{R}, \quad g^{(n+1)}(x) = \left(g^{(n)}\right)'(x) = \left(e^{-x^2} p_n(x)\right)' = e^{-x^2}\left(p_n'(x) - 2x p_n(x)\right).$$

Ainsi, (G.3) est vraie à l'ordre $n+1$ à condition de définir le polynôme p_{n+1} par (G.5). On en déduit donc que g est de classe C^∞ sur $\mathbb{R}$ et qu'elle vérifie (G.3), (G.4) et (G.5).

REMARQUE G.1. Il est immédiat que la fonction g est de classe C^∞ sur $\mathbb{R}$, mais la preuve par récurrence nous a permis de démontrer la propriété (G.4)-(G.5), qui nous permettra de calculer les dérivées successives de g.

(b) On en déduit successivement les quatre premières dérivées de f. Après calculs, d'après la relation de récurrence de la question 2a, il vient

$$p_4(x) = 16x^4 - 48x^2 + 12.$$

On en déduit donc

$$f^{(4)}(x) = 4\left(4x^4 - 12x^2 + 3\right)e^{-x^2}. \tag{G.11}$$

REMARQUE G.2. On pouvait procéder aussi comme cela était fait dans l'exercice de TD 3.1 (voir équation (3.10) du corrigé de TD) et calculer directement $f^{(4)}$ à la main, sans récurrence !

(c) Étudions la fonction $f^{(4)}$ sur l'intervalle $[0, A]$. D'après la relation de récurrence précédente, on a

$$f^{(5)}(x) = p_5(x)e^{-x^2} \tag{G.12}$$

où

$$p_5(x) = -32x^5 + 160x^4 - 120x = 8x\left(-4x^4 + 20x^2 - 15\right).$$

Ainsi, le signe de $f^{(5)}$ est égal au signe de l'expression $-4x^4 + 20x^2 - 15$, expression bicarrée qui se factorise par le biais de l'expression $g(X) = -4X^2 + 20X - 15$, où $X = x^2$. Les racines de g valent

$$X_1 = \frac{5 - \sqrt{10}}{2} \text{ et } X_2 = \frac{5 + \sqrt{10}}{2},$$

qui sont strictement positives. On en déduit donc le signe de la dérivée de $f^{(4)}$ ainsi que son tableau de variation sur $\mathbb{R}_+$ (cf. figure G.1).

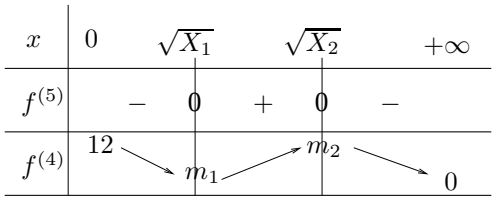


FIGURE G.1. Tableau de variation de la fonction $f^{(4)}$.

On peut vérifier que m_1 et m_2 sont majorés en valeur absolue par 12. Ainsi, on en déduit que $\left|f^{(4)}\right|$ est majorée par 12 sur $\mathbb{R}_+$ et en particulier sur $[0, A]$. On a donc, pour tout A positif,

$$M = 12. \tag{G.13}$$

REMARQUE G.3. On pouvait procéder aussi comme cela était fait dans l'exercice de TD 3.1 et obtenir plus rapidement la majoration plus brutale donnée par l'équation (3.18) du corrigé de TD, ou, ce qui revient au même qu'ici, utiliser les résultats de l'annexe C du corrigé de TD.

REMARQUE G.4. Il existe une majoration plus grossière de $f^{(4)}$ sur $[0, A]$: on écrit

$$\begin{aligned}\forall x \in \mathbb{R}_+^*, \quad \left|f^{(4)}(x)\right| &= \left|4\left(4x^3 - 12x^2 + 3\right)e^{-x^2}\right|, \\ &\leq 4e^{-x^2}\left|4x^3 - 12x^2 + 3\right|, \\ &\leq 4\left(4x^3 + 12x^2 + 3\right), \\ &\leq 4\left(4x^3 + 12x^3 + 3x^3\right), \\ &\leq 4\left(4A^3 + 12A^3 + 3A^3\right),\end{aligned}$$

puisque A est minoré par 1. On en déduit donc

$$M = 76A^3.$$

L'inconvénient de ce majorant est qu'il dépend de A.

(d) En utilisant (3.49) ou le tableau 3.3, on connaît un majorant de l'erreur d'intégration commise par la méthode de Simpson sur l'intervalle $[0, A]$

$$\left|E_N^S\right| \leq \frac{1}{2880} A\, h^4 M.$$

Pour obtenir une erreur inférieure à $\varepsilon/2$, il suffit donc que

$$h_{\max} = \left(\frac{1440\varepsilon}{AM}\right)^{1/4}. \tag{G.14}$$

(3) (a) La fonction $x \mapsto e^{-x^2}$ est continue sur $[A, X]$ donc intégrable et $R(X)$ est défini. En outre, puisque A est minoré par 1, sur $[A, X]$, on a $x \leq x^2$ et donc $e^{-x^2} \leq e^{-x}$, ce qui implique (G.6).

(b) Il est immédiat que

$$b(X) = e^{-A} - e^{-X}, \tag{G.15}$$

ce qui implique

$$l = e^{-A}. \tag{G.16}$$

(c) De (G.6) et (G.15), on déduit que

$$R(X) \leq e^{-A} - e^{-X},$$

ce qui implique (G.7).

(d) Puisque la fonction $x \mapsto e^{-x^2}$ est continue et positive, on sait que (G.7) implique que l'intégrale I est convergente et (G.8)

(e) On déduit de (G.8) que pour avoir (G.9), il suffit que

$$e^{-A} \leq \frac{\varepsilon}{2},$$

ce qui est équivalent à

$$A \geq -\ln\varepsilon + \ln 2. \tag{G.17}$$

REMARQUE G.5. On peut utiliser une méthode plus rapide, en intégration de Lebesgue : puisque la fonction $x \mapsto e^{-x^2}$ est mesurable (car continue), positive et majorée par la fonction $x \mapsto e^{-x}$ sur $[A, +\infty[$, on peut écrire dans $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$,

$$\int_A^{+\infty} e^{-x^2} dx \leq \int_A^{+\infty} e^{-x} dx.$$

Cette dernière intégrale vaut e^{-A} ; ainsi, dans $\overline{\mathbb{R}}_+$,

$$\int_A^{+\infty} e^{-x^2} dx \leq e^{-A} < +\infty,$$

ce qui implique que I est convergente et l'inégalité (G.8).

(4) (a) D'après (G.17), il suffit de prendre

$$A = 9,9035. \tag{G.18}$$

(b) D'après (G.13) et (G.14), il vient donc

$$h_{\max} \approx \left(\frac{1440 \times 10^{-4}}{9,9035 \times 12}\right)^{1/4};$$

soit, numériquement,

$$h_{\max} \approx 0,18657. \tag{G.19}$$

(c) De

$$N \geq N_{\min} \text{ où } N_{\min} = \mathrm{E}\left(\frac{A}{h_{\max}}\right) + 1, \tag{G.20}$$

on déduit donc que

$$N_{\min} = 54. \tag{G.21}$$

Si on appelle I_N^S la valeur approchée de l'intégrale, calculée par la méthode de Simpson sur l'intervalle $[0, A]$ avec N intervalles, on est sûr que $\left|I - I_N^S\right| \leq \varepsilon$. En effet, on a

$$\begin{aligned}
\left|I - I_N^S\right| &\leq \left|I - \int_0^A e^{-x^2} dx\right| + \left|\int_0^A e^{-x^2} dx - I_a\right|, \\
&= \left|\int_A^{+\infty} e^{-x^2} dx\right| + \left|\int_0^A e^{-x^2} dx - I_a\right|, \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.
\end{aligned}$$

En utilisant les sources matlab distribuée à l'url habituelle, (voir `int_fcn.m`), on déduit la valeur approchée de I avec $a = 0$, $b = A$ et $N = N_{\min}$, avec la méthode de Simpson fournit la valeur numérique suivante :

$$I_N^S = 8.862269254527579 \times 10^{-1} \tag{G.22}$$

Remarque G.6. Pour information, nous donnons aussi les valeurs fournies sur le même intervalle, avec le même nombre de points en utilisant respectivement la méthode des rectangles, du point milieux et du trapèze :

$$\begin{aligned}
I_N^R &= 9.779259995268316 \times 10^{-1} \\
I_N^M &= 8.862269254527581 \times 10^{-1} \\
I_N^T &= 8.862269254527575 \times 10^{-1}
\end{aligned}$$

Puisque la valeur exacte [1] de l'intégrale I est égale à

$$I = \frac{\sqrt{\pi}}{2} \approx 8.862269254527579 \times 10^{-1}$$

1. Formellement (mais ce calcul est totalement justifié *a posteriori*), on écrit successivement

$$\begin{aligned}
I^2 &= \int_0^{+\infty} e^{-x^2} dx \times \int_0^{+\infty} e^{-y^2} dy, \\
&= \iint_{\mathbb{R}_+ \times \mathbb{R}_+} e^{-(x^2+y^2)} dx dy,
\end{aligned}$$

nous pouvons donner les valeurs (calculées sous matlab) des erreurs commises en utilisant respectivement la méthode des rectangles, du milieu, des trapèzes et de Simpson :

$$E_N^R \approx 9.1699 \times 10^{-2} \tag{G.24}$$

$$E_N^M \approx 1.1102 \times 10^{-16} \tag{G.25}$$

$$E_N^T \approx 4.44089 \times 10^{-16} \tag{G.26}$$

$$E_N^S \approx 0. \tag{G.27}$$

Nous constatons que l'erreur obtenue est numériquement nulle (c'est à dire inférieure au zéro machine, de l'ordre de 10^{-16}) et donc que la majoration par $\varepsilon = 10^{-4}$ était correcte mais très pessimiste.

REMARQUE G.7. On peut utiliser la méthode de Gauss-Hermite, ce qui permet d'évaluer directement I sur l'intervalle $\mathbb{R}_+$, d'autant qu'elle est beaucoup plus efficace (voir [BM03, exercice 3.8 et section 3.3]).

REMARQUE G.8. *Remarque ultérieure à [BM03, exercice 3.3].*

La fonction `erf` de matlab [2] permet de calculer (pour l'utilisation de la loi normale centrée réduite) :

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{G.28}$$

(a) On peut en déduire

$$\int_0^x e^{-t^2} dt = \frac{\sqrt{\pi}}{2} \mathrm{erf}(x),$$

et en particulier

$$\int_0^{+\infty} e^{-t^2} dt = \frac{\sqrt{\pi}}{2} \mathrm{erf}(+\infty),$$

et, puisque, sous matlab, $\mathrm{erf}(+\infty) = 1$, on a

$$\int_0^{+\infty} e^- t^2 dt = \frac{\sqrt{\pi}}{2},$$

résultat que l'on connaissait déja (voir (G.23)).

(b) On peut aussi grâce à la fonction réciproque de `erf`, `erfinv` refaire les calculs de l'intégrale approchée de I au plus juste. On remplace (G.9) et le calcul de la question 3e (voir (G.17)) par la détermination de A, pour ε donné, défini par

$$\int_A^{+\infty} e^{-x^2} dx = \frac{\varepsilon}{2}. \tag{G.29}$$

Par ailleurs, la définition `erfinv` fournit, pour $x = \mathrm{erfinv}(y)$,

$$y = \mathrm{erf}(x), \tag{G.30}$$

---

ce que l'on écrit en polaire sous la forme (puisque $dxdy = dS = rdrd\theta$) :

$$\begin{aligned} I^2 &= \iint_{\theta \in [0,\pi/2], r \in \mathbb{R}_+} e^{-r^2} r dr d\theta, \\ &= \int_0^{\pi/2} d\theta \int_0^{+\infty} e^{-r^2} r dr, \\ &= \frac{\pi}{2} \left[-\frac{1}{2} e^{-r^2} \right]_{r=0}^{r=+\infty} \\ &= \frac{\pi}{4}, \end{aligned}$$

d'où l'on déduit

$$I = \frac{\sqrt{\pi}}{2}. \tag{G.23}$$

2. Cette remarque est pédagogique et un peu "hypocrite" car on utilise la valeur exacte de $\int_0^{+\infty} e^{-x^2} dx$ et des valeurs numériques de $\int_A^{+\infty} e^{-x^2} dx$ pour déterminer l'erreur permettant le calcul de I !

et donc, successivement

$$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt = y \Longleftrightarrow \int_0^x e^{-t^2}dt = \frac{\sqrt{\pi}}{2}y,$$
$$\Longleftrightarrow \int_0^{+\infty} e^{-t^2}dt - \int_x^{+\infty} e^{-t^2}dt = \frac{\sqrt{\pi}}{2}y,$$

et donc (G.30) est équivalent à

$$\int_x^{+\infty} e^{-t^2}dt = \frac{\sqrt{\pi}}{2}(1-y). \tag{G.31}$$

Ainsi, pour résoudre (G.29), par comparaison avec (G.31), il suffit de poser

$$A = \text{erfinv}(y),$$

avec

$$\frac{\varepsilon}{2} = \frac{\sqrt{\pi}}{2}(1-y)$$

soit encore

$$A = \text{erfinv}\left(1 - \frac{\varepsilon}{\sqrt{\pi}}\right). \tag{G.32}$$

Une fois A déterminé, on détermine $h_{\max}$ grâce à (G.14) où M est donné par (G.13), et enfin $N_{\min}$ est déterminé par (G.20), ce qui s'écrit encore

$$N_{\min} = \text{E}\left(A\left(\frac{AM}{1440\varepsilon}\right)^{1/4}\right) + 1,$$

et donc, compte tenu de (G.32), on obtient

$$N_{\min} = \text{E}\left(\left(\text{erfinv}\left(1 - \frac{\varepsilon}{\sqrt{\pi}}\right)\right)^{5/4}\left(\frac{M}{1440\varepsilon}\right)^{1/4}\right) + 1. \tag{G.33}$$

Numériquement, pour M donné par (G.13) et pour ε donné par

$$\varepsilon = 10^{-4}, \tag{G.34}$$

on obtient

$$A = 2.847737230440994, \tag{G.35}$$

et

$$N_{\min} = 12, \tag{G.36}$$

ce qui est plus faible que les valeurs données par (G.18) et (G.21), puisque l'on calcule au plus juste. De nouveau, pour $N = N_{\min}$, on obtient

$$I_N^S = 8.86176877051535\,10^{-1}, \tag{G.37}$$

du même ordre que le résultat fourni précédemment (voir (G.22)). Enfin, en utilisant la valeur exacte donnée par (G.23), on obtient une erreur égale à

$$E_N^S = 5.0048\,10^{-5}, \tag{G.38}$$

ce qui plus élevé que (G.27) mais plus faible que (G.34), ce qui est tout à fait normal de nouveau.

♢

Annexe H

Formules d'intégration élémentaires à 0, 1 et 2 points

Dans cette annexe, nous établissons simultanément les formules d'intégration élémentaires à 0, 1 et 2 points équirépartis, c'est-à-dire celle du rectangle à gauche (voir (3.26)), du trapèze (voir (3.33)) et de Simpson (voir (3.36)) sur l'intervalle $[a, b]$, grâce à la formule générale (3.21). Pour cela, on considère le support $\{x_0, x_1, x_2\}$ de $[a, b]$ défini par

$$x_0 = a, \tag{H.1a}$$

$$x_1 = b, \tag{H.1b}$$

$$x_2 = m = \frac{a+b}{2}. \tag{H.1c}$$

On détermine tout d'abord le polynôme Π_0 de f sur le support $\{x_0\}$, puis Π_1 de f sur le support $\{x_0, x_1\}$ et enfin Π_2 de f sur le support $\{x_0, x_1, x_2\}$.

$x_i \setminus k$	0	1	2
$x_0 = a$	$\boxed{f(a)}$		
		$\boxed{\frac{f(b)-f(a)}{b-a}}$	
$x_1 = b$	$f(b)$		$\boxed{2\,\frac{-2\,f(m)+f(b)+f(a)}{(-b+a)^2}}$
		$-2\,\frac{-f(m)+f(b)}{-b+a}$	
$x_2 = 1/2\,a + 1/2\,b$	$f(m)$		

TABLE H.1. Différences divisées de f.

Voir le tableau H.1, où figurent les différences divisées $f[a] = f[x_0]$, $f[a, b] = f[x_0, x_1]$ et $f[a, b, m] = f[x_0, x_1, x_2]$.

On obtient successivement de façon immédiate

$$\Pi_0(x) = f(a),$$

et donc

$$\int_a^b \Pi_0(x) = f(a)(b-a).$$

Ensuite, d'après (2.47) avec $n = 1$, on a

$$\int_a^b \Pi_1(x)dx = \int_a^b \Pi_0(x)dx + \int_a^b f[a, b](x-a)dx,$$

et donc, on obtient successivement

$$\begin{aligned}\int_a^b \Pi_1(x)dx &= f(a)(b-a) + \int_a^b f[a,b](x-a)dx,\\ &= f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\frac{(b-a)^2}{2},\\ &= \frac{b-a}{2}\left(2f(a)+f(b)-f(a)\right)\end{aligned}$$

et donc

$$\int_a^b \Pi_1(x)dx = \frac{b-a}{2}\left(f(a)+f(b)\right).$$

Enfin, d'après (2.47) avec $n=2$, on a

$$\int_a^b \Pi_2(x)dx = \int_a^b \Pi_1(x)dx + \int_a^b f[a,b,m](x-a)(x-b)dx,$$

et donc, on obtient successivement

$$\begin{aligned}\int_a^b \Pi_2(x)dx &= \frac{b-a}{2}\left(f(a)+f(b)\right) + f[a,b,m]\int_a^b (x-a)(x-b)dx,\\ &= \frac{b-a}{2}\left(f(a)+f(b)\right) - \frac{1}{6}f[a,b,m](b-a)^3,\\ &= \frac{b-a}{2}\left(f(a)+f(b)\right) - \frac{1}{6}2\,\frac{-2\,f\,(m)+f\,(b)+f\,(a)}{(-b+a)^2}(b-a)^3,\\ &= (b-a)\left(\frac{f(a)+f(b)}{2}\right) - \frac{(b-a)}{3}\left(f\,(a)-2f(m)+f(b)\right),\end{aligned}$$

et donc

$$\int_a^b \Pi_2(x)dx = \frac{b-a}{6}\left(f(a)+4f(m)+f(b)\right).$$

Bref, on a

$$\int_a^b \Pi_0(x)dx = f\,(a)\,(b-a)\,, \tag{H.2a}$$

$$\int_a^b \Pi_1(x)dx = \frac{b-a}{2}\left(f(a)+f(b)\right), \tag{H.2b}$$

$$\int_a^b \Pi_2(x)dx = \frac{b-a}{6}\left(f(a)+4f(m)+f(b)\right). \tag{H.2c}$$

Annexe I

Formule d'intégration élémentaires à 4 points

Comme dans l'annexe H, nous établissons simultanément la formule d'intégration élémentaire à 4 points équirépartis sur l'intervalle $[a, b]$, grâce à la formule générale (3.21). Pour cela, on considère le support $\{x_0, x_1, x_2\}$ de $[a, b]$ défini par

$$x_0 = a, \tag{I.1a}$$

$$x_1 = a + \frac{b-a}{3}, \tag{I.1b}$$

$$x_2 = a + 2\frac{b-a}{3}, \tag{I.1c}$$

$$x_3 = b. \tag{I.1d}$$

$x_i \setminus k$	0	1	2	3
$x_0 = a$	$\boxed{f0}$			
		$\boxed{3\,\frac{-f1+f0}{-b+a}}$		
$x_1 = 2/3\,a + 1/3\,b$	$f1$		$\boxed{9/2\,\frac{f2-2\,f1+f0}{(-b+a)^2}}$	
		$3\,\frac{-f2+f1}{-b+a}$		$\boxed{9/2\,\frac{-f3+3\,f2-3\,f1+f0}{(-b+a)^3}}$
$x_2 = 1/3\,a + 2/3\,b$	$f2$		$9/2\,\frac{f3-2\,f2+f1}{(-b+a)^2}$	
		$3\,\frac{-f3+f2}{-b+a}$		
$x_3 = b$	$f3$			

TABLE I.1. Différences divisées de f.

Voir le tableau I.1, où figurent les quatre différences divisées utiles, où, pour $0 \le i \le 3$, $fi = f(x_i)$.
On obtient successivement :

$$\int_a^b \Pi_0(x)dx = f(a)\,(b-a)\,, \tag{I.2a}$$

$$\int_a^b \Pi_1(x)dx = 1/2\,\left(f(a) - 3\,f\left(2/3\,a + 1/3\,b\right)\right)(-b+a)\,, \tag{I.2b}$$

$$\int_a^b \Pi_2(x)dx = -1/4\,\left(f(a) + 3\,f\left(1/3\,a + 2/3\,b\right)\right)(-b+a)\,, \tag{I.2c}$$

$$\int_a^b \Pi_3(x)dx = -1/8\,\left(f(b) + 3\,f\left(2/3\,a + 1/3\,b\right) + 3\,f\left(1/3\,a + 2/3\,b\right) + f(a)\right)(-b+a)\,. \tag{I.2d}$$

On a donc

$$\int_a^b \Pi_3(x)dx = (b-a)\sum_{i=0}^{3} W_i f(x_i), \tag{I.3}$$

où

$$W_0 = 1/8,$$
$$W_1 = 3/8,$$
$$W_2 = 3/8,$$
$$W_3 = 1/8.$$

Annexe J

Formules d'intégration élémentaires à 3 et 4 points et erreur associées (sous forme de problèmes corrigés)

Ces deux problèmes sont très proches d'un exercice donné à l'examen de MNB à l'automne 2019. Ils permettent d'établir à la fois les formules d'intégration élémentaires à 3 (méthode de Simpson) et 4 points et de montrer les formules d'erreurs associées.

Premier énoncé

Soient a et b deux réels tels que $a < b$. Considérons $n = 2$ et le support d'interpolation à 3 points, défini par

$$x_0 = a, \quad x_1 = \frac{a+b}{2}, \quad x_2 = b. \tag{J.1}$$

Soit f, une fonction continue sur l'intervalle $[a, b]$. On considère la formule de quadrature

$$Q(f) = W_0 f(x_0) + W_1 f(x_1) + W_2 f(x_2) \tag{J.2}$$

pour approcher numériquement l'intégrale

$$\mathcal{I}(f) = \int_a^b f(x)dx. \tag{J.3}$$

(1) (a) Montrer que le degré d'exactitude de la formule de quadrature Q est *supérieur ou égal* à l'entier n ssi

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_a^b l_i(x)dx, \tag{J.4}$$

où pour tout $i \in \{0, ..., n\}$, l_i désigne le polynôme d'interpolation de Lagrange sur le support $\{x_0, ..., x_n\}$.

(b) Déterminer les polynômes de Lagrange, en supposant, pour simplifier que

$$a = 0, \quad b = 1. \tag{J.5}$$

(c) On suppose que le degré d'exactitude de la formule de quadrature est au moins n. En déduire la valeur des poids W_i pour tout $i \in \{0, ..., n\}$, toujours sous l'hypothèse (J.5).

(d) *Pour toute la suite, on admettra*, que sans l'hypothèse (J.5), les poids W_i pour tout $i \in \{0, ..., n\}$ sont donnés par

$$W_i = (b-a)\widetilde{W}_i, \tag{J.6}$$

où les poids $\widetilde{W}_i$ pour tout $i \in \{0, ..., n\}$, sont ceux déterminés dans la question 1c, sous l'hypothèse (J.5).

Quelle formule du cours reconnaissez-vous ?

(2) (a) La formule de quadrature Q étudiée peut-elle être de degré d'exactitude strictement plus grand que n ? Déterminer dans ce cas le degré $r \in \mathbb{N}$ d'exactitude.

(b) (i) *On admet* que si le degré d'exactitude de la formule de quadrature est égal à r, alors, on a :

Il existe $\alpha \neq 0$ tel que, pour toute fonction $f \in C^{r+1}[a,b]$,

$$\mathcal{I}(f) - Q(f) = \alpha(b-a)^{r+2} f^{(r+1)}(\xi), \text{ où } \xi \in [a,b]. \quad \text{(J.7)}$$

Comment ce résultat et les calculs d'ordre de la question 2a permettent-ils de déterminer la constante α de ce résultat ? On utilisera la formule (J.7) pour une fonction f judicieusement choisie.

(ii) Grâce à cela retrouver le résultat du cours sur l'erreur commise dans la méthode élémentaire de Simpson.

Premier corrigé

On pourra consulter :
— l'exercice K.1 des TD qui propose une manière alternative pour étudier les formules de quadrature.
— Pour plus de détails sur les formules de Newton-Cotes, qui généralise les formule de quadrature étudiée ici on pourra consulter [CM84] et plus particulièrement le chapitre 2 et l'exercice 2.3 ou encore
`http://fr.wikipedia.org/wiki/Formule_de_Newton-Cotes`
`http://utbmjb.chez-alice.fr/UTBM/mt40/medianMT40_A03.zip`
`http://utbmjb.chez-alice.fr/UTBM/mt40/mediancorrigeMT40_A03.zip`
Dans tout ce corrigé, on note $n = 2$.

(1) (a)
- Si la formule de quadrature $Q(f)$ est de degré d'exactitude au moins $r \geq n$, alors elle est exacte pour l_j, pour j fixé, le polynôme de Lagrange de f sur le support $\{x_0, ..., x_n\}$ puisqu'il est de degré n. On a donc
$$\mathcal{I}(l_j) = Q(l_j),$$
et donc
$$\int_a^b l_j(x)dx = \sum_{i=0}^{n} W_i l_j(x_i).$$
On utilise le fait que
$$\forall j \in \{0, ..., n\}, \quad l_i(x_j) = \delta_{ij},$$
et on obtient donc
$$\int_a^b l_j(x)dx = \sum_{i=0}^{n} W_i \delta_{ij},$$
et donc
$$\int_a^b l_j(x)dx = W_j.$$
- Réciproquement, soit P un polynôme de degré au plus n. Montrons que $\mathcal{I}(P) = Q(P)$. On a successivement
$$\begin{aligned} Q(p) &= \sum_{i=0}^{n} W_i P(x_i), \\ &= \sum_{i=0}^{n} \left(\int_a^b l_i(x)dx \right) P(x_i), \\ &= \int_a^b \left(\sum_{i=0}^{n} P(x_i) l_i(x) \right) dx. \end{aligned}$$

D'après l'équation

$$\Pi_n(P)(x) = \sum_{i=0}^{n} P(x_i) l_i(x),$$

on a donc

$$= \int_a^b \Pi_n(P)(x) dx.$$

Puisque P est un polynôme de degré au plus n, on a donc $\Pi_n(P) = P$, ce qui achève la preuve.

(b) Sous l'hypothèse (J.5) de l'énoncé, on obtient après calculs :

$$l_0(x) = 2\,x^2 - 3\,x + 1\,; \tag{J.8a}$$

$$l_1(x) = -4\,x^2 + 4\,x\,; \tag{J.8b}$$

$$l_2(x) = 2\,x^2 - x. \tag{J.8c}$$

(c) Puisque que le degré d'exactitude de la formule de quadrature est au moins n, d'après le résultat de la question 1a, on a

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_a^b l_i(x) dx, \tag{J.9}$$

et donc, d'après l'expression obtenue dans la question 1b et sous l'hypothèse (J.5) de l'énoncé, on obtient après calculs :

$$W_0 = 1/6\,; \tag{J.10a}$$

$$W_1 = 2/3\,; \tag{J.10b}$$

$$W_2 = 1/6. \tag{J.10c}$$

(d) Si on admet l'équation (J.6) de l'énoncé (dont on pourra par exemple trouver une preuve dans `http://utbmjb.chez-alice.fr/UTBM/mt40/medianMT40_A03.zip` et `http://utbmjb.chez-alice.fr/UTBM/mt40/mediancorrigeMT40_A03.zip`), grâce à (J.10), on obtient donc l'expression suivante de $Q(f)$:

$$Q(f) = \frac{1}{6}(b-a)(f(a) + 4f((a+b)/2) + f(b)), \tag{J.11}$$

ce qui d'après le tableau 3.2 est exactement la méthode élémentaire de Simpson.

(2) (a) D'après la question 1a, les poids W_i ont été déterminés de telle sorte que la méthode soit de degré au moins n. Rien n'interdit qu'il ne soit pas plus grand. On essaye les différentes valeurs $p \geq n+1$ et pour ces valeurs, on détermine successivement pour $f = X^p$, $\mathcal{I}(f) - Q(f)$. Si $\mathcal{I}(f) - Q(f)$ est non nulle pour $f = X^{n+1}$, le degré est exactement n. Sinon, on détermine la plus grande valeur de p pour laquelle $\mathcal{I}(f) - Q(f)$ est nulle ce qui fournit le degré. Ici, on obtient, après calculs,

$$\text{pour } p = 3, \quad \mathcal{I}\left(X^p\right) - Q\left(X^p\right) = 0, \tag{J.12a}$$

$$\mathcal{I}\left(X^4\right) - Q\left(X^4\right) = \frac{1}{120}\,(a-b)^5 \neq 0. \tag{J.12b}$$

On obtient donc le degré r défini par

$$r = 3. \tag{J.13}$$

Remarque J.1. Ce degré est conforme au tableau de la remarque 3.15.

(b) (i) Si on applique la formule (J.7) de l'énoncé à la fonction $f = X^{r+1}$, on a $f^{(r+1)}(\xi) = (r+1)!$, indépendamment de ξ et donc

$$\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right) = \alpha(b-a)^{r+2}(r+1)!$$

Le terme $\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right)$ est exactement la première erreur non nulle, déterminé en question 2a. On a donc

$$\alpha = \frac{\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right)}{(b-a)^{r+2}(r+1)!}. \tag{J.14}$$

(ii) Ainsi, en utilisant (J.12b) et (J.13), on aboutit donc à

$$\alpha = \frac{\frac{1}{120}\,(a-b)^5}{\alpha(b-a)^{r+2}(r+1)} = \frac{\frac{1}{120}\,(a-b)^5}{(b-a)^{r+2}(r+1)!} = \frac{\frac{1}{120}\,(a-b)^5}{(b-a)^5 4!},$$

et donc, après simplification,

$$\alpha = -\frac{1}{2880}. \tag{J.15}$$

Grâce à la formule (J.7) de l'énoncé et (J.15), on vient donc de montrer l'expression de l'erreur commise dans la méthode élémentaire de Simpson, comme annoncé dans le tableau 3.3. Cette façon de procéder peut se programmer informatiquement et constitue une preuve rigoureuse (en admettant néanmoins l'équation (J.7) de l'énoncé) de l'expression de l'erreur commise.

Plus de détails pourront être trouvés dans l'annexe K et particulièrement dans la remarque K.5.

Second énoncé

Soient a et b deux réels tels que $a < b$. Considérons $n = 3$ et le support d'interpolation à 4 points, défini par

$$x_0 = a, \quad x_1 = a + \frac{1}{3}(b-a), \quad x_2 = a + \frac{2}{3}(b-a), \quad x_3 = b. \tag{J.16}$$

Soit f, une fonction continue sur l'intervalle $[a, b]$. On considère la formule de quadrature

$$Q(f) = W_0 f(x_0) + W_1 f\,(x_1) + W_2 f\,(x_2) + W_3 f\,(x_3) \tag{J.17}$$

pour approcher numériquement l'intégrale

$$\mathcal{I}(f) = \int_a^b f(x)dx. \tag{J.18}$$

(1) (a) Montrer que le degré d'exactitude de la formule de quadrature Q est *supérieur ou égal* à l'entier n ssi

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_a^b l_i(x)dx, \tag{J.19}$$

où pour tout $i \in \{0, ..., n\}$, l_i désigne le polynôme d'interpolation de Lagrange sur le support $\{x_0, ..., x_n\}$.

(b) Déterminer les polynômes de Lagrange, en supposant, pour simplifier que

$$a = 0, \quad b = 1. \tag{J.20}$$

(c) On suppose que le degré d'exactitude de la formule de quadrature est au moins n. En déduire la valeur des poids W_i pour tout $i \in \{0, ..., n\}$, toujours sous l'hypothèse (J.20).

Pour toute la suite, on admettra, que sans l'hypothèse (J.20), les poids W_i pour tout $i \in \{0, ..., n\}$ sont donnés par

$$W_i = (b-a)\widetilde{W}_i, \tag{J.21}$$

où les poids $\widetilde{W}_i$ pour tout $i \in \{0, ..., n\}$, sont ceux déterminés dans la question 1c, sous l'hypothèse (J.20).

(2) (a) La formule de quadrature Q étudiée peut-elle être de degré d'exactitude strictement plus grand que n ? Déterminer dans ce cas le degré $r \in \mathbb{N}$ d'exactitude.

(b) (i) *On admet* que si le degré d'exactitude de la formule de quadrature est égal à r, alors, on a :

Il existe $\alpha \neq 0$ tel que, pour toute fonction $f \in C^{r+1}[a,b]$,

$$\mathcal{I}(f) - Q(f) = \alpha (b-a)^{r+2} f^{(r+1)}(\xi), \text{ où } \xi \in [a,b]. \quad \text{(J.22)}$$

Comment ce résultat et les calculs d'ordre de la question 2a permettent-ils de déterminer la constante α de ce résultat ? On utilisera la formule (J.22) pour une fonction f judicieusement choisie.

(ii) Grâce à cela déterminer la constante α de la formule de quadrature étudiée.

Second corrigé

Dans tout ce corrigé, on note $n = 3$.

(1) (a) Voir preuve dans la démonstration du lemme 3.34.

(b) Sous l'hypothèse (J.20) de l'énoncé, on obtient après calculs :

$$l_0(x) = -9/2\, x^3 + 9\, x^2 - 11/2\, x + 1\,; \quad \text{(J.23a)}$$

$$l_1(x) = \frac{27}{2}\, x^3 - \frac{45}{2}\, x^2 + 9\, x\,; \quad \text{(J.23b)}$$

$$l_2(x) = -\frac{27}{2}\, x^3 + 18\, x^2 - 9/2\, x\,; \quad \text{(J.23c)}$$

$$l_3(x) = 9/2\, x^3 - 9/2\, x^2 + x. \quad \text{(J.23d)}$$

(c) Puisque que le degré d'exactitude de la formule de quadrature est au moins n, d'après le résultat de la question 1a, on a

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_a^b l_i(x) dx, \quad \text{(J.24)}$$

et donc, d'après l'expression obtenue dans la question 1b et sous l'hypothèse (J.20) de l'énoncé, on obtient après calculs :

$$W_0 = 1/8\,; \quad \text{(J.25a)}$$

$$W_1 = 3/8\,; \quad \text{(J.25b)}$$

$$W_2 = 3/8\,; \quad \text{(J.25c)}$$

$$W_3 = 1/8. \quad \text{(J.25d)}$$

(2) (a) D'après la question 1a, les poids W_i ont été déterminés de telle sorte que la méthode soit de degré au moins n. Rien n'interdit qu'il ne soit pas plus grand. On essaye les différentes valeurs $p \geq n+1$ et pour ces valeurs, on détermine successivement pour $f = X^p$, $\mathcal{I}(f) - Q(f)$. Si $\mathcal{I}(f) - Q(f)$ est non nulle pour $f = X^{n+1}$, le degré est exactement n. Sinon, on détermine la plus grande valeur de p pour laquelle $\mathcal{I}(f) - Q(f)$ est nulle ce qui fournit le degré. Ici, on obtient, après calculs,

$$\mathcal{I}\left(X^4\right) - Q\left(X^4\right) = \frac{1}{270}\, (a-b)^5 \neq 0. \quad \text{(J.26)}$$

On obtient donc le degré r défini par

$$r = 3. \quad \text{(J.27)}$$

(b) (i) Si on applique la formule (J.22) de l'énoncé à la fonction $f = X^{r+1}$, on a $f^{(r+1)}(\xi) = (r+1)!$, indépendamment de ξ et donc

$$\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right) = \alpha (b-a)^{r+2} (r+1)!$$

Le terme $\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right)$ est exactement la première erreur non nulle, déterminé en question 2a. On a donc

$$\alpha = \frac{\mathcal{I}\left(X^{r+1}\right) - Q\left(X^{r+1}\right)}{(b-a)^{r+2}(r+1)!}. \tag{J.28}$$

(ii) Ainsi, en utilisant (J.26) et (J.27), on aboutit donc à

$$\alpha = \frac{\frac{1}{270}\,(a-b)^5}{\alpha(b-a)^{r+2}(r+1)} = \frac{\frac{1}{270}\,(a-b)^5}{(b-a)^{r+2}(r+1)!} = \frac{\frac{1}{270}\,(a-b)^5}{(b-a)^5 4!},$$

et donc, après simplification,

$$\alpha = -\frac{1}{6480}. \tag{J.29}$$

Annexe K

Formules d'intégration élémentaires de Newton-Cotes (sous forme d'exercice corrigé)

Cet exercice a été donné à l'examen de MNB en Mécanique à l'automne 2017.

EXERCICE K.1.
Soit f, une fonction continue sur l'intervalle $[0,1]$. On considère la formule de quadrature

$$Q(f) = W_0 f(0) + W_1 f\left(1/2\right) + W_2 f\left(1\right), \tag{K.1}$$

pour approcher numériquement l'intégrale

$$\mathcal{I}(f) = \int_0^1 f(x)dx.$$

(1) (a) Trouver les poids W_0, W_1, W_2 tels que la formule intègre exactement les polynômes jusqu'au degré 2. On exprimera ces coefficients sous forme rationnelle.

(b) On propose dans cette question de procéder autrement pour éviter le calcul de l'inverse d'une matrice.

(i) Calculer $\Pi_2(f)$, le polynôme d'interpolation de f aux nœuds 0, 1/2, 1 en fonction de $f\left(0\right)$, $f(1/2)$, $f(1)$.

(ii) Calculer

$$\mathcal{I}\big(\Pi_2(f)\big) = \int_0^1 \Pi_2(f)(x)dx, \tag{K.2}$$

en fonction de $f\left(0\right)$, $f(1/2)$, $f(1)$.

(iii) En justifiant et en utilisant l'égalité de $Q\left(\Pi_2(f)\right)$ et de $\mathcal{I}\big(\Pi_2(f)\big)$, en déduire de nouveau l'expression des poids W_0, W_1, W_2 .

(2) Calculer le degré d'exactitude de cette formule.

(3) (a) Utiliser la formule de quadrature trouvée pour donner une approximation numérique de l'intégrale

$$\mathcal{I} = \int_0^1 e^{-x^2} dx.$$

(b) Quelle est l'erreur alors commise ?

Correction de l'exercice K.1.
On renverra aussi à l'exercice de TD 3.3 proche de cet exercice !
Pour toute la suite, les nœuds interpolant sont notés x_0, ..., x_n.

(1) (a) Par linéarité des fonctions $f \mapsto I(f)$ et $f \mapsto Q(f)$, l'exactitude de la formule de quadrature sur l'espace vectoriel des fonctions polynomiales de degré inférieur à 2, de dimension 3, est équivalente à l'exactitude de la formule de quadrature pour les vecteurs de la base canonique de cet espace vectoriel, c'est-à-dire pour les fonctions définies par

$$e_0(x) = 1, \quad e_1(x) = x, \quad e_2(x) = x^2.$$

Cela donne le système linéaire suivant

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 1/4 & 1 \end{pmatrix} C = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \end{pmatrix}, \tag{K.3}$$

où C est le vecteur des coefficients W_i recherchés. On peut le résoudre à la main ou matriciellement pour obtenir :

$$C = \begin{pmatrix} 1/6 \\ 2/3 \\ 1/6 \end{pmatrix}. \tag{K.4}$$

Remarque K.1.
La matrice intervenant dans le système linéaire (K.3) est en fait la matrice de Vandermonde suivante, correspondant aux points x_i donnés par

$$h = (b-a)/n, \quad \forall i \in \{0, ..., n\}, \quad x_i = a + ih, \tag{K.5}$$

avec, ici $n = 2$ et $a = 0$, $b = 1$:

$$D_n = \begin{pmatrix} 1 & 1 & 1 & ... & 1 & 1 \\ x_0 & x_1 & x_2 & ... & x_{n-1} & x_n \\ x_0^2 & x_1^2 & x_2^2 & ... & x_{n-1}^2 & x_n^2 \\ & \vdots & & ... & \vdots & \\ x_0^{n-1} & x_1^{n-1} & x_2^{n-1} & ... & x_{n-1}^{n-1} & x_n^{n-1} \\ x_0^n & x_1^n & x_2^n & ... & x_{n-1}^n & x_n^n \end{pmatrix}. \tag{K.6}$$

Notons que cette matrice peut s'inverser en théorie en utilisant les polynômes de Lagrange ! Voir [BM03, Exercice 2.5 p. 55]. Notons que dans le cas $n = 2$, l'inverse explicite a été calculé dans cet exercice :

$$D_2^{-1} = \begin{pmatrix} \dfrac{x_1 x_2}{(x_0 - x_1)(x_0 - x_2)} & -\dfrac{x_1 + x_2}{(x_0 - x_1)(x_0 - x_2)} & \dfrac{1}{(x_0 - x_1)(x_0 - x_2)} \\ \dfrac{x_0 x_2}{(x_1 - x_0)(x_1 - x_2)} & -\dfrac{x_0 + x_2}{(x_1 - x_0)(x_1 - x_2)} & \dfrac{1}{(x_1 - x_0)(x_1 - x_2)} \\ \dfrac{x_0 x_1}{(x_2 - x_0)(x_2 - x_1)} & -\dfrac{x_0 + x_1}{(x_2 - x_0)(x_2 - x_1)} & \dfrac{1}{(x_2 - x_0)(x_2 - x_1)} \end{pmatrix},$$

où les x_i sont donnés par (K.5). Ici, on a explicitement

$$D_2^{-1} = \begin{pmatrix} 1 & -3 & 2 \\ 0 & 4 & -4 \\ 0 & -1 & 2 \end{pmatrix}. \tag{K.7}$$

Si on veut l'inverser directement, on notera que cette matrice est sensible à l'inversion numérique (car elle a un conditionnement important quand n grandit). Il est préférable d'utiliser la méthode de la question 1b pour la calculer pour n grand.

(b) (i) Pour calculer $\Pi_2(f)$, le polynôme d'interpolation de f aux nœuds 0, 1/2, 1 en fonction de $f(0)$, $f(1/2)$, $f(1)$, on utilise la forme de Newton, en prenant bien soin de laisser les valeurs de $f(0)$, $f(1/2)$, $f(1)$ générique, comme le montre la suite.

$x_i \setminus k$	0	1	2
$x_0 = 0$	$\boxed{f(0)}$		
		$\boxed{2f(1/2) - 2f(0)}$	
$x_1 = 1/2$	$f(1/2)$		$\boxed{2f(1) - 4f(1/2) + 2f(0)}$
		$2f(1) - 2f(1/2)$	
$x_2 = 1$	$f(1)$		

TABLE K.1. Différences divisées de f.

Pour calculer le polynôme sous la forme de Newton, on détermine tout d'abord les différences divisées $f[x_i, ..., x_{i+k}]$ données dans le tableau K.1. Ensuite, on n'utilise plus que les différences divisées qui sont encadrées et le polynôme interpolateur de degré 2, $\Pi_2(f)$, est donné par la formule :

$$\Pi_2(f)(x) = \sum_{i=0}^{n} f[x_0, ..., x_i](x - x_0)...(x - x_{i-1}). \tag{K.8}$$

Ici, on a donc :

$$\Pi_2(f)(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

On a successivement

$$x - x_0 = x,$$
$$(x - x_0)(x - x_1) = x^2 - 1/2\, x.$$

Après calculs, il vient :

$$\Pi_2(f)(x) = 2\,x^2 f(1) - 4\,x^2 f(1/2) + 2\,x^2 f(0) + 4\,x f(1/2) - 3\,x f(0) - x f(1) + f(0). \tag{K.9}$$

(ii) L'intégrale $\mathcal{I}\big(\Pi_2(f)\big)$ définie par

$$\mathcal{I}\big(\Pi_2(f)\big) = \int_0^1 \Pi_2(f)(x)dx, \tag{K.10}$$

s'obtient en intégrant le polynôme $\Pi_2(f)$ qui vient d'être déterminé. En intégrant les fonctions 1, x et x^2, sur $[0, 1]$, on obtient donc finalement

$$\mathcal{I}\big(\Pi_2(f)\big) = 1/6\, f(1) + 2/3\, f(1/2) + 1/6\, f(0). \tag{K.11}$$

(iii) On cherche à trouver les coefficients W_i de telle sorte que la formule de quadrature intègre exactement les polynômes jusqu'au degré 2. Cela est donc équivalent à ce qu'elle soit exacte pour $\Pi_2(f)$, pour toute fonction f, puisque $\Pi_2(f)$ est un polynôme de degré au plus 2. C'est donc équivalent à l'égalité de $Q(\Pi_2(f))$ et de $\mathcal{I}\big(\Pi_2(f)\big)$. Dans cette dernière égalité, chacune des valeurs de $\Pi_2(f)(x_i)$ est remplacée par définition par $f(x_i)$. D'après (K.11), on voit donc apparaître les coefficients W_i qui correspondent bien à ceux donnés par (K.4).

(2) Pour déterminer le degré d'exactitude (que l'on appelle aussi l'ordre) de la méthode, nous avons deux méthodes.

(a) Soit, on remarque que, en posant $a = 0$ et $b = 1$, la formule de quadrature établie (K.11) est équivalente à

$$Q(f) = \frac{1}{6}(b-a)\big(f(a) + 4f((a+b)/2) + f(b)\big),$$

et on reconnaît la méthode de Simpson, dont on connaît l'erreur, d'après les formules rappelées dans l'énoncé. Voir tableau K.2. Ainsi, ici l'erreur est proportionnelle à $f^{(4)}(\eta)$, et donc nulle pour les

méthode	erreur
rectangle	$\frac{(b-a)^2}{2} f'(\eta)$
milieu	$\frac{(b-a)^3}{24} f''(\eta)$
trapèze	$-\frac{(b-a)^3}{12} f''(\eta)$
Simpson	$-\frac{(b-a)^5}{2880} f^{(4)}(\eta)$

TABLE K.2. Erreurs pour les méthodes élémentaires sur $[a, b]$; η appartient à $]a, b[$.

polynômes de degrés au plus trois, et non nulle, par exemple, pour le polynôme x^4. Ainsi, le degré d'exactitude (appelé aussi l'ordre) de la méthode, qui correspond au plus haut degré du polynôme exactement intégré par la formule de quadrature, vaut donc 3.

(b) Si on ne reconnaît pas la méthode vue en cours, il suffit de vérifier que le plus haut degré du polynôme exactement intégré par la formule de quadrature vaut 3. Ainsi, le degré d'exactitude (que l'on appelle aussi l'ordre) de la méthode est 3.

(3) (a) Pour donner une approximation numérique de l'intégrale, on utilise la formule de quadrature déterminée ; on a donc

$$\begin{aligned} \mathcal{I} &= \begin{pmatrix} 1/6 \\ 2/3 \\ 1/6 \end{pmatrix} . \begin{pmatrix} 1 \\ e^{-1/4} \\ e^{-1} \end{pmatrix}, \\ &= 1/6 + 2/3\, e^{-1/4} + 1/6\, e^{-1}, \\ &\approx 0.747180428909510. \end{aligned}$$

(b) On a ici $a = 0$ et $b = 1$. D'après le tableau donné dans l'énoncé (voir tableau K.2), l'erreur de la méthode de quadrature est donnée par

$$E(f) = \beta f^{(4)}(\xi), \text{ où } \xi \in [0, 1], \tag{K.12}$$

avec, ici,

$$\beta = -\frac{1}{2880}. \tag{K.13}$$

On détermine alors la dérivée à l'ordre exigé :

$$f^{(4)}(x) = 4\, e^{-x^2} \left(3 - 12\, x^2 + 4\, x^4\right).$$

On majore la valeur absolue de cette fonction sur $[0, 1]$ par M donné par

$$M = 12 \approx 12. \tag{K.14}$$

Ici, cette valeur a été obtenue grâce à la fonction fournie `maxabsfun`. Compte tenu de (K.12), (K.13) et (K.14), on a donc une erreur finalement majorée par

$$\varepsilon = \frac{1}{240} \approx 4.166667\,10^{-3}. \tag{K.15}$$

On peut aussi déterminer la valeur exacte (par une approximation très précise !) : on obtient

$$\mathcal{I} = 0.746824132812427,$$

et donc une erreur réellement commise de

$$E = 3.562961\,10^{-4},$$

qui est bien inférieure à celle donnée par (K.15).

Remarque K.2.

La façon de déterminer les coefficients de la forme de quadrature de la question 1b est la plus efficace. C'est cette méthode qui permet à la fois de déterminer les méthodes élémentaires usuelles (rectangles, trapèzes, ...) mais aussi l'erreur de convergence associée. Voir par exemple [BM03, Chapitre 3, p. à 81 à 89].

Remarque K.3.

En raisonnant comme dans la remarque K.1, une autre façon de déterminer les poids W_i est de remarquer que, pour un polynôme p de degré au plus $n = 2$,

$$\begin{aligned}\int_0^1 p(x)dx &= \int_0^1 \sum_{i=0}^{2} p(x_i) l_i(x) dx, \\ &= \sum_{i=0}^{2} p(x_i) \int_0^1 l_i(x) dx,\end{aligned}$$

où les l_i sont les polynômes de Lagrange, et donc

$$\forall i \in \{0, ..., n\}, \quad W_i = \int_0^1 l_i(x) dx.$$

Par exemple, on a

$$\begin{aligned}W_0 &= \int_0^1 2\,x^2 - 3\,x + 1 dx, \\ &= 1/6,\end{aligned}$$

ce qui est le premier coefficient déjà déterminé.

Remarque K.4.

De façon plus générale, la méthode de quadrature étudiée, pour n quelconque est appelée la méthode de Newton-Cotes fermée. On peut calculer les coefficients sur tout intervalle $[a, b]$. Les coefficients et les ordres pour les premières valeurs de n sont donnés dans le tableau K.3. Pour obtenir les coefficients sur un autre intervalle $[a, b]$ quelconque, il suffit de multiplier chacun de ces coefficients par $b - a$. Les points x_i correspondant sont naturellement donnés par $x_0 = 0$ si $n = 0$ et sinon

$$\forall i \in \{0, ..., n\}, \quad x_i = a + ih \text{ où } h = (b - a)/n.$$

En pratique, on choisit des valeurs de n pas trop élevées et souvent la méthode de Simpson suffira largement. De plus, au-delà, d'un certaine valeur de n_0 de n (pour $n_0 = 8$) les coefficients changent de signe ce qui favorise les propagations d'erreur d'arrondis.

valeurs de n	nom	coefficients	ordre
0	rectangle	1	0
1	trapèze	1/2, 1/2	1
2	Simpson	1/6, 2/3, 1/6	3
3	Simpson 3/8	1/8, 3/8, 3/8, 1/8	3
4	Boole-Villarceau	$\frac{7}{90}$, $\frac{16}{45}$, 2/15, $\frac{16}{45}$, $\frac{7}{90}$	5
5		$\frac{19}{288}$, $\frac{25}{96}$, $\frac{25}{144}$, $\frac{25}{144}$, $\frac{25}{96}$, $\frac{19}{288}$	5
6	Weddle-Hardy	$\frac{41}{840}$, $\frac{9}{35}$, $\frac{9}{280}$, $\frac{34}{105}$, $\frac{9}{280}$, $\frac{9}{35}$, $\frac{41}{840}$	7
7		$\frac{751}{17280}$, $\frac{3577}{17280}$, $\frac{49}{640}$, $\frac{2989}{17280}$, $\frac{2989}{17280}$, $\frac{49}{640}$, $\frac{3577}{17280}$, $\frac{751}{17280}$	7
8		$\frac{989}{28350}$, $\frac{2944}{14175}$, $-\frac{464}{14175}$, $\frac{5248}{14175}$, $-\frac{454}{2835}$, $\frac{5248}{14175}$, $-\frac{464}{14175}$, $\frac{2944}{14175}$, $\frac{989}{28350}$	9
9		$\frac{2857}{89600}$, $\frac{15741}{89600}$, $\frac{27}{2240}$, $\frac{1209}{5600}$, $\frac{2889}{44800}$, $\frac{2889}{44800}$, $\frac{1209}{5600}$, $\frac{27}{2240}$, $\frac{15741}{89600}$, $\frac{2857}{89600}$	9

TABLE K.3. Noms, coefficients et ordres des 10 premières méthodes (sur l'intervalle $[0,1]$).

REMARQUE K.5.

On peut montrer que toute méthode de Newton-Cotes fermée à $n+1$ (avec n non nul) points de support est d'ordre n si n est impair et d'ordre $n+1$ si n est pair. Cela généralise les propriétés vue en cours (voir par exemple [BM03, remarque 3.20]) : la méthode du trapèze ($n=1$) est d'ordre 1, celle de Simpson ($n=2$) est d'ordre 3. On peut aussi montrer que comme pour les méthode déjà vues en cours (comme dans le tableau K.2) une méthode de Newton-Cotes d'ordre m et à $n+1$ points possède une erreur $E_n(f)$ qui vérifie : il existe $\alpha_n > 0$ tel que, pour tout a, b, pour toute fonction $f \in C^{m+1}([a,b])$,

$$E_n(f) = \alpha_n (b-a)^{m+2} f^{(m+1)}(\xi), \text{ où } \xi \in [a,b]. \tag{K.16}$$

Plus précisément, $m = n$ si n est impair et $m = n+1$ si n est pair, c'est-à-dire : si n est pair (non nul) il existe $\alpha_n > 0$ tel que, pour tout a, b, pour toute fonction $f \in C^{n+2}([a,b])$,

$$E_n(f) = \alpha_n (b-a)^{n+3} f^{(n+2)}(\xi), \text{ où } \xi \in [a,b], \tag{K.17}$$

et si n est impair, il existe $\alpha_n > 0$ tel que, pour tout a, b, pour toute fonction $f \in C^{n+1}([a,b])$,

$$E_n(f) = \alpha_n (b-a)^{n+2} f^{(n+1)}(\xi), \text{ où } \xi \in [a,b]. \tag{K.18}$$

Par exemple, d'après le tableau K.2, pour $n=1$, on a $\alpha_1 = -1/12$ et pour $n=2$, on a $\alpha_2 = -1/2880$. De façon plus générale, on peut calculer les constantes α_n en remarquant que l'erreur commise pour le polynôme x^{m+1} (si m est l'ordre de la formule et n le degré non nul) vaut

$$\widetilde{E}_n = \int_a^b x^{m+1} dx - \sum_{i=0}^{n} W_i x_i^{m+1}, \tag{K.19}$$

cette erreur non nulle pouvant être calculée informatiquement. En utilisant (K.16), cette erreur vaut aussi

$$\widetilde{E}_n = \alpha_n (b-a)^{m+2} (m+1)!. \tag{K.20}$$

Bref, grâce à (K.20), on a l'expression explicite et algébrique de α_n, pour n non nul :

$$\alpha_n = \frac{\widetilde{E}_n}{(b-a)^{m+2}(m+1)!}, \tag{K.21}$$

où $\widetilde{E}_n$ est déterminée de façon informatique graçe à (K.19). Si $n=0$, on obtient α_0 grâce au tableau K.2.

Voir par exemple le tableau K.4.

valeurs de n	α_n
0	$1/2$
1	$-1/12$
2	$-\frac{1}{2880}$
3	$-\frac{1}{6480}$
4	$-\frac{1}{1935360}$
5	$-\frac{11}{37800000}$
6	$-\frac{1}{1567641600}$
7	$-\frac{167}{426924691200}$
8	$-\frac{37}{627836977715200}$
9	$-\frac{173}{458209960750080}$

TABLE K.4. Coefficients α_n des 10 premières méthodes.

REMARQUE K.6.

Pour plus de détails sur les formules de Newton-Cotes, on pourra consulter [CM84] et plus particulièrement le chapitre 2 et l'exercice 2.3 ou encore

`http://fr.wikipedia.org/wiki/Formule_de_Newton-Cotes`

`http://utbmjb.chez-alice.fr/UTBM/mt40/medianMT40_A03.zip`

`http://utbmjb.chez-alice.fr/UTBM/mt40/mediancorrigeMT40_A03.zip`

Annexe L

Méthode de dichotomie ou de bissection (sous la forme d'un exercice corrigé)

Donnons dans cette annexe, [BM03, Exercice 4.1].

Énoncé

Voir [BM03, TP 4.A]

Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ continue. On suppose $f(a)f(b) \leq 0$; la fonction f admet une racine dans $[a, b]$ dont on cherche une approximation.

(1) On construit les trois [1] suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ définies de la façon suivante par récurrence : on pose $a_0 = a$, $b_0 = b$ et pour tout $n \in \mathbb{N}$, $x_n = (a_n + b_n)/2$ avec

$$\begin{aligned} &\text{si } f(a_n)f(x_n) \leq 0, \quad a_{n+1} = a_n, \quad b_{n+1} = x_n, \\ &\text{si } f(a_n)f(x_n) > 0, \quad a_{n+1} = x_n, \quad b_{n+1} = b_n. \end{aligned}$$

Vérifier que les trois suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont définies et que, pour tout $n \in \mathbb{N}$,

$$f(a_n)f(b_n) \leq 0. \tag{L.1}$$

(2) Montrer que, pour tout $n \in \mathbb{N}$,

$$|a_{n+1} - b_{n+1}| \leq \frac{1}{2}|a_n - b_n|, \tag{L.2}$$

et

$$a \leq a_n \leq x_n \leq b_n \leq b, \qquad a_{n+1} \geq a_n, \quad b_{n+1} \leq b_n. \tag{L.3}$$

(3) En déduire que, pour tout $n \in \mathbb{N}$,

$$|a_n - b_n| \leq \frac{1}{2^n}(b - a). \tag{L.4}$$

(4) En déduire que, si f est continue sur $[a, b]$, alors les trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ convergent vers une racine de f, notée α. Montrer que, pour tout $n \in \mathbb{N}$,

$$|x_n - \alpha| \leq \frac{1}{2^n}(b - a). \tag{L.5}$$

(5) Si $\varepsilon > 0$ est donné, déterminer en fonction de ε, le plus petit entier n tel que

$$|x_n - \alpha| \leq \varepsilon. \tag{L.6}$$

(6) Quel est l'avantage de la méthode de la dichotomie ?

1. Souvent, on ne s'intéresse qu'aux deux suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ qui encadrent la racine recherchée. On rajoute ici la suite $(x_n)_{n\in\mathbb{N}}$, par analogie aux autres méthodes qui font intervenir une telle suite.

Corrigé

(1) On montre aisément par récurrence sur n, que, pour tout $n \in \mathbb{N}$, $f(a_n)f(b_n) \leq 0$, ce qui implique que les suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont correctement définies.

(2) L'inégalité (L.2) est immédiate par récurrence sur n, puisqu'à chaque itération, l'intervalle $[a_n, b_n]$ est découpé en deux (d'où le nom de la méthode [2]). Quant aux inégalités (L.3), elles résultent de la définition de x_n, a_n et b_n.

(3) L'égalité (L.4) se démontre à partir de (L.2) par récurrence sur n.

(4) La suite $(a_n)_{n\in\mathbb{N}}$ (resp. $(b_n)_{n\in\mathbb{N}}$) est donc croissante (resp. décroissante) et majorée par b (resp. minorée par a). Ainsi, elles admettent chacune une limite notée α et α'. D'après (L.4), à la limite $\alpha = \alpha'$. D'autre part, selon (L.1), puisque f est continue, à la limite $f(\alpha)^2 \leq 0$. Ainsi, $f(\alpha) = 0$ et les deux suite $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ convergent vers une racine de f. Ainsi, selon (L.3), la suite $(x_n)_{n\in\mathbb{N}}$ converge vers la même racine, d'après le théorème des gendarmes [3].

D'autre part, puisque $(a_n)_{n\in\mathbb{N}}$ est croissante, on a nécessairement, pour tout n, $a_n \leq \alpha$. De même, on a $b \geq \alpha$. Ainsi, pour tout n, $\alpha \in [a_n, b_n]$. Selon (L.3) et (L.4), on en déduit (L.5).

Remarque L.1. Attention, f peut avoir plusieurs racines et les suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ peuvent tendre vers l'une ou l'autre de ces racines.

Remarque L.2. Les deux suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont adjacentes : elles vérifient la définition L.3 et le lemme L.4 (pour plus de complément, consulter [RDO88, section 1.2.2. 2°)]) :

Définition L.3. Deux suites de réels $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont dites adjacentes si et seulement si la suite $(a_n)_{n\in\mathbb{N}}$ est croissante, la suite $(b_n)_{n\in\mathbb{N}}$ est décroissante et si $(a_n - b_n)_{n\in\mathbb{N}}$ tend vers zéro.

Lemme L.4. *Deux suites adjacentes convergent vers la même limite.*

On peut donc montrer grâce à (L.3) et (L.4) directement que les suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont adjacentes et en déduire grâce au lemme L.4 leur convergence ainsi que celle x_n, vers la même limite qui est une racine de f.

(5) Pour que (L.6) soit vrai, il suffit que $1/2^n(b-a) \leq \varepsilon$ ce qui donne (4.21) et (4.22).

(6) La méthode de dichotomie assure la convergence des suites (a_n) et (b_n) et on est sûr de ne pas «sortir» de l'intervalle initial $[a, b]$ grâce à (L.3).

Remarque L.5. On verra plus tard que la convergence (L.5) est «lente». Néanmoins, cette méthode est globalement convergente sur $[a, b]$, puisque $x_n \in [a, b]$. Ainsi, la méthode de la dichotomie permet de s'approcher lentement mais sûrement d'une racine de f. «Une fois assez proche», on pourra alors utiliser d'autres méthodes plus rapides, d'ordre supérieur à un. Leur convergence nécessite en général cette approche préalable.

2. Dichotomie est emprunté en 1750 du grec *dikhotomia*, «division en deux parties égales», élément correspondant à l'adverbe *dikha*, «en deux», et de *-tomia*, «division», «section». cf [Rey98]

3. Les deux gendarmes a_n et b_n encadrent x_n, qui n'a pas d'autre choix que de converger, contraint et forcé, vers la limite commune.

Annexe M

Dichotomie discontinue

M.1. Introduction

Supposons que l'on observe un segment dont une partie (à partir d'une extrémité) est peinte en rouge et l'autre (donc à partir de l'autre extrémité) est peinte en vert. On cherche à déterminer la limite rouge/vert (endroit où la couleur peut ne pas être définie !). On sait qu'à chaque extrémité, les deux couleurs sont différentes et nous allons, par dichotomie (voir par exemple section 4.3 page 79) couper ce segment en deux, un grand nombre de fois, de façon que les deux extrémités des segment obtenus soient toutes les deux de couleurs différentes. Au début, on considère le segment initial. Si la couleur n'est pas définie au milieu, c'est que la limite rouge/vert est atteinte. Sinon, la couleur y est définie et on certain que l'un des demi-segment présente une couleur différente à ses extrémités. On considère ce demi-segment-là et on réitère. On soit obtient la limite rouge/vert en un nombre fini d'itérations soit on obtient une suite de segment de longueur qui tend vers zéro (puisque divisée à chaque étape en deux) donc les extrémités tendent vers la limite recherchée.

Voilà donc le principe de la dichotomie appliqué à une fonction typiquement discontinue ! Nous allons le formaliser et constater aussi qu'il contient la dichotomie habituelle !

M.2. Principe

On remplace la dichotomie usuelle (voir par exemple l'annexe L page 194 ou [BM03, Exercice 4.1]) par la définition suivante :

Définition M.1. Soient a, b réels tels que $a < b$ et f une fonction de $[a, b]$ dans $\mathbb{R}$, dont le domaine de définition D_f est inclus dans $[a, b]$. On suppose que

$$a, b \in D_f, \tag{M.1a}$$

$$f(a) = \alpha, \quad f(b) = \beta \text{ avec } \alpha \neq \beta. \tag{M.1b}$$

On construit les trois suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ définies de la façon suivante par récurrence : on pose $a_0 = a$, $b_0 = b$ et pour tout $n \in \mathbb{N}$, $x_n = (a_n + b_n)/2$ avec

$$\begin{cases} 1) \text{ si } x_n \notin D_f \text{ ou } f(x_n) \notin \{\alpha, \beta\}, & \text{alors, } a_{n+1} = x_n \text{ et } b_{n+1} = x_n, \\ 2) \text{ sinon } (x_n \in D_f \text{ et } f(x_n) \in \{\alpha, \beta\}), & \text{alors,} \begin{cases} \text{si } f(x_n) = \alpha, & \text{alors } a_{n+1} = x_n, \quad b_{n+1} = b_n, \\ \text{si } f(x_n) = \beta, & \text{alors } a_{n+1} = a_n, \quad b_{n+1} = x_n. \end{cases} \end{cases} \tag{M.2}$$

Voir l'algorithme M.1.

On pourra comparer avec l'algorithme usuel de dichotomie qui prend aussi en compte le test d'arrêt si $b - a < \varepsilon$ ou $f(a)f(b) = 0$ (voir l'algorithme 4.1 page 80).

On a alors le résultat très simple suivant

Algorithme M.1 Algorithme de dichotomie discontinue *dichotomie-discontinue*(a, b, α, β, f, n, $\longrightarrow x$)

Entrée :

a, b réels tels que $a < b$ et f une fonction, vérifiant (M.1).

n, entier strictement positif.

Sortie :

x tels que $x \notin D_f$ ou $f(x) \notin \{\alpha, \beta\}$ ou $f(x) \in \{\alpha, \beta\}$ ($x = x_n$, après n itérations de l'algorithme).

```
p ⟵ 0
test ⟵ "vrai"
tant que p < n et test="vrai" faire
  p ⟵ p + 1
  x ⟵ (a + b)/2
  si x ∉ D_f ou f(x) ∉ {α, β} alors
    b ⟵ x
    a ⟵ x
    test ⟵ "faux"
  sinon
    si f(x) = α alors
      a ⟵ x
    sinon
      b ⟵ x
    fin si
  fin si
fin tant que
```

LEMME M.2. *Les trois suites* $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ *et* $(x_n)_{n\in\mathbb{N}}$ *de la définition M.1 convergent vers un réel* $l \in [a, b]$ *qui vérifie*

$$\forall n \in \mathbb{N}, \quad |b_n - a_n| \leq \frac{b-a}{2^n}, \tag{M.3a}$$

$$|l - x_n| \leq \frac{b-a}{2^n}, \tag{M.3b}$$

De plus,

(1) si la limite l *est atteinte en nombre fini d'itérations, alors, on a*

$$l \notin D_f \text{ ou } f(l) \notin \{\alpha, \beta\}. \tag{M.4}$$

(2) sinon,

$$l \text{ appartient à l'adhérence de } D_f, \tag{M.5a}$$

et

$$(f \text{ est discontinue à droite}) \text{ ou } (f \text{ est discontinue à gauche}). \tag{M.5b}$$

Enfin,

$$\text{si } l \text{ appartient à } D_f, \ f \text{ est discontinue en } l. \tag{M.5c}$$

DÉMONSTRATION. Elle est imédiate et est proche du raisonnement de l'annexe L page 194.

(1) L'idée essentielle est tant que $(x_n)_{n\in\mathbb{N}}$ appartient à D_f et que $f(x_n)$ appartient à $\{\alpha,\beta\}$, l'algorithme est donné par le cas 2) de (M.2) et on montre dans ce cas que l'algorithme est bien posé, c'est-à-dire que

$$f(a_n)=\alpha \text{ et } f(b_n)=\beta. \tag{M.6}$$

C'est en effet le cas pour $n=0$ d'après (M.1b). De plus, si c'est vrai au rang n, alors c'est aussi vrai au rang $n+1$, d'après le cas 2) de (M.2), si $f(x_n)=\alpha$, alors $f(a_{n+1})=f(x_n)=\alpha$ et $f(b_{n+1})=f(b_n)=\beta$. Sinon, $f(x_n)=\beta$, alors $f(a_{n+1})=f(a_n)=\alpha$ et $f(b_{n+1})=f(x_n)=\beta$. Dans cas, on a alors de façon immédiate

$$b_{n+1}-a_{n+1}=\frac{b_n-a_n}{2}, \tag{M.7}$$

et

$$b_{n+1}-a_{n+1}\leq\frac{b_n-a_n}{2}, \tag{M.8a}$$

$$a_{n+1}\geq a_n, \tag{M.8b}$$

$$b_{n+1}\leq b_n, \tag{M.8c}$$

$$a_n\leq x_n\leq b_n. \tag{M.8d}$$

Si au contraire, il existe un entier n pour lequel $x_n\notin D_f$ ou $f(x_n)\notin\{\alpha,\beta\}$, alors, on considère le plus petit entier n_0 pour lequel

$$x_{n_0}\notin D_f \text{ ou } f(x_{n_0})\notin\{\alpha,\beta\}. \tag{M.9}$$

Ainsi, les suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ deviennent stationnaires à partir de cet indice (toutes égales à x_{n_0}). Dans ce cas, (M.8) est toujours vrai. Ainsi, dans tous les cas, les suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont adjacentes (voir par exemple la remarque L.2 page 195 de l'annexe L ou [RDO88, section 1.2.2. 2°)]) et convergent donc vers une limite commune l, qui d'après (M.8d) est aussi la limite de $(x_n)_{n\in\mathbb{N}}$. Cette limite vérifie aussi

$$\forall n,\quad a_n\leq l\leq b_n. \tag{M.10}$$

Ainsi, (M.3) provient de (M.8a) et de (M.8d).

(2) Dans le cas 1) de (M.2), pour l'indice n_0 par définition, on a (M.9), ce qui implique (M.4) puisque $l=x_{n_0}$. Sinon, on est dans le cas 2) pour tout n, on a donc, par définition de cette méthode, (M.6) et (M.7), pour tout n. Dans les deux cas, (M.5a) est immédiat. Enfin, les deux suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ ne peuvent être simultanément constantes à partir d'un certain rang, d'après (M.7).

(a) Supposons, par exemple, que la suite $(a_n)_{n\in\mathbb{N}}$ soit constante à partir d'un certain rang. Ainsi, $l=a_n$ pour $n\geq q$ est dans D_f et donc

$$f(l)=\alpha. \tag{M.11}$$

Ainsi, il est nécessaire que $(b_n)_{n\in\mathbb{N}}$ soit même strictement décroissante (à partir d'un certain rang). D'après (M.6), on a donc

$$\lim_{n\to\infty} b_n=l \text{ avec } \forall n,\quad f(b_n)=\beta. \tag{M.12}$$

Ainsi, f est discontinue à droite et on a donc montré (M.5b). Si l appartient D_f, alors, de plus f n'est pas continue en l (sinon la limite de l'image de toute suite doit être égale à $f(l)$) et (M.5c) est montrée.

(b) Supposons maintenant qu'aucune des suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ ne soit constante à partir d'un certain rang. Ainsi, les deux suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont respectivement croissantes et décroissantes et jamais constante à partir d'un certain rang. Ainsi, d'après (M.6), si f admet une limite à droite et à gauche, elles sont différentes (puisqu'égales nécessairement à α et β) et on a donc montré (M.5b). En particulier, si l est dans D_f, f est discontinue en l. En effet, si f était continue en l, les deux limites à droite et à gauche de f en l seraient égales et on a donc montré (M.5c).

□

REMARQUE M.3. L'idée fondamentale de cet algorithme est qu'il détecte une discontinuité éventuelle de f et la continuité de f n'est donc pas requise !

Traitons le cas particulier suivant : on supposera que f ne prend que deux valeurs possibles sur une partition finie de $[a,b]$ par des intervalles avec une valeur quelconque (voir non définie) à la frontière entre deux intervalles.

LEMME M.4. *Supposons qu'il existe α, $\beta \in \mathbb{R}$ distincts et $p \in \mathbb{N}^*$ et $(c_k)_{1\leq k\leq p} \subset [a,b]^p$ tels que*

$$f(a) = \alpha \text{ et } f(b) = \beta, \tag{M.13a}$$

$$c_1 < c_2 < .... < c_p, \tag{M.13b}$$

$$\forall k \in \{1,...,p-1\}, \quad (\forall x \in]c_k, c_{k+1}[, \quad f(x) = \alpha) \text{ ou } (\forall x \in]c_k, c_{k+1}[, \quad f(x) = \beta), \tag{M.13c}$$

$$\forall k \in \{1,...,p\}, \quad f(c_k - 0) \neq f(c_k + 0), \tag{M.13d}$$

en tout point c_k, f peut être définie (auquel cas, elle peut prendre toute valeur) ou non. (M.13e)

Alors, les trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1 convergent vers l'un des c_k.

DÉMONSTRATION. Il suffit d'utiliser le lemme M.2. Si on est dans le cas 1), les seuls réels de $[a,b]$ où f n'est soit pas définie, soit différentes de α et β sont les c_k. Si on est dans le cas 2), alors les seuls réels de $[a,b]$ vérifiant (M.5b) et (M.5c) sont encore les c_k. □

LEMME M.5. *Le lemme M.4 est valable quelle que soit l'éventuelle valeur de f aux points c_k (où f peut ne pas être définie).*

DÉMONSTRATION. En effet, si on change la valeur de f en l'un des c_k ou en excluant l'un des c_k de D_f, les valeurs des suites $(x_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ peuvent être modifiées mais leur limite sera encore l'un des c_k. □

Le cas particulier du lemme M.4 où $p = 1$ correspond en fait à l'exemple donné dans l'introduction. Il correspond au lemme suivant

LEMME M.6. *Supposons qu'il existe α, $\beta \in \mathbb{R}$ distincts et $c \in [a,b]$ tels que*

$$f(a) = \alpha \text{ et } f(b) = \beta, \tag{M.14a}$$

$$(\forall x \in]a,c[, \quad f(x) = \alpha \text{ et } \forall x \in]c,b[, \quad f(x) = \beta) \text{ ou } (\forall x \in]a,c[, \quad f(x) = \beta \text{ et } \forall x \in]c,b[, \quad f(x) = \alpha), \tag{M.14b}$$

en c, f peut être définie (auquel cas, elle peut prendre toute valeur) ou non. (M.14c)

Alors, les trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1 convergent vers c.

EXEMPLE M.7. Reprenons le lemme (M.6) en étudiant les conséquences vues au lemme M.5 quand on modifie la valeur de f en c.

Supposons d'abord qu'en c, f soit non définie ou ait une valeur différente de α et β. Dans, ce cas, si l'un des x_n est égal à c, les suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$, deviennent stationnaires et c est donc atteint en un nombre fini d'itérations. Sinon, ces suites ne sont jamais stationnaires et c est donc atteint en un nombre infini d'itérations. Si au contraire, on suppose, qu'en c, f soit égale à α ou β. Dans ce cas, l'algorithme ne pourra plus "détecter" automatiquement la valeur de c et on est de nouveau dans le cas où c est donc atteint en un nombre infini d'itérations. En effet, dans le cas où aucun des x_n est égal à c, c est donc atteint en un nombre infini d'itérations. Supposons néanmoins que $f(c) = \alpha$ et que l'un des x_n soit égal à c. Dans ce cas, à partir de cet indice, tous les a_n seront égaux à c, les b_n seront égaux à x_n et la suite $(b_n)_{n\in\mathbb{N}}$ tendra vers c en étant toujours strictement décroissante. Si $f(c) = \beta$ et l'un des x_n est égal à c, alors à partir d'un indice, tous les

b_n seront égaux à c, les a_n seront égaux à x_n et la suite $(a_n)_{n\in\mathbb{N}}$ tendra vers c en étant toujours strictement croissante.

Remarque M.8. Le comportement mis en évidence dans l'exemple M.7 dans le cas où l'une des suites $(a_n)_{n\in\mathbb{N}}$ ou $(b_n)_{n\in\mathbb{N}}$ est constante est spécifique à la dichotomie discontinue. Cela ne peut arriver pour la dichotomie usuelle quand on impose d'arrêter l'algorithme si le zéro de f est atteint. En effet, supposons que cela se produise avec $(a_n)_{n\in\mathbb{N}}$ constante. Cela signifie que $b_n \to r$ avec $f(r) = 0$. Or, $f(a_n) = f(r) = 0$ et cela n'est pas possible.

M.3. Applications

M.3.1. Dichotomie usuelle (continue)

Lemme M.9. *Soit g une fonction continue sur $[a, b]$ vérifiant $g(a)g(b) \leq 0$. On considère la fonction f définie de $[a, b]$ dans $\mathbb{R}$ par*

$$\forall x \in \mathbb{R}, \quad f(x) = \textit{signe}(g(x)). \tag{M.15}$$

On rappelle que

$$\forall x \in \mathbb{R}, \quad \textit{signe}(x) = \begin{cases} 1 & \textit{si } x > 0, \\ -1 & \textit{si } x < 0, \\ 0 & \textit{si } x = 0. \end{cases} \tag{M.16}$$

On considère les trois suites correspondantes $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1. Alors, ces trois suites ne sont autres que celles de la dichotomie usuelle (voir l'annexe L) (modifiée de telle sorte que les trois suites deviennent stationnaires et égales à x_n, si pour un indice n_0, on a $f(x_{n_0}) = 0$) et elle convergent toutes les trois vers un zéro de f.

Démonstration. Ce lemme constitue donc une preuve alternative du lemme 4.1 page 79 !

On peut supposer sans perte de généralité que

$$g(a)g(b) < 0. \tag{M.17}$$

Si cette quantité est nulle, g a un zéro en a ou en b et il n'est pas nécessaire de mettre en place la dichotomie.

(1) Appliquons le lemme M.2 à la fonction f définie par (M.15). Il est clair que (M.1) a lieu puisque $D_f = [a, b]$ et que, d'après (M.17), (M.1b) est vérifiée avec $\alpha = \text{signe}(g(a))$ et $\beta = \text{signe}(g(b))$.

Les trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1 convergent vers un réel l qui vérifie soit (M.4), soit (M.5b) et (M.5c).

(a) Si on suppose (M.4), puisque $D_f = [a, b]$, on a donc $f(l) \notin \{-1, 1\}$ et donc (car le signe ne peut prendre que trois valeurs, 0, 1 ou -1)

$$0 = f(l) = \text{signe}(g(l)), \tag{M.18}$$

et donc que

$$g(l) = 0. \tag{M.19}$$

(b) Sinon, (M.5b) et (M.5c) impliquent que

$$f \text{ est discontinue en } l. \tag{M.20}$$

Supposons que

$$g(l) = \eta \in \{-1, 1\}. \tag{M.21}$$

Alors par définition $\eta g(l) > 0$ et par continuité de g en l, il existe $\delta > 0$ tel que ηg est strictement positive sur $]l - \delta, l + \delta[$. Autrement dit, signe(g) est constant et vaut η sur $]l - \delta, l + \delta[$ et donc f est continue en l, ce qui contredit (M.20). Ainsi, (M.21) est faux et on a de nouveau (M.18) et donc (M.19).

(2) Au laisse au lecteur, le soin de vérifier qu'avec la définition de f donnée par (M.15), les trois suites correspondantes $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ ne sont autres que celles de la dichotomie usuelle.

□

REMARQUE M.10. Si on remplace le signe de la définition (M.16) par (avec $\eta > 0$)

$$\forall x \in \mathbb{R}, \quad \mathrm{s}_\eta(x) = \begin{cases} 1 & \text{si } x > \eta, \\ -1 & \text{si } x < -\eta, \\ 0 & \text{si } x \in [-\eta, \eta], \end{cases} \tag{M.22}$$

le lemme M.4 est toujours valablre mais dans ce cas, la dichotomie présentée s'arrête quand un zéro de g est trouvé avec une précision η, ce qui se fait souvent pour la dichotomie habituelle.

REMARQUE M.11. On peut aussi considérer un test d'arrêt $b - a < \varepsilon$ dans l'algorithme de dichotomie pour l'interrompre à la précision ε sur l, comme la dichotomie habituelle.

REMARQUE M.12. Si on suppose que f admet un nombre fini de zéro dans $[a, b]$, la dichotomie habituelle est couverte par les lemmes M.4 et M.5 et on peut remplacer la valeur du signe de zéro par toute valeur, voire décréter qu'il n'est pas défini, ce qui ne modifiera pas les limites des suites (qui seront l'un des zéros de f), en vertu du lemme M.4. De plus, on peut même, ce qui revient au même, supposer que g n'est pas continue en ses zéros (elle doit l'être sur les intervalles ouverts où elle est non nulle). Le lecteur vérifiera alors que les limites des suites sont encore les zéros, cette fois-ci du prolongement continue de f sur $[a, b]$ (c'est-à-dire, là où f change de signe, on prolongera f par $\widetilde{f}$, nulle en ce point-là.

M.3.2. Détection de changement de signe

Appliquons le lemme M.4 au cas où $[a, b]$ est décomposé en un nombre fini d'intervalles ouverts où f est soit positive ou nulle sur les uns et strictement négative sur les autres soit strictement positive sur les uns et négative ou nulle sur les autres.

LEMME M.13. *Supposons qu'il existe $p \in \mathbb{N}^*$ et $(c_k)_{1\le k\le p} \subset [a, b]^p$ tels que*

$$(g(a) \ge 0 \text{ et } g(b) < 0) \text{ ou } (g(a) > 0 \text{ et } g(b) \le 0) \text{ ou } (g(a) \le 0 \text{ et } g(b) > 0 \text{ ou } (g(a) < 0 \text{ et } g(b) \ge 0) \tag{M.23a}$$

$$c_1 < c_2 < .... < c_p, \tag{M.23b}$$

on a l'un des cas suivants

$$\begin{cases} \forall k \in \{1, ..., p-1\}, & (\forall x \in]c_k, c_{k+1}[, \quad g(x) \ge 0) \text{ ou } (\forall x \in]c_k, c_{k+1}[, \quad g(x) < 0), \\ \forall k \in \{1, ..., p-1\}, & (\forall x \in]c_k, c_{k+1}[, \quad g(x) > 0) \text{ ou } (\forall x \in]c_k, c_{k+1}[, \quad g(x) \le 0), \\ \forall k \in \{1, ..., p-1\}, & (\forall x \in]c_k, c_{k+1}[, \quad g(x) \le 0) \text{ ou } (\forall x \in]c_k, c_{k+1}[, \quad g(x) > 0), \\ \forall k \in \{1, ..., p-1\}, & (\forall x \in]c_k, c_{k+1}[, \quad g(x) < 0) \text{ ou } (\forall x \in]c_k, c_{k+1}[, \quad g(x) \ge 0), \end{cases} \tag{M.23c}$$

pour tout $k \in \{1, ..., p\}$, on a l'un des cas suivants :

$$\begin{cases} g \text{ est positive ou nulle strictement à gauche de } c_k \text{ et strictement négative strictement à droite de } c_k, \\ g \text{ est strictement positive strictement à gauche de } c_k \text{ et négative ou nulle strictement à droite de } c_k, \\ g \text{ est négative ou nulle strictement à gauche de } c_k \text{ et strictement positive strictement à droite de } c_k, \\ g \text{ est strictement négative strictement à gauche de } c_k \text{ et positive ou nulle strictement à droite de } c_k, \end{cases} \tag{M.23d}$$

en tout point c_k, g peut être définie (auquel cas, elle peut prendre toute valeur) ou non. (M.23e)

Alors, les trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1 associée à la fonction g définie par de différentes façons (en fonction des diverses hypothèses vu ci-dessus) :

$$\forall x \in [a,b], \quad f(x) = \begin{cases} 1 & si\ g(x) \geq 0, \\ 0 & si\ g(x) < 0, \end{cases} \tag{M.24a}$$

ou bien

$$\forall x \in [a,b], \quad f(x) = \begin{cases} 1 & si\ g(x) > 0, \\ 0 & si\ g(x) \leq 0. \end{cases} \tag{M.24b}$$

convergent vers l'un des c_k.

Démonstration. Il suffit de remarquer que l'on applique le lemme M.4 à la bonne fonction f. □

M.3.3. Détection de changement de valeur

C'est exactement le cas du lemme M.4 et le cas particulier du lemme M.6.

M.4. Calcul de l en base 2

Définition M.14. Aux définitions des suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ de la définition M.1, on adjoint la suite $(\eta_n)_{n\in\mathbb{N}^*}$ donnée par : pour tout $n \in \mathbb{N}$

$$\text{Si } a_n < b_n, \text{ alors} \begin{cases} \text{si } c_n \in D_f \text{ et } f(c_n) \in \{\alpha,\beta\}, & \text{alors } \begin{cases} \text{si } a_{n+1} = a_n, & \eta_{n+1} = 0, \\ \text{si } b_{n+1} = b_n, & \eta_{n+1} = 1, \end{cases} \\ \text{si } c_n \notin D_f \text{ ou } f(c_n) \notin \{\alpha,\beta\}, & \text{alors } \eta_{n+1} = 1, \end{cases} \tag{M.25a}$$

$$\text{Si } a_n = b_n, \text{ alors } \eta_{n+1} = 0. \tag{M.25b}$$

Lemme M.15. *La limite l introduite dans le lemme M.4 vérifie*[1]

$$l = a + (b-a)\sum_{k=1}^{\infty} \frac{\eta_k}{2^k}, \tag{M.27}$$

ce qui signifie encore

$$\frac{l-a}{b-a} = \sum_{k=1}^{\infty} \frac{\eta_k}{2^k}, \tag{M.28}$$

et donc que $(\eta_n)_{n\in\mathbb{N}^*}$ correspond aux chiffres de l'écriture en base 2 illimitée de $(l-a)/(b-a)$. De plus, on a

$$\forall n \in \mathbb{N}, \quad \sum_{k=1}^{n} \frac{\eta_k}{2^k} \leq \frac{l-a}{b-a} \leq \sum_{k=1}^{n-1} \frac{\eta_k}{2^k} + \frac{1+\eta_n}{2^n}, \tag{M.29}$$

ce qui signifie que les chiffres $(\eta_k)_{1\leq k\leq n}$ et $\Big((\eta_k)_{1\leq k\leq n-1}, 1+\eta_n\Big)$ correspondent respectivement à l'approximation de $(l-a)/(b-a)$ par défaut et par excès.

Démonstration. Remarquons, tout d'abord que

$$\forall n \in \mathbb{N}, \quad a_{n+1} = a_n + \frac{\eta_{n+1}(b-a)}{2^{n+1}}. \tag{M.30}$$

1. ce qui signifie :

$$l = a + (b-a)\lim_{n\to+\infty} \sum_{k=1}^{n} \frac{\eta_k}{2^k}. \tag{M.26}$$

(1) Supposons tout d'abord que l'on soit toujours dans le cas 2) de (M.2). On a donc, pour tout n, $a_n < b_n$. Si on a $f(x_n) = \alpha$, alors $b_{n+1} = b_n$ et selon la définition (M.25a), on a $\eta_{n+1} = 1$ et donc, grâce à (M.7) :

$$\begin{aligned} a_{n+1} - a_n - \frac{\eta_{n+1}(b-a)}{2^{n+1}} &= x_n - a_n - \frac{b-a}{2^{n+1}}, \\ &= \frac{b_n - a_n}{2} - \frac{b-a}{2^{n+1}}, \\ &= \frac{b-a}{2^{n+1}} - \frac{b-a}{2^{n+1}}, \\ &= 0. \end{aligned}$$

Si, au contraire, on a $f(x_n) = \beta$, alors $a_{n+1} = b_n$ et selon la définition (M.25a), on a $\eta_{n+1} = 0$ et donc, grâce à (M.7) :

$$a_{n+1} - a_n - \frac{\eta_{n+1}(b-a)}{2^{n+1}} = 0.$$

et donc, (M.30) est vrai dans tous les cas. Ainsi, (M.30) implique, pour tout $p \in \mathbb{N}$, par sommation pour n allant de 0 à $p-1$:

$$a_p = a_0 + \sum_{n=0}^{p-1} \frac{\eta_{n+1}(b-a)}{2^{n+1}} = a + (b-a) \sum_{n=1}^{p} \frac{\eta_n}{2^n},$$

soit

$$\forall n \in \mathbb{N}, \quad a_n = a + (b-a) \sum_{n=1}^{p} \frac{\eta_k}{2^k}. \tag{M.31}$$

On on déduit aussi

$$\begin{aligned} \forall n \in \mathbb{N}, \quad b_n &= a_n + (b_n - a_n), \\ &= a + (b-a) \sum_{n=1}^{p} \frac{\eta_k}{2^k} + \frac{b-a}{2^n}, \end{aligned}$$

et donc

$$\forall n \in \mathbb{N}, \quad b_n = a + (b-a) \sum_{n=1}^{p-1} \frac{\eta_k}{2^k} + \frac{1+\eta_n}{2^n}. \tag{M.32}$$

De (M.10), (M.31) et (M.32), on déduit donc

$$\forall n \in \mathbb{N}, \quad a + (b-a) \sum_{n=1}^{p} \frac{\eta_k}{2^k} \leq l \leq a + (b-a) \sum_{n=1}^{p-1} \frac{\eta_k}{2^k} + \frac{1+\eta_n}{2^n}, \tag{M.33}$$

et donc, on a (M.29). On constate que la série géométrique de terme général $1/2^n$ est convergente et, puisque $\eta_n/2^n \leq 1/2^n$, la série de terme général $\eta_n/2^n$ est, elle aussi, convergente Si on passe à la limite $n \to +\infty$ dans (M.31) et (M.32), on obtient (M.27).

(2) Si au contraire, il existe un indice n_0 pour lequel, on est dans le cas 1) de (M.2), et on a donc (en prenant comme d'habitude le plus petit n_0)

$$\forall n \geq n_0 + 1, \quad a_n = b_n. \tag{M.34}$$

Dans, ce cas, d'après la définition (M.25a), on a

$$\eta_{n_0+1} = 1, \tag{M.35}$$

puis, d'après (M.25b), on a

$$\forall n \geq n_0 + 2, \quad \eta_{n+1} = 0. \tag{M.36}$$

Tous les calculs précédents sont valables à condition de remplacer les sommes $\sum_{k=1}^{\infty}$ par des sommes finies $\sum_{k=1}^{n_0+1}$; il vient donc

$$l = a + (b-a)\sum_{k=1}^{n_0+1} \frac{\eta_k}{2^k}. \tag{M.37}$$

□

Remarque M.16. L'écriture (M.27) permet donc $(\eta_k)_{1\leq k\leq n}$ étant données, de prévoir *a priori*, le comportement de l'algorithme de dichotomie discontinue, ce qui nous sera utile dans la section suivante. On fixera donc n_0 tel que (M.35) et (M.36) aient lieu de sorte que l'on ait (M.37), ce qui nous sera utile dans la section suivante.

Remarque M.17. Pour l'exemple (M.7), si on est dans le cas où f vaut α ou β et que l'un des c est atteint par les x_n, les valeurs de η_k après cet indice sont constamment égales à 0 ou constamment égales à 1. Attention, de façon analogue à ce qu'il se passe pour l'écriture décimale illimitée, il est en fait interdit d'avoir une suite de 1 illimitée en base 2. Dans ce cas-là, on remplacera tous les 1 par des zéros et l'on incrémentera de 1 le premier chiffre égal à 0 qui précède, voire l'unité auquel cas, on aura

$$0, \overline{a_1 a_2 .....a_p 1111111111111....} = 0, \overline{a_1 a_2 .....(a_p+1)}. \tag{M.38}$$

Pour plus de détails, voir par exemple [RDO88, section 1.3.2. 2°)] ou [Bas14a, Transparents 13 et 14] ce qui permettra d'expliquer le paradoxe apparent de l'égalité (M.38) ou, en base 10 :

$$0, 999999999999.... = 1.$$

Nous rencontrerons cette situation dans l'exemple M.23.

Remarque M.18. Puisque la dichotomie usuelle est couverte par la dichotomie discontinue, les résultats de cette section justifient donc la remarque 4.7 page 80.

M.5. Simulations numériques

Voir la fonction `http://utbmjb.chez-alice.fr/Polytech/MNBif/fichiers_matlab/bisection.m`. Tous les exemples donnés dans cette section figurent dans l'aide de cette fonction.

M.5.1. Détection de changement de valeur

Commençons par la détection d'un changement de valeur (voir la section M.3.3)

Exemple M.19.

Commençons par l'exemple de la détection d'un changement de valeur (voir la section M.3.3) On considère la fonction signe définie par (M.15) dont l'unique changement de valeur a lieu en

$$l = 0. \tag{M.39}$$

On se donne a et b, définis par

$$a = -\frac{1}{2}, \quad b = \pi, \tag{M.40}$$

ce qui assurera, vu l'aspect irrationnel de π qu'en principe (aux arrondis de calcul près) on atteindra jamais la valeur rationnelle de l donné par (M.39) avec un nombre fini d'itérations. On se donne pour $\varepsilon > 0$, la valeur :

$$\varepsilon = \texttt{eps} = 2^{-52} = 2.220445\,10^{-16}. \tag{M.41}$$

On obtient pour

$$n = 53, \tag{M.42a}$$

$$x_n = 2.907638\,10^{-17}, \tag{M.42b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.42c}$$

Exemple M.20.

Si on définit

$$a = 0, \quad b = 1 \tag{M.43}$$

et l définit par (M.37) avec les valeurs de $(\eta_k)_{k\in\mathbb{N}^*}$ et

$$n_0 + 1 = 11, \tag{M.44a}$$

$$\forall k \geq 11, \quad \eta_k = 0, \tag{M.44b}$$

$$(\eta_k)_{1\leq k\leq 10} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}, \tag{M.44c}$$

c'est-à-dire, numériquement

$$l = 0.586914062500. \tag{M.44d}$$

ε est donné par

$$\varepsilon = 0. \tag{M.45}$$

On considère f donnée par

$$\forall x \in \mathbb{R}, \quad f(x) = \text{signe}(x - l), \tag{M.46}$$

(où la fonction signe est donnée par (M.16)) dont l'unique changement de valeur correspond à $x = l$. On obtient pour

$$n = 9, \tag{M.47a}$$

$$x_n = 0.586914062500, \tag{M.47b}$$

et on vérifie que

$$x_n = l. \tag{M.47c}$$

ce qui est bien conforme aux résultats de la section M.4. On est dans le cas où la limite est atteinte en un nombre fini d'itérations.

Exemple M.21.

On reprend les mêmes données de l'exemple M.20 avec f donnée par (M.46) mais on suppose que le signe de 0 n'est pas défini. On obtient pour

$$n = 9, \tag{M.48a}$$

$$x_n = 0.586914062500, \tag{M.48b}$$

et on vérifie que

$$x_n = l, \tag{M.48c}$$

ce qui est identique à (M.47). Cela est conforme à ce que l'on observé dans l'exemple M.7 : la valeur de f en c n'a pas d'influence sur la limite.

Exemple M.22.

On reprend les mêmes données de l'exemple M.20 avec f donnée par (M.46), mais la fonction signe est modifiée de la façon suivante :

$$\forall x \in \mathbb{R}, \quad \text{signe}(x) = \begin{cases} 1 & \text{si } x > 0, \\ -1 & \text{si } x < 0, \\ 3 & \text{si } x = 0. \end{cases} \tag{M.49}$$

On obtient pour

$$n = 9, \tag{M.50a}$$
$$x_n = 0.586914062500, \tag{M.50b}$$

et on vérifie que

$$x_n = l, \tag{M.50c}$$

ce qui est identique à (M.47). Cela est conforme à ce que l'on observé dans l'exemple M.7 : la valeur de f en c n'a pas d'influence sur la limite.

Exemple M.23.

On reprend les mêmes données de l'exemple M.20 avec ε donné par (M.41), avec f donnée par (M.46), et la fonction signe est modifiée de la façon suivante :

$$\forall x \in \mathbb{R}, \quad \text{signe}(x) = \begin{cases} 1 & \text{si } x > 0, \\ -1 & \text{si } x < 0, \\ 1 & \text{si } x = 0. \end{cases} \tag{M.51}$$

On obtient pour

$$n = 51, \tag{M.52a}$$
$$x_n = 0.586914062500, \tag{M.52b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.52c}$$

La limite est la même que celle obtenue dans les exemples (M.20), (M.21) et (M.22) mais avec un nombre d'itération plus élevé. En effet, on obtient numériquement la valeur de η définie par

$$n_0 + 1 = 53, \tag{M.53a}$$
$$\forall k \geq 53, \quad \eta_k = 0, \tag{M.53b}$$
$$(\eta_k)_{1 \leq k \leq 10} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}, \tag{M.53c}$$
$$\forall k \in \{11, ..., 52\}, \quad \eta_k = 1. \tag{M.53d}$$

ce qui est *a priori* différent de (M.37), (M.44). En fait, c'est la même valeur à ε près. En effet, d'après l'exemple M.7, (M.37), (M.44) et (M.53), on a

$$\begin{aligned}
l - x_n &= (b-a)\left(\sum_{k=1}^{\widetilde{n_0}+1}\frac{\widetilde{\eta}_k}{2^k} - \sum_{k=1}^{n_0+1}\frac{\eta_k}{2^k}\right),\\
&= (b-a)\left(\frac{1}{2^{10}} - \sum_{k=11}^{52}\frac{1}{2^k}\right),\\
&= (b-a)\left(\frac{1}{2^{10}} - \frac{1}{2^{11}}\sum_{k=11}^{52}\frac{1}{2^{k-11}}\right),\\
&= (b-a)\left(\frac{1}{2^{10}} - \frac{1}{2^{11}}\sum_{k=0}^{41}\frac{1}{2^k}\right),\\
&= (b-a)\left(\frac{1}{2^{10}} - \frac{1}{2^{11}}\frac{1-\frac{1}{2^{42}}}{1-\frac{1}{2}}\right),\\
&= (b-a)\left(\frac{1}{2^{10}} - \frac{1}{2^{10}}\left(1-\frac{1}{2^{42}}\right)\right),\\
&= (b-a)\left(\frac{1}{2^{52}}\right),
\end{aligned}$$

ce qui, d'après (M.43), est bien inférieur ou égal à ε défini par (M.41). En fait, on a des valeurs identiques à celles obtenues dans les exemples (M.20), (M.21) et (M.22) à ε près. Cela est conforme à ce que l'on observé dans l'exemple M.7 : la valeur de f en c n'a pas d'influence sur la limite. On est ici dans le cas où en c, f est égale à β et l'un des x_n est égal à c. Dans ce cas, à partir d'un indice, tous les b_n seront égaux à c, les a_n seront égaux à x_n et la suite $(a_n)_{n\in\mathbb{N}}$ tendra vers c en étant toujours strictement croissante. On peut vérifier numériquement que les x_n obtenus forment bien une suite strictement croissante à partir de $n = 11$. De plus, on a aussi un unique indice $n = 10$, pour lequel $x_n = l$. Cela confirme bien ce que l'on a observé dans l'exemple M.7 : pour cet indice, l'algorithme ne pourra plus "détecter" automatiquement la valeur de c (atteint pour $n = 10$). Ensuite, il y a convergence des x_n vers l par valeurs strictement croissantes. Voir la figure M.1 où on a tracé le logarithme décimal de $|x_n - l|$ en fonction de n (sauf en $n = 10$ où il n'est pas défini et il a été pris par défaut à la valeur $\log(\texttt{eps})$, repéré par une étoile), ce qui confirme tout cela.

Enfin, d'après la remarque M.17, on a en fait, en reprenant le calcul précédent

$$\begin{aligned}
l &= \lim_{n\to+\infty} x_n,\\
&= a + (b-a)\left(\lim_{n\to+\infty}\left(\sum_{k=1}^{n}\frac{\eta_k}{2^k}\right)\right),\\
&= a + (b-a)\left(\lim_{n\to+\infty}\left(\sum_{k=1}^{10}\frac{\eta_k}{2^k} + \sum_{k=11}^{n}\frac{1}{2^k}\right)\right),\\
&= a + (b-a)\left(\sum_{k=1}^{10}\frac{\eta_k}{2^k} + \frac{1}{2^{10}}\right),
\end{aligned}$$

Autrement dit, les chiffres de l'écriture décimale illimitée en base 2 de $(l-b)/(b-a)$ sont

$$\begin{pmatrix}1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1\end{pmatrix}$$

et non

$$\begin{pmatrix}1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0\end{pmatrix}$$

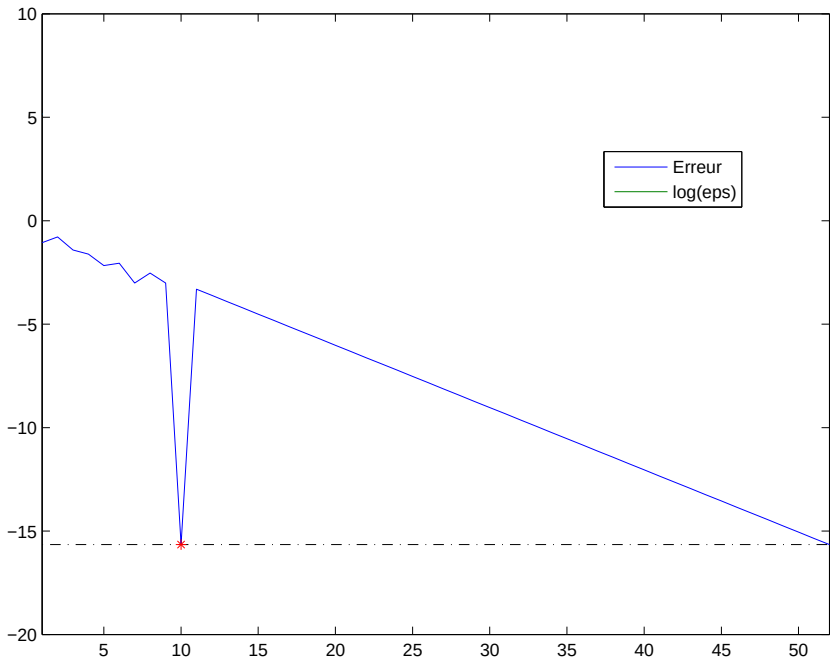


FIGURE M.1. Logarithme décimal de l'erreur en fonction de n.

suivi d'une infinité de 1.

M.5.2. Détection de changement de signe

Continuons par la détection d'un changement de signe (voir la section M.3.2)

EXEMPLE M.24.
On considère la fonction f définie par

$$\forall x \in \mathbb{R}, \quad f(x) = \begin{cases} x & \text{si } x > 0, \\ \min\left(0, x^2 \sin\left(\frac{1}{x}\right)\right) & \text{si } x < 0, \\ 0 & \text{si } x = 0. \end{cases} \tag{M.54}$$

Cette fonction est strictement positive sur $\mathbb{R}_+^*$ et négative ou nulle sur $\mathbb{R}_-$. Elle présente même un nombre infini d'intervalles où elle est strictement négative. Numériquement, ces intervalles sont en nombre fini, puisque leur taille tend vers zéro au voisinage de zéro. On est bien dans le cadre du lemme M.13. L'unique changement de signe a lieu en

$$l = 0. \tag{M.55}$$

(1) On considère a, b et $\varepsilon > 0$ donnés par (M.40) et (M.41). On obtient pour

$$n = 53, \tag{M.56a}$$

$$x_n = 2.907638\,10^{-17}, \tag{M.56b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.56c}$$

(2) On procède maintenant comme dans l'exemple M.23 en choisissant a, b et l définis par (M.37) et (M.44), ε défini par (M.41). On obtient pour

$$n = 51, \tag{M.57a}$$
$$x_n = 0.586914062500, \tag{M.57b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.57c}$$

ce qui est bien conforme aux résultats de la section M.4. On peut vérifier que l'on est dans un cas identique à celui de l'exemple M.23 avec cette fois-ci, les valeurs finales de η_k égales à 0. Voir la figure

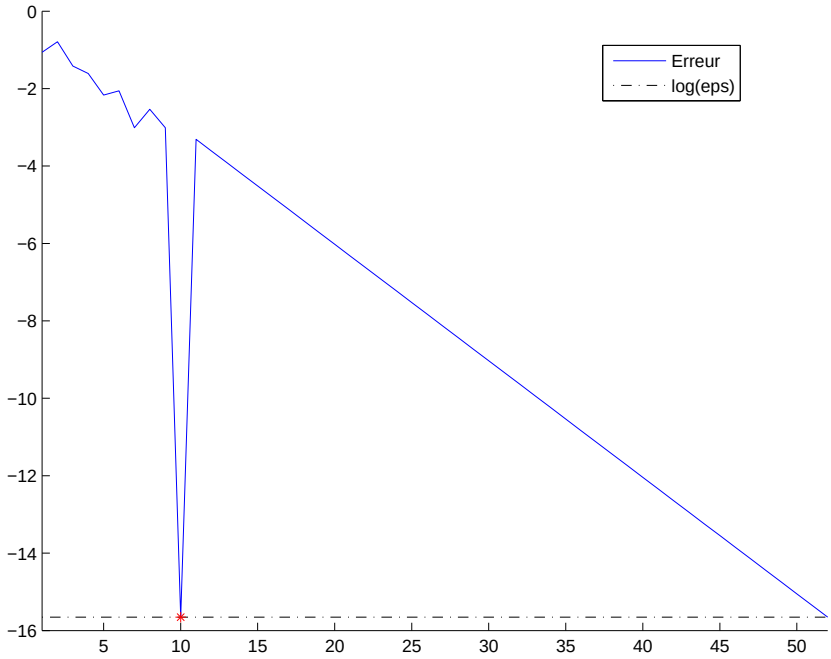


FIGURE M.2. Logarithme décimal de l'erreur en fonction de n.

M.2.

(3) Si on reprend l'exemple du cas (2) en changeant de f en la posant égale à 3 en 0, dans la définition (M.54), On obtient pour

$$n = 51, \tag{M.58a}$$
$$x_n = 0.586914062500, \tag{M.58b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.58c}$$

ce qui est bien conforme aux résultats de la section M.4. Cela est conforme à ce que l'on observé dans l'exemple M.7 : la valeur de f en c n'a pas d'influence sur la limite.

(4) Si on reprend l'exemple du cas (2) en changeant de f en la posant égale à 3 en 0, dans la définition (M.54), on obtient pour

$$n = 51, \tag{M.59a}$$
$$x_n = 0.586914062500, \tag{M.59b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon, \tag{M.59c}$$

et donc des résultats identiques à ceux du cas 3 de cet exemple.

(5) Si on reprend l'exemple du cas (2) en excluant la valeur 0 de D_f, par rapport à la définition (M.54), on obtient pour

$$n = 9, \tag{M.60a}$$
$$x_n = 0.586914062500, \tag{M.60b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon, \tag{M.60c}$$

et donc des résultats identiques à ceux du cas 3 de cet exemple, sauf que la limite est atteinte cette fois-ci en un nombre fini d'itérations.

M.5.3. Dichotomie usuelle (continue)

Finissons enfin par la détection d'un zéro d'une fonction (voir la section M.3.1)

Exemple M.25.

(1) On reprend a, b et ε définis par (M.40) et (M.41) et f donnée par

$$\forall x \in \mathbb{R}, \quad f(x) = \arctan(x), \tag{M.61}$$

dont la seule racine réellle est encore donnée par (M.39). On obtient pour

$$n = 53, \tag{M.62a}$$
$$x_n = -1.008866\,10^{-16}, \tag{M.62b}$$

et on vérifie que

$$|x_n - l| \leq \varepsilon. \tag{M.62c}$$

(2) Si on reprend, comme dans l'exemple M.20, a et b définis par (M.43), ε défini par (M.45), l défini par (M.37), (M.44) et (M.44d). On considère f donnée par

$$\forall x \in \mathbb{R}, \quad f(x) = \arctan(x - l), \tag{M.63}$$

dont l'unique zéro correspond à $x = l$. On obtient pour

$$n = 9, \tag{M.64a}$$
$$x_n = 5.869140\,10^{-1}, \tag{M.64b}$$

et on vérifie que

$$x_n = l. \tag{M.64c}$$

(3) Si on modifie la valeur de f en l ou si on exclut l de D_f, conformément à la remarque M.12, on obtient encore la même limite avec un nombre d'itérations fini ou non.

Annexe N

Convergence globale de la méthode du point fixe

En cours de rédaction.

Nous donnons dans cette annexe des résultats de convergence globale de la méthode du point fixe, qui peuvent compléter la proposition 4.19.

N.1. Cas particuliers

Proposition N.1. *Fonction g continue ayant un nombre fini de points fixes α_i avec chaque intervalle $[\alpha_i, \alpha_{i+1}]$ g-stables.*

Exemple : résultat relatif à g^2 du cas 2b de l'annexe P.

Dans ce cas, entre chaque point fixe, on a $g(x) > x$ ou $g(x) < x$. De plus, cela implique, avec la g stabilité de $[\alpha_i, \alpha_{i+1}]$ une majoration ou une minoration de la valeur absolue de la dérivée de g par rapport à 1 (et donc point attractif ou répulsif, dont l'aspect seul ne montre pas *a priori* la convergence ou la divergence de la méthode.

Proposition N.2. *Fonction g continue ayant un point fixe unique α appartenant à l'intérieur de l'intervalle I, g-stable, tel que*

$$\forall x \in I, x \leq \alpha \Longrightarrow g(x) \geq \alpha,$$
$$x \geq \alpha \Longrightarrow g(x) \leq \alpha,$$

telle que $g^2 = g \circ g$ vérifie les hypothèse de la proposition N.1.

Exemples : Annexes P ou V.

N.2. Cas général

Théorème N.3. *À rédiger*

Annexe O

Étude de la convergence globale de la suite définie par $u_0 \in \mathbb{R}_+^*$ et $u_{n+1} = 1/2(u_n + A/u_n)$, sous la forme d'un problème corrigé

Nous montrons dans cette annexe, la convergence globale sur $\mathbb{R}_+^*$ de la suite définie par $u_{n+1} = 1/2(u_n + A/u_n)$, c'est-à-dire sa convergence pour tout $u_0 \in \mathbb{R}_+^*$.

Énoncé

Pour $A \in \mathbb{R}_+^*$, on considère la suite u_n définie par $u_0 \in \mathbb{R}$ et, pour tout $n \in \mathbb{N}$, $u_{n+1} = 1/2\,(u_n + A/u_n)$.

(1) Étudier la convergence de la suite (u_n) pour $u_0 \geq \sqrt{A}$.

(2) Montrer qu'il y a encore convergence de la suite pour $u_0 \leq \sqrt{A}$.

(3) En introduisant la suite auxiliaire définie par $w_n = (u_n - \sqrt{A})/(u_n + \sqrt{A})$, montrer que $w_{n+1} = w_n^2$ et Conclure.

Pour $u_0 = 3$ et $A = 3$, comment choisir n pour que $\left|u_n - \sqrt{A}\right| \leq \varepsilon = 10^{-9}$? Calculer les valeurs de u_n correspondant et conclure.

Corrigé

On pourra consulter [BM03, Exercice 4.4] qui a inspiré cet exercice.

(1)

Voir, par exemple, la figure O.1.

Commençons par étudier la fonction $g(x) = 1/2(x + A/x)$ sur $\mathbb{R}_+^*$. On a

$$g'(x) = \frac{1}{2}\left(1 - \frac{A}{x^2}\right) = \frac{1}{2x^2}\left(x^2 - A\right),$$

strictement négative sur $]0, \sqrt{A}[$ et strictement positive pour $]\sqrt{A}, +\infty[$. Ainsi g est strictement décroissante sur $]0, \sqrt{A}[$ et strictement croissante pour $]\sqrt{A}, +\infty[$. De plus,

$$\begin{aligned} g(\sqrt{A}) &= \sqrt{A}, \\ \lim_{x \to 0^+} g(x) &= +\infty, \\ \lim_{x \to +\infty} g(x) &= +\infty. \end{aligned}$$

Voir le tableau de variation O.1 page suivante. L'asymptote de g en $+\infty$ est la droite d'équation $y = x/2$. On en déduit le tracé de g.

Voir par exemple la figure O.2. On déduit en particulier de l'étude de g que si $J = [\sqrt{A}, +\infty[$, alors J est g-stable (voir définition 4.18 page 86).

Remarquons aussi que si on pose $f(x) = g(x) - x$, on a

$$f(x) = \frac{1}{2x}\left(x^2 + A - 2x^2\right) = \frac{1}{2x}\left(A - x^2\right),$$

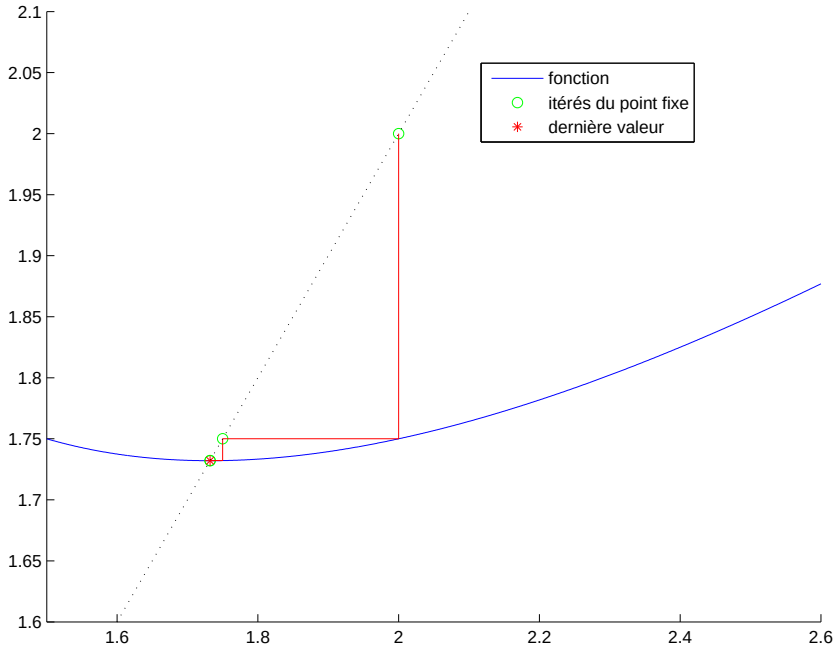


FIGURE O.1. Le tracé graphique des valeurs de u_n pour $u_0 = 2$ et $A = 3$.

x	0		$\sqrt{A}$		$+\infty$
signe de $g'(x)$		$-$	0	$+$	
variations de g	$+\infty$	↘	$\sqrt{A}$	↗	$+\infty$

TABLE O.1. Tableau de variation de g

nulle en $\sqrt{A}$, strictement négative sur $[\sqrt{A}, +\infty[$, strictement postive sur $[\sqrt{A}, +\infty[$. g n'admet donc qu'un seul point fixe, $\sqrt{A}$. Ainsi, pour tout $x \geq \sqrt{A}$, $g(x) \leq x$. Si $u_0 \in I$, alors d'après ce qui précède, pour tout n, u_n appartient à I et donc, $u_{n+1} \leq u_n$. La suite u_n est décroissante et minorée par $\sqrt{A}$; elle est donc convergente vers l'unique point fixe de g, qui est $\sqrt{A}$.

On peut aussi, de façon alternative, remarquer que pour tout n, $u_n \in I = [\sqrt{A}, u_1]$ pour tout n. On a aussi, pour tout $x \in I$, $0 < g'(x)$ et

$$g'(x) - 1/2 = -\frac{1}{2}\frac{A}{x^2},$$

et, pour tout $x \in I$, on a $1/x^2 \leq A$ et donc

$$g'(x) - 1/2 \geq -\frac{1}{2}.$$

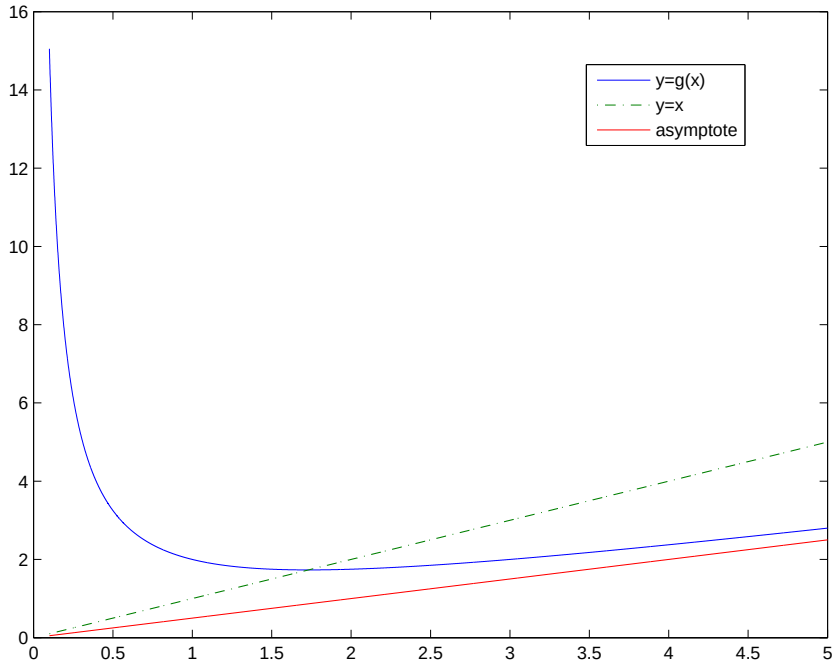


FIGURE O.2. La fonction g pour $A = 3$.

On peut donc appliquer la proposition 4.19 page 87, avec $k = 1/2$.

(2)

Voir, par exemple, la figure O.3.

Si $0 < u_0 \leq \sqrt{A}$, on montre par récurrence que pour tout n, $u_n > 0$ et que la suite est correctement définie. De plus, puisque $g(]0, \sqrt{A}]) = [\sqrt{A}, +\infty[$, alors u_1 appartient à $[\sqrt{A}, +\infty[$ et il suffit d'appliquer la question 1.

REMARQUE O.1. La convergence de la suite u_n pour toute valeur de u_0 pouvait aussi se déduire du théroème Y.1 page 287 de l'annexe Y.

(3) Par définition,

$$\begin{aligned} w_{n+1} &= \frac{\frac{1}{2}\left(u_n + \frac{A}{u_n}\right) - \sqrt{A}}{\frac{1}{2}\left(u_n + \frac{A}{u_n}\right) + \sqrt{A}}, \\ &= \frac{u_n^2 - 2\sqrt{A}u_n + A}{u_n^2 + 2\sqrt{A}u_n + A}, \\ &= \frac{\left(u_n - \sqrt{A}\right)^2}{\left(u_n + \sqrt{A}\right)^2}. \end{aligned}$$

et donc

$$\forall n \in \mathbb{N}, \quad w_{n+1} = w_n^2. \tag{O.1}$$

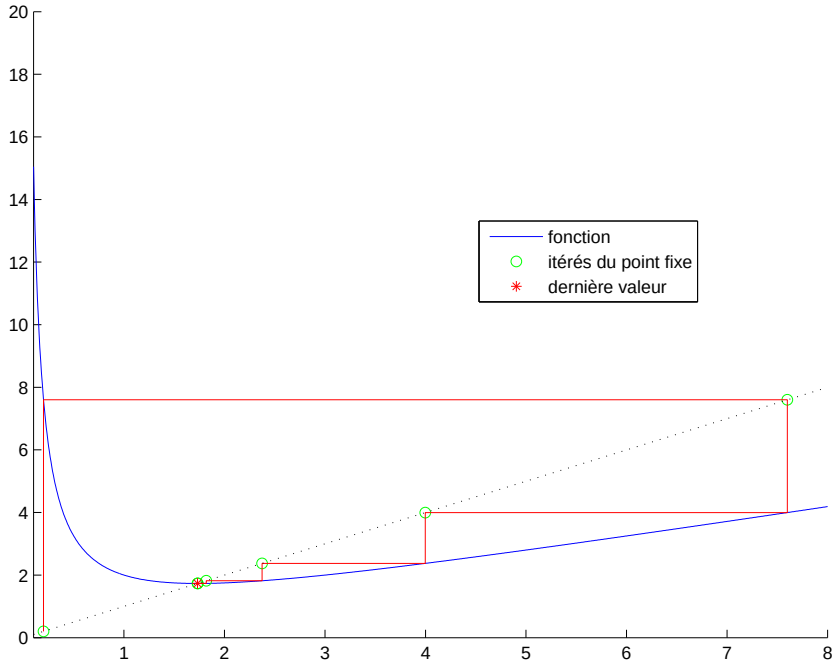


FIGURE O.3. Le tracé graphique des valeurs de u_n pour $u_0 = 0.200$ et $A = 3$.

On retrouve donc l'ordre deux de la convergence de la méthode de Newton. En effet, on peut réécrire (O.1) :

$$\frac{\left|u_n - \sqrt{A}\right|}{\left|u_{n+1} - \sqrt{A}\right|^2} = \frac{\left|u_n + \sqrt{A}\right|}{\left|u_{n+1} + \sqrt{A}\right|^2},$$

et donc

$$\lim_{n\to+\infty} \frac{\left|u_n - \sqrt{A}\right|}{\left|u_{n+1} - \sqrt{A}\right|^2} = \frac{1}{2\sqrt{A}}.$$

On déduit aussi de (O.1) par récurrence sur n que

$$\forall n \in \mathbb{N}, \quad w_n = w_0^{(2^n)} = \left(\frac{u_0 - \sqrt{A}}{u_0 + \sqrt{A}}\right)^{(2^n)}. \tag{O.2}$$

Puisque le nombre $\frac{u_0-\sqrt{A}}{u_0+\sqrt{A}}$ est dans $[0, 1[$, on en déduit que $w_n \to 0$ quand n tend vers l'infini. En revenant à la définition de w_n :

$$w_n = \frac{u_n - \sqrt{A}}{u_n + \sqrt{A}},$$

on a

$$u_n = \left(u_n + \sqrt{A}\right) w_n + \sqrt{A},$$

et puisque u_n est bornée, cela nous permet de retrouver que u_n tend vers $\sqrt{A}$ quand n tend vers l'infini.

(4) On peut déduire de (O.2) que, puisque $(u_n)_{n\in\mathbb{N}}$ est décroissante,

$$\left|u_n - \sqrt{A}\right| \leq \left(u_n + \sqrt{A}\right) w_n \leq \left(L_0 + \sqrt{A}\right) \left(\frac{L_0 - \sqrt{A}}{L_0 + \sqrt{A}}\right)^{(2^n)}.$$

Ainsi, pour que $\left|u_n - \sqrt{A}\right| \leq \varepsilon$, il suffit que

$$\left(L_0 + \sqrt{A}\right)\left(\frac{L_0 - \sqrt{A}}{L_0 + \sqrt{A}}\right)^{(2^n)} \leq \varepsilon, \tag{O.3}$$

donc que

$$(L_0 + 2)\left(\frac{L_0}{L_0 + 1}\right)^{(2^n)} \leq \varepsilon,$$

car $u_0 = 3$ et $A = 3$. Soit encore

$$n \geq \frac{\ln\left(\frac{\ln(\varepsilon/5)}{\ln(3/4)}\right)}{\ln 2}.$$

Numériquement, pour $\varepsilon = 10^{-9}$, on trouve $n \geq 7$. Indiquons les différentes valeurs de u_n et de $u_n - \sqrt{A}$, calculées sous matlab :

n	u_n	$\lvert u_n - \sqrt{A}\rvert$
0	3	1.26794
1	2	0.26794
2	1.75	0.01795
3	1.73214285714286	9.20496×10^{-5}
4	1.73205081001473	2.44585×10^{-9}
5	1.73205080756888	0
6	1.73205080756888	0
7	1.73205080756888	0

Les résultats sont bien conformes à ce que on avait prévu (mais, comme souvent, la majoration était pessimiste, en raison du caractère grossier de la majoration utilisée).

Annexe P

Étude du zéro de la fonction $e^x + x - K$, sous la forme d'un problème corrigé

Énoncé

Soit $K \in \mathbb{R}$. Dans cet exercice, on étudie l'équation

$$e^x = -x + K. \tag{P.1}$$

(1) Montrer que l'équation (P.1) admet une unique solution sur $\mathbb{R}$, notée r.

(2) On définit la fonction g par

$$\forall x \in \mathbb{R}, \quad g(x) = K - e^x, \tag{P.2}$$

On met désormais l'équation (P.1) sous la forme

$$g(x) = x, \tag{P.3}$$

et on considère la méthode du point fixe associée sur un intervalle $[a, b]$ (définie par $x_0 \in [a, b]$ et $x_{n+1} = g(x_n)$).

(a) On pose, dans cette question :

$$a = -2, \tag{P.4a}$$
$$b = -1/2, \tag{P.4b}$$
$$K = -1. \tag{P.4c}$$

(i) Montrer qu'avec ces valeurs la méthode du point fixe converge pour tout $x_0 \in [a, b]$.

(ii) Déterminer la valeur de n telle que $|x_n - r| \leq \varepsilon$ avec

$$\varepsilon = 10^{-2}. \tag{P.5}$$

(iii) Pour cette valeurs de n, calculer les valeurs de $(x_p)_{0 \leq p \leq n}$ avec $x_0 = -5/4$.

(b) On pose, dans cette question :

$$a = 1/4, \tag{P.6a}$$
$$b = 1, \tag{P.6b}$$
$$K = 5/2. \tag{P.6c}$$

(i) Est-ce que la méthode du point fixe converge ?

(ii)

On a affiché sur la figure P.1 page suivante, les 30 premières valeurs calculées avec les paramètres donnés par (P.6) et x_0 donné par

$$x_0 = 5/8. \tag{P.7}$$

On a indiqué dans le tableau P.1 page suivante, les valeurs correspondantes, en séparant les termes d'indices impairs et pairs.

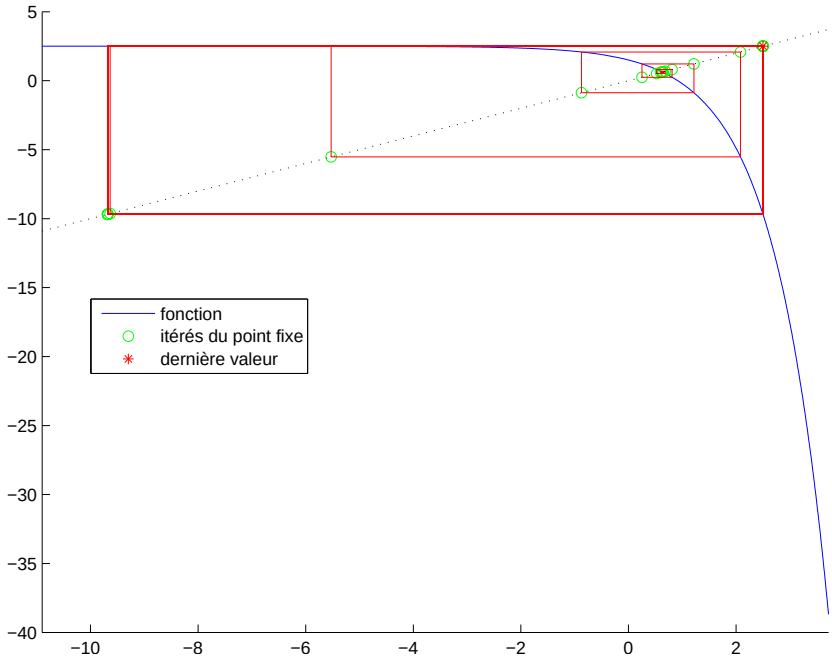


FIGURE P.1. Les 30 premières valeurs de x_n pour $x_0 = 5/8$ et $K = 5/2$.

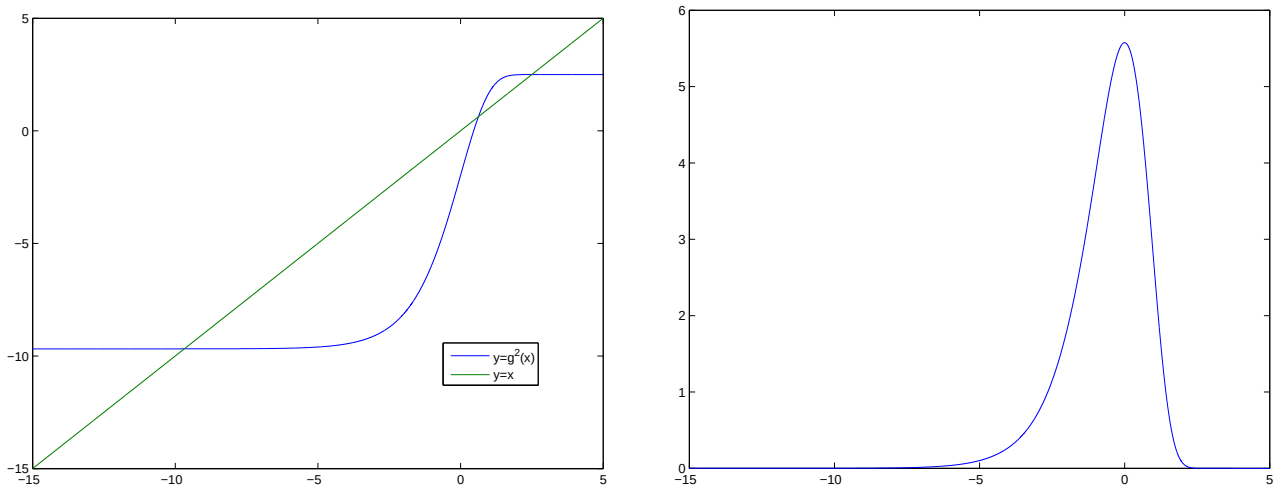
n	x_{2n}	x_{2n+1}
0	0.62500000000000	0.63175404256778
1	0.61909312177470	0.64275701533088
2	0.59828327950539	0.68100658482373
3	0.52413439208549	0.81100379305089
4	0.24983444603184	1.21618714121982
5	−0.87429745515325	2.08284501315852
6	−5.52727415891270	2.49602318550744
7	−9.63414264798875	2.49993454467100
8	−9.68169657765000	2.49993758447888
9	−9.68173360772356	2.49993758679009
10	−9.68173363587809	2.49993758679185
11	−9.68173363589949	2.49993758679185
12	−9.68173363589951	2.49993758679185
13	−9.68173363589951	2.49993758679185
14	−9.68173363589951	2.49993758679185

TABLE P.1. Les 30 premières valeurs de x_n pour $x_0 = 5/8$ et $K = 5/2$.

(A)

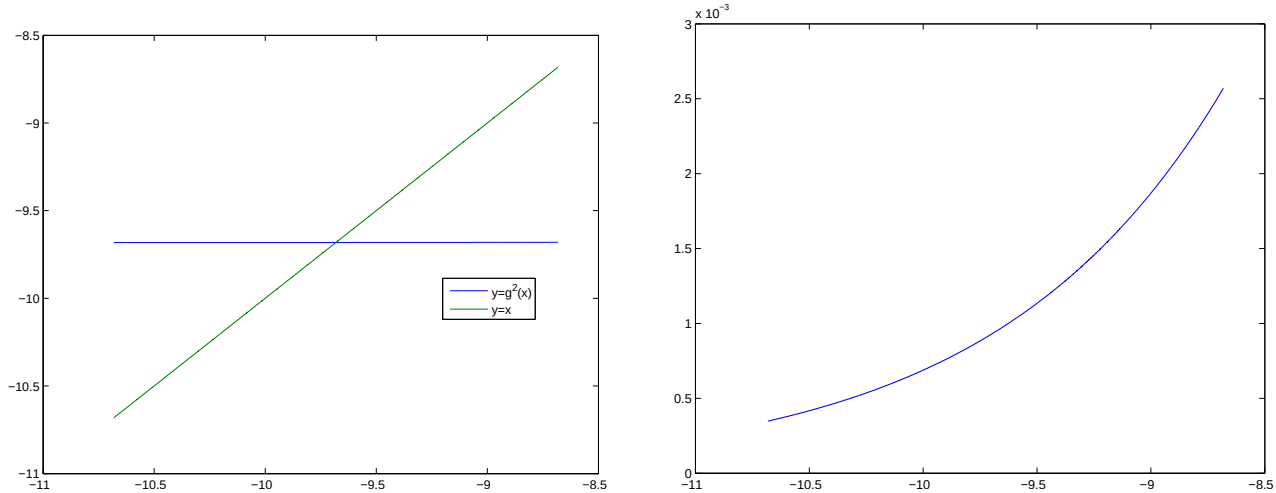
Que constatez-vous sur cette figure et sur ce tableau ?

(B)



(a) g^2 (b) $(g^2)'$

FIGURE P.2. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[-15, 5]$.



(a) g^2 (b) $(g^2)'$

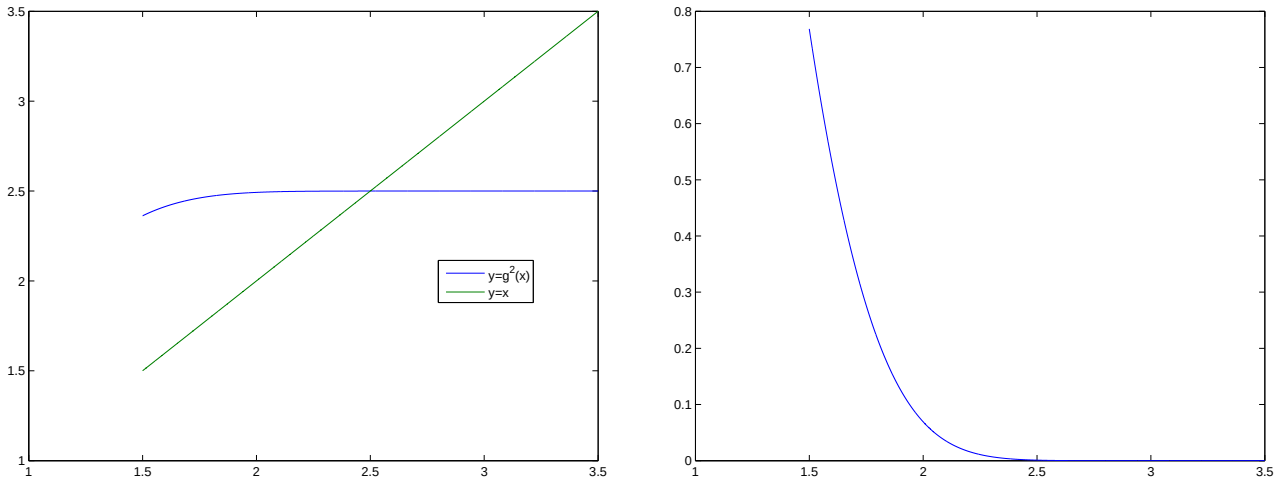
FIGURE P.3. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[-10.68173, -8.68173]$.

On définit la fonction g^2 par

$$\forall x \in \mathbb{R}, \quad g^2(x) = g(g(x)). \tag{P.8}$$

On a tracé sur la figure P.2, les graphes des fonctions g^2 et $(g^2)'$ sur l'intervalle $[-15, 5]$, sur la figure P.3, les graphes des fonctions g^2 et $(g^2)'$ sur l'intervalle $[-10.68173, -8.68173]$ et sur la figure P.4 page suivante les graphes des fonctions g^2 et $(g^2)'$ sur l'intervalle $[1.49994, 3.49994]$. Au vu de ces figures, essayez d'expliquer les observations faites dans la question 2(b)iiA.

(iii) Quelle méthode pourriez-vous utiliser pour résoudre (P.1) dans ce cas ?



(a) g^2 (b) $(g^2)'$

FIGURE P.4. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[1.49994, 3.49994]$.

Corrigé

(1) Soit $K \in \mathbb{R}$. Montrons que l'équation

$$e^x = -x + K. \tag{P.9}$$

admet une unique solution sur $\mathbb{R}$, notée r. Pour cela, on définit la fonction f sur $\mathbb{R}$ par

$$\forall x \in \mathbb{R}, \quad f(x) = e^x + x - K, \tag{P.10}$$

dont les zéros sont exactement les solutions de (P.9). La fonction f est dérivable sur $\mathbb{R}$ et on a

$$\forall x \in \mathbb{R}, \quad f'(x) = e^x + 1, \tag{P.11}$$

qui est strictement positive. La fonction f est donc strictement croissante sur $\mathbb{R}$. De plus, on a

$$\lim_{x \to -\infty} f(x) = -\infty, \tag{P.12a}$$

$$\lim_{x \to +\infty} f(x) = +\infty, \tag{P.12b}$$

dont on déduit, d'après la continuité de f et le théorème des valeurs intermédiaires que

$$f \text{ admet un unique zéro sur } \mathbb{R}, \text{ noté } r \text{ (ou } r(K) \text{ en cas d'ambiguïté).} \tag{P.13}$$

Par définition, on a donc

$$e^{r(K)} + r(K) - K = 0. \tag{P.14}$$

REMARQUE P.1. Remarquons que la stricte croissance de f sur $\mathbb{R}$ implique, puisque r est le zéro de f :

$$\begin{aligned} \forall x \in \mathbb{R}, \quad x > r &\implies f(x) > f(r) = 0, \\ x < r &\implies f(x) < 0, \\ x = r &\implies f(x) = 0, \end{aligned}$$

et donc, puisque

$$\forall x \in \mathbb{R}, \quad f(x) = x - g(x), \tag{P.15}$$

on obtient naturellement que

$$g \text{ n'a qu'un seul point fixe sur } \mathbb{R}, \text{ qui est } r \text{ (noté } r(K) \text{ en cas d'ambiguïté).} \tag{P.16}$$

et de plus

$$\forall x \in \mathbb{R}, \quad x > r \Longrightarrow g(x) < x, \tag{P.17a}$$
$$x < r \Longrightarrow g(x) > x, \tag{P.17b}$$
$$x = r \Longrightarrow g(x) = x. \tag{P.17c}$$

Remarque P.2. On a aussi

$$\forall K < 1, \quad |g'(r(K)| < 1, \tag{P.18a}$$
$$\forall K > 1, \quad |g'(r(K)| > 1, \tag{P.18b}$$
$$g'(r(1)) = -1. \tag{P.18c}$$

En effet, on a

$$\forall x \in \mathbb{R}, \quad g'(x) = -e^x, \tag{P.19}$$

et donc

$$\forall K \in \mathbb{R}, \quad g'(r(K)) = -e^{r(K)}.$$

Remarquons que d'après la stricte croissance de f, on a

$$r(K) < 0 \Longleftrightarrow f(r(K)) < f(0) \Longleftrightarrow 0 < 1 - K$$

et donc

$$r(K) < 0 \Longleftrightarrow K < 1, \tag{P.20}$$

ce qui montre (P.18a) et (P.18b). Enfin, il est clair que

$$r(1) = 0, \tag{P.21}$$

ce qui montre (P.18c).

(2) On définit la fonction g par

$$\forall x \in \mathbb{R}, \quad g(x) = K - e^x, \tag{P.22}$$

et on met désormais l'équation (P.9) sous la forme

$$g(x) = x, \tag{P.23}$$

et on considère la méthode du point fixe associée sur un intervalle $[a, b]$, définie par

$$x_0 \in [a, b] \text{ et } x_{n+1} = g(x_n). \tag{P.24}$$

Remarque P.3. On peut en fait, directement étudier la convergence ou la divergence de la méthode du point fixe (selon les valeurs de K) sans utiliser le théorème 4.12, comme le suggérait la suite de l'énoncé (et dont nous donnerons la solution plus bas, c'est-à-dire à partir du point 2a page 234) . Pour cela, on utilise des techniques proches de celles utilisées dans les annexes O et V.

La preuve se fait en plusieurs points.

(a) Commençons par définir la fonction g^2 par

$$g^2(x) = g(g(x) = K - e^{K-e^x}. \tag{P.25}$$

et considérons f_2 définie par (de façon analogue à (P.15))

$$f_2(x) = x - g^2(x) = x - K + e^{K-e^x}. \tag{P.26}$$

On a donc

$$\left(g^2\right)'(x) = e^{x+K-e^x}, \tag{P.27}$$

et donc

$$\forall x \in \mathbb{R}, \quad \left(g^2\right)'(x) > 0. \tag{P.28}$$

On déduit de (P.28) que

$$g^2 \text{ est strictement croissante sur } \mathbb{R}. \tag{P.29}$$

et que

$$f_2'(x) = 1 - \left(e^x e^{K-e^x}\right),$$

et donc

$$f_2'(x) = 1 - e^{x+K-e^x}, \tag{P.30}$$

et donc

$$f_2'(x) > 0 \text{ ssi } x + K - e^x < 0,$$

ce qui est équivalent à

$$f_2'(x) > 0 \text{ ssi } h(x) > 0, \text{ où } \forall x \in \mathbb{R}, \quad h(x) = e^x - x - K \tag{P.31}$$

Étudions la fonction h. On a

$$\forall x \in \mathbb{R}, \quad h'(x) = e^x - 1,$$

qui est strictement positive sur $\mathbb{R}_+^*$ et strictement négative sur $\mathbb{R}_-^*$.

x	$-\infty$		0		$+\infty$
signe de $h'(x)$		$-$	0	$+$	
variations de h	$+\infty$	↘	$1-K$	↗	$+\infty$

TABLE P.2. Tableau de variation de h.

Voir le tableau de variation de h dans le tableau P.2. La valeur minimale de h est donc donnée par

$$h_{\min} = h(0) = 1 - K.$$

Nous avons alors deux cas :

(i)

x	$-\infty$		0		$+\infty$
signe de $h'(x)$		$-$	0	$+$	
variations de h	$+\infty$	↘	$1-K \geq 0$	↗	$+\infty$

TABLE P.3. Tableau de variation de h dans le cas où $K \leq 1$.

Premier cas : on a

$$K \leq 1, \tag{P.32}$$

ce qui est équivalent à $h_{\min} \geq 0$ et donc, d'après le tableau P.3, h est strictement positive sur $\mathbb{R}^*$, et ainsi, d'après (P.31),

$$f_2 \text{ est strictement croissante sur } \mathbb{R}. \tag{P.33}$$

Enfin, puisque

$$\lim_{x\to-\infty} f_2(x) = -\infty, \tag{P.34a}$$
$$\lim_{x\to+\infty} f_2(x) = +\infty, \tag{P.34b}$$

on en déduit que f_2 admet un zéro unique sur $\mathbb{R}$. Or, d'après (P.16), on sait que

$$\text{le point fixe } r(K) \text{ de } g \text{ est aussi point fixe de } g^2, \tag{P.35}$$

(puisque $g^2(r) = g(r) = r$) et donc

$$\text{le point fixe } r(K) \text{ de } g \text{ est aussi zéro de } f_2, \tag{P.36}$$

et donc ici

$$\text{le point fixe } r(K) \text{ de } g \text{ est aussi l'unique zéro de } f_2. \tag{P.37}$$

On aboutit à des conclusions analogues à celles de la remarque P.1 : la stricte croissance de f_2 sur $\mathbb{R}$ implique, puisque $r(K)$ est le zéro de f_2 :

$$\forall x \in \mathbb{R}, \quad x > r(K) \Longrightarrow f_2(x) > f_2(r) = 0, \tag{P.38a}$$
$$x < r(K) \Longrightarrow f_2(x) < 0, \tag{P.38b}$$
$$x = r(K) \Longrightarrow f_2(x) = 0, \tag{P.38c}$$

et donc, d'après (P.26), on obtient naturellement que

$$g^2 \text{ n'a qu'un seul point fixe sur } \mathbb{R}, \text{ qui est } r \text{ (noté } r(K) \text{ en cas d'ambiguïté).} \tag{P.39}$$

et de plus

$$\forall x \in \mathbb{R}, \quad x > r(K) \Longrightarrow g^2(x) < x, \tag{P.40a}$$
$$x < r(K) \Longrightarrow g^2(x) > x, \tag{P.40b}$$
$$g^2(r(K)) = r(K). \tag{P.40c}$$

Remarque P.4. On a aussi

$$\forall K < 1, \quad \left|\left(g^2\right)'(r(K))\right| < 1, \tag{P.41}$$

et

$$\left(g^2\right)'(r(1)) = 1. \tag{P.42}$$

On a effet d'après (P.27),

$$\left(g^2\right)'(r(K)) = e^{r(K)+K-e^{r(K)}}$$

et la définition (P.14), de $r(K)$ implique

$$\forall K \in \mathbb{R}, \quad \left(g^2\right)'(r(K)) = e^{2r(K)}. \tag{P.43}$$

Enfin, (P.20) et (P.43) impliquent (P.41) et (P.42) vient de (P.21).

(ii)

Second cas : on a

$$K > 1. \tag{P.44}$$

Dans ce sous-cas, $h_{\min} = 1 - K > 0$ et la fonction h s'annule en deux réels δ et η, vérifiant $\delta < 0 < \eta$. Voir le tableau de variation P.4 dont on déduit que h est strictement positif sur $]-\infty, \delta[\cup]\eta, +\infty[$ et strictement négatif sur $]\delta, \eta[$, dont on déduit d'après (P.31) que f_2 est strictement croissante sur $]-\infty, \delta] \cup [\eta, +\infty[$ et strictement décroissante sur $[\delta, \eta]$. On rappelle aussi que l'on a (P.34).

On en déduit le tableau de variation P.5.

Par ailleurs, remarquons que

$$f_2(0) = e^{K-1} - K, \tag{P.45a}$$
$$f_2(\ln K) = 1 + \ln K - K. \tag{P.45b}$$
$$\tag{P.45c}$$

La fonction logarithme étant concave sur $\mathbb{R}_+^*$, elle est en dessous de sa tangente en tout point et en particulier au point 1, on a donc

$$\forall h > -1, \quad \ln(1+h) < h,$$

x	$-\infty$		δ		0		η		$+\infty$
signe de $h'(x)$			$-$		0		$+$		
variations de h	$+\infty$	$\searrow$	0	$\searrow$	$1-K<0$	$\nearrow$	0	$\nearrow$	$+\infty$

TABLE P.4. Tableau de variation de h dans le cas où $K > 1$.

x	$-\infty$		δ		0		$\ln K$		η		$+\infty$
signe de $f_2'(x)$		$+$		0		$-$		0		$+$	
variations de f_2	$-\infty$	$\nearrow$	$f_2(\delta)$	$\searrow$	$f_2(0)>0$	$\searrow$	$f_2(\ln(K))<0$	$\searrow$	$f_2(\eta)$	$\nearrow$	$+\infty$

TABLE P.5. Tableau de variation de f_2 dans le cas où $K > 1$.

résultat qui peut aussi être obtenu par une étude de fonction ou par une formule de Taylor-Lagrange à l'ordre 2. Si on pose $1 + h = u$ où $u > 0$, on en déduit

$$\forall u > 0, \quad \ln(u) < u - 1,$$

et en particulier, en $K > 0$, on a donc

$$\forall K > 0, \quad 1 + \ln(K) - K < 0, \tag{P.46}$$

ce qui nous prouve, grâce à (P.45b) :

$$\forall K > 0, \quad f_2(\ln K) < 0. \tag{P.47}$$

De la même façon, la fonction exponentielle étant convexe sur $\mathbb{R}$, elle est au-dessus de sa tangente en tout point et en particulier au point 0, on a donc

$$\forall h \neq 0, \quad e^h > h + 1,$$

résultat qui peut aussi être obtenu par une étude de fonction ou par une formule de Taylor-Lagrange à l'ordre 2. Si on pose $1 + h = u$ où $u \in \mathbb{R} \setminus \{1\}$, on en déduit

$$\forall u \neq 1, \quad e^{u-1} > h,$$

et en particulier, en $K \neq 0$, on a donc

$$\forall K \neq 0, \quad e^{K-1} > K \tag{P.48}$$

ce qui nous prouve, grâce à (P.45a) :

$$\forall K \neq 0, \quad f_2(0) > 0. \tag{P.49}$$

De cela et du tableau de variation P.5, on déduit que (car $\delta < 0 < \eta$),

$$f_2 \text{ ne s'anulle qu'une fois sur } \mathbb{R}_-^*. \tag{P.50}$$

Enfin, remarquons que $\eta > \ln K$, car c'est équivalent (car $\ln K > 0$) à $h(\eta) > h(\ln K)$ soit à $0 > -\ln K$, ce qui est vrai car $K > 1$. D'après (P.47), on peut donc compléter le tableau de variation P.5, dont on déduit que

$$f_2 \text{ ne s'anulle qu'une fois sur }]0, \ln K[\text{ et une fois sur }]\ln K, +\infty[. \tag{P.51}$$

Autrement dit, d'après (P.50) et (P.51), que f_2 possède trois zéros, deux à deux distincts, le premier dans l'intervalle $]-\infty, 0[$, le deuxième dans l'intervalle $]0, \ln K[$, et le troisième dans l'intervalle $]\ln(K), +\infty[$. Notons que l'unique zéro $r(k)$ de f est aussi dans l'intervalle $]0, \ln K[$. En effet, cela est équivalent à

$$f(0)f(\ln K) < 0,$$

ce qui est équivalent, compte tenu de la définition de f à

$$(1 - K)(K + \ln K - K) < 0,$$

soit encore à

$$(1 - K)\ln K < 0,$$

ce qui est vrai d'après (P.44). D'après (P.36), et par unicité du zéro de f dans $]0, \ln K[$, on en déduit que le deuxième zéro de f_2 (dans l'intervalle $]0, \ln K[$) ne peut être que $r(K)$ et d'après ce qui précède :

f_2 possède trois zéros, deux à deux distincts, le premier, noté $\alpha(K)$, dans l'intervalle $]-\infty, 0[$,
le second, égal à $r(K)$ dans l'intervalle $]0, \ln K[$, et le troisième, noté $\beta(K)$, dans l'intervalle $]\ln(K), +\infty[$. (P.52)

On en déduit le tableau P.6. En revenant à la la fonction g^2, on en déduit que

g^2 possède trois points fixes, deux à deux distincts, le premier, noté $\alpha(K)$, dans l'intervalle $]-\infty, 0[$,
le second, égal à $r(K)$ dans l'intervalle $]0, \ln K[$, et le troisième, noté $\beta(K)$, dans l'intervalle $]\ln(K), +\infty[$. (P.53)

x	$-\infty$	$\alpha(K)$	δ	$r(K)$	η	$\beta(K)$	$+\infty$
signe de $f_2'(x)$		$+$	0	$-$	0	$+$	
variations de f_2	$-\infty$	0	$f_2(\delta)$	0	$f_2(\eta)$	0	$+\infty$

TABLE P.6. Tableau de variation de f_2 dans le cas où $K > 1$.

On aboutit à des conclusions analogues à celles des équations (P.38) et (P.40) : le tableau de variation P.6 implique

$$\forall x \in \mathbb{R}, \quad x < \alpha(K) \Longrightarrow f_2(x) < 0, \tag{P.54a}$$
$$\alpha(K) < x < r(K) \Longrightarrow f_2(x) > 0, \tag{P.54b}$$
$$r(K) < x < \beta(K) \Longrightarrow f_2(x) < 0, \tag{P.54c}$$
$$x > r(K) \Longrightarrow f_2(x) > 0, \tag{P.54d}$$
$$x \in \{\alpha(K), r(K), \beta(K)\} \Longrightarrow f_2(x) = 0, \tag{P.54e}$$

et

$$\forall x \in \mathbb{R}, \quad x < \alpha(K) \Longrightarrow g^2(x) > x, \tag{P.55a}$$
$$\alpha(K) < x < r(K) \Longrightarrow g^2(x) < x, \tag{P.55b}$$
$$r(K) < x < \beta(K) \Longrightarrow g^2(x) > x, \tag{P.55c}$$
$$x > r(K) \Longrightarrow g^2(x) < x, \tag{P.55d}$$
$$x \in \{\alpha(K), r(K), \beta(K)\} \Longrightarrow g^2(x) = x, \tag{P.55e}$$

Remarque P.5. Notons que l'on a dans ce cas :

$$\left|\left(g^2\right)'(\alpha(K))\right| < 1, \tag{P.56a}$$
$$\left|\left(g^2\right)'(\beta(K))\right| < 1, \tag{P.56b}$$
$$\left|\left(g^2\right)'(r(K))\right| > 1. \tag{P.56c}$$

— La preuve de (P.56c) se fait comme celle de (P.18b) en utilisant (P.20) et (P.43).
— Montrons (P.56b). Pour cela, nous allons utiliser la convexité de g^2. D'après (P.27), on a

$$\left(g^2\right)''(x) = (1 - e^x)e^{x+K-e^x},$$

et donc

$$\forall x > 0, \left(g^2\right)''(x) < 0, \tag{P.57a}$$
$$\forall x < 0, \left(g^2\right)''(x) > 0. \tag{P.57b}$$

Or, on a $0 < r(K) < \beta(K)$ donc $g\left(g^2\right)''$ est strictement négative sur l'intervalle $[r(K), \beta(K)]$ et donc g^2 est strictement concave, c'est-à-dire, $-g^2$ strictement convexe sur l'intervalle $[r(K), \beta(K)]$. On renvoie à [RDO88, section 4.5.1] Cette convexité implique que la dérivée de $-g^2$ en $\beta(K)$ est strictement supérieure à la pente de $-g^2$ entre les points $r(K$ et $\beta(K)$:

$$\left(-g^2\right)'(\beta(K) > \frac{-g^2(\beta(K)) + g^2(r(K))}{\beta(K) - r(K)},$$

ce qui implique

$$\left(-g^2\right)'(\beta(K) > \frac{-\beta(K) + r(K)}{\beta(K) - r(K)} = -1,$$

et donc

$$\left(g^2\right)'(\beta(K) < 1.$$

Ainsi, compte tenu de (P.28), on a

$$0 < \left(g^2\right)'(\beta(K) < 1,$$

ce qui montre (P.56b).
Illustrons cela en traçant les divers éléments calculés sur la figure P.5.
— Montrons (P.56a). Cette fois, nous n'utiliserons pas la convexité de g^2 sur la totalité de l'intervalle $[\alpha(K), r(K)]$. Mais, on a $\alpha(K) < 0$ donc d'après (P.57b), g^2 est strictement convexe sur l'intervalle $[\alpha(K), \alpha(K)/2]$ et comme précédemment, on en déduit

$$0 < \left(g^2\right)'(\alpha(K)) < \frac{g^2(\alpha(K)/2) - g^2(\alpha(K))}{\alpha(K)/2 - \alpha(K)}. \tag{P.58}$$

Or, on a $\alpha(K)/2 \in]\alpha(K), r(K)[$ et d'après (P.55b), on a $g^2(\alpha(K)/2) < \alpha(K)/2$ et d'après (P.58), on a

$$0 < \left(g^2\right)'(\alpha(K)) < \frac{\alpha(K)/2 - g^2(\alpha(K))}{\alpha(K)/2 - \alpha(K)} = \frac{\alpha(K)/2 - \alpha(K)}{\alpha(K)/2 - \alpha(K)} = 1,$$

ce qui nous permet de conclure.
Illustrons cela en traçant les divers éléments calculés sur la figure P.6.

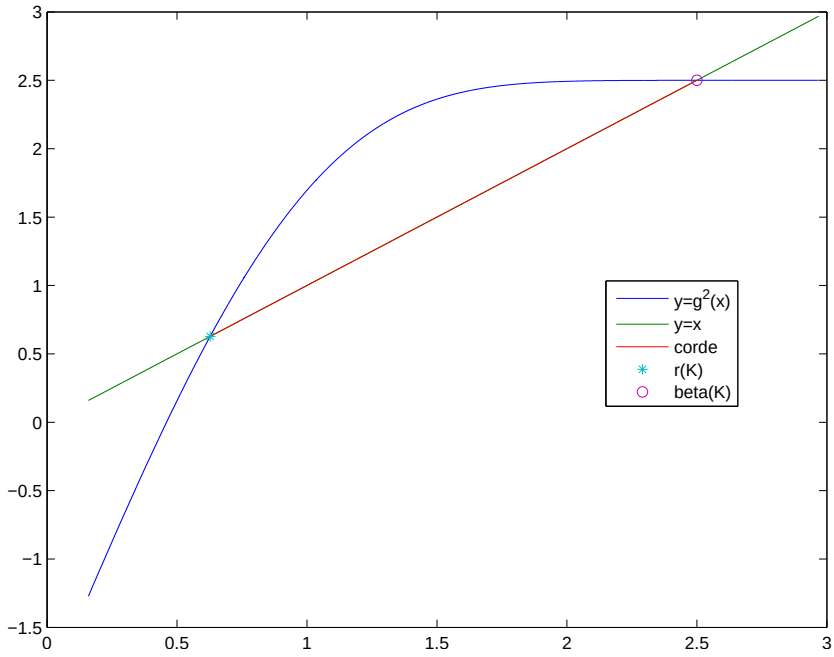


FIGURE P.5. La fonction g^2 sur l'intervalle $[r(K), \beta(K)]$ pour $K = 5/2$.

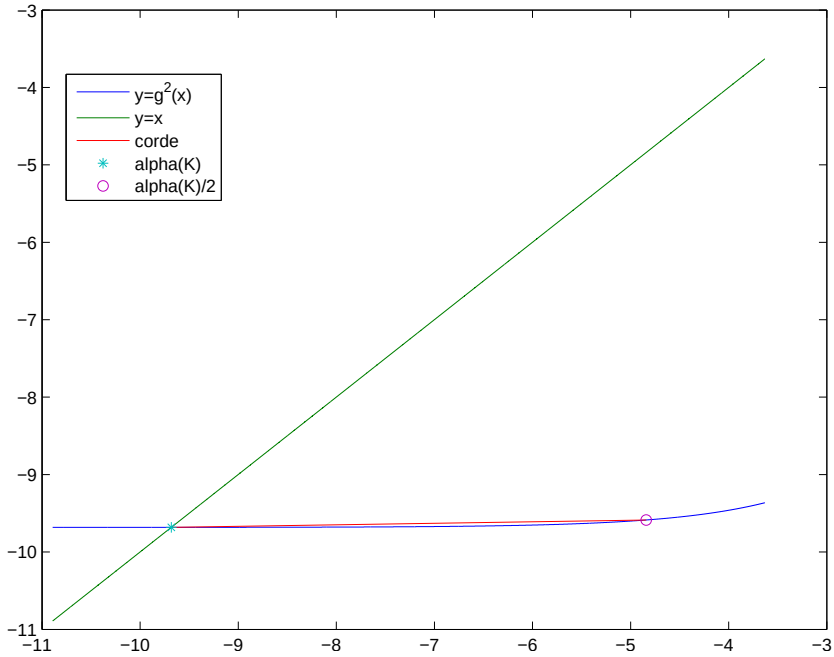


FIGURE P.6. La fonction g^2 sur l'intervalle $[\alpha, \alpha/2]$ pour $K = 5/2$.

Sur la figure (P.7), on constate que $\left(g^2\right)'(\alpha(K))$ et $\left(g^2\right)'(\beta(K))$ semblent appartenir à $]0, 1[$, pour tout K, ce qui est confirmé numériquement par

$$\min_{K\in]1,15]} \left(g^2\right)'(\alpha(K)) = 0,$$

$$\max_{K\in]1,15]} \left(g^2\right)'(\alpha(K)) = 0.997202351770,$$

$$\min_{K\in]1,15]} \left(g^2\right)'(\beta(K)) = 0,$$

$$\max_{K\in]1,15]} \left(g^2\right)'(\beta(K)) = 0.997202351770.$$

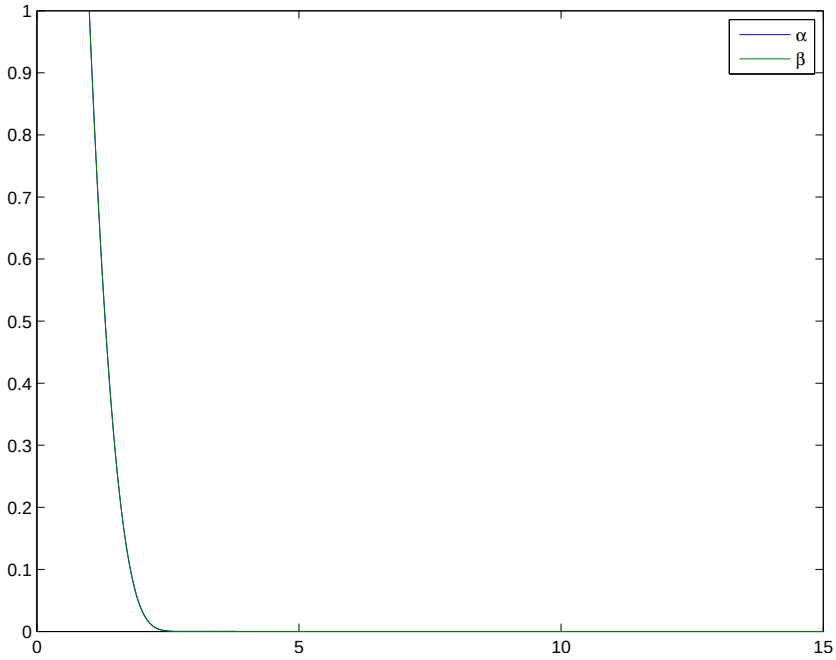


FIGURE P.7. Les courbes $\big(g^2\big)'(\alpha(K))$ et $\big(g^2\big)'(\beta(K))$ en fonction de K pour $K \in]1, 15]$.

De plus, il semblerait sur cette figure que

$$\forall K, \quad \big(g^2\big)'(\alpha(K)) = \big(g^2\big)'(\beta(K)),$$

ce qui est confirmé numériquement par

$$\max_{[K\in]1,15]} \Big|\big(g^2\big)'(\alpha(K)) - \big(g^2\big)'(\beta(K))\Big| = 2.55351\,10^{-15}.$$

Illustrons les trois types de points fixes de g^2 selon les valeurs de K donnés par $K = 1$, $K = 5/2$ et $K = -1$, par la figure P.8.

(b) Concluons maintenant sur la convergence de la suite (v_n) du point fixe associée à g^2, c'est-à-dire définie par

$$v_{n+1} = g^2(v_n) \text{ et } v_0 \in \mathbb{R}. \tag{P.59}$$

Considèrons les intervalles de $\mathbb{R}$ définis par

- Si $K \leq 1$,

$$I_1 =]-\infty, r(K)[, \tag{P.60a}$$
$$I_2 =]r(K), +\infty[, \tag{P.60b}$$

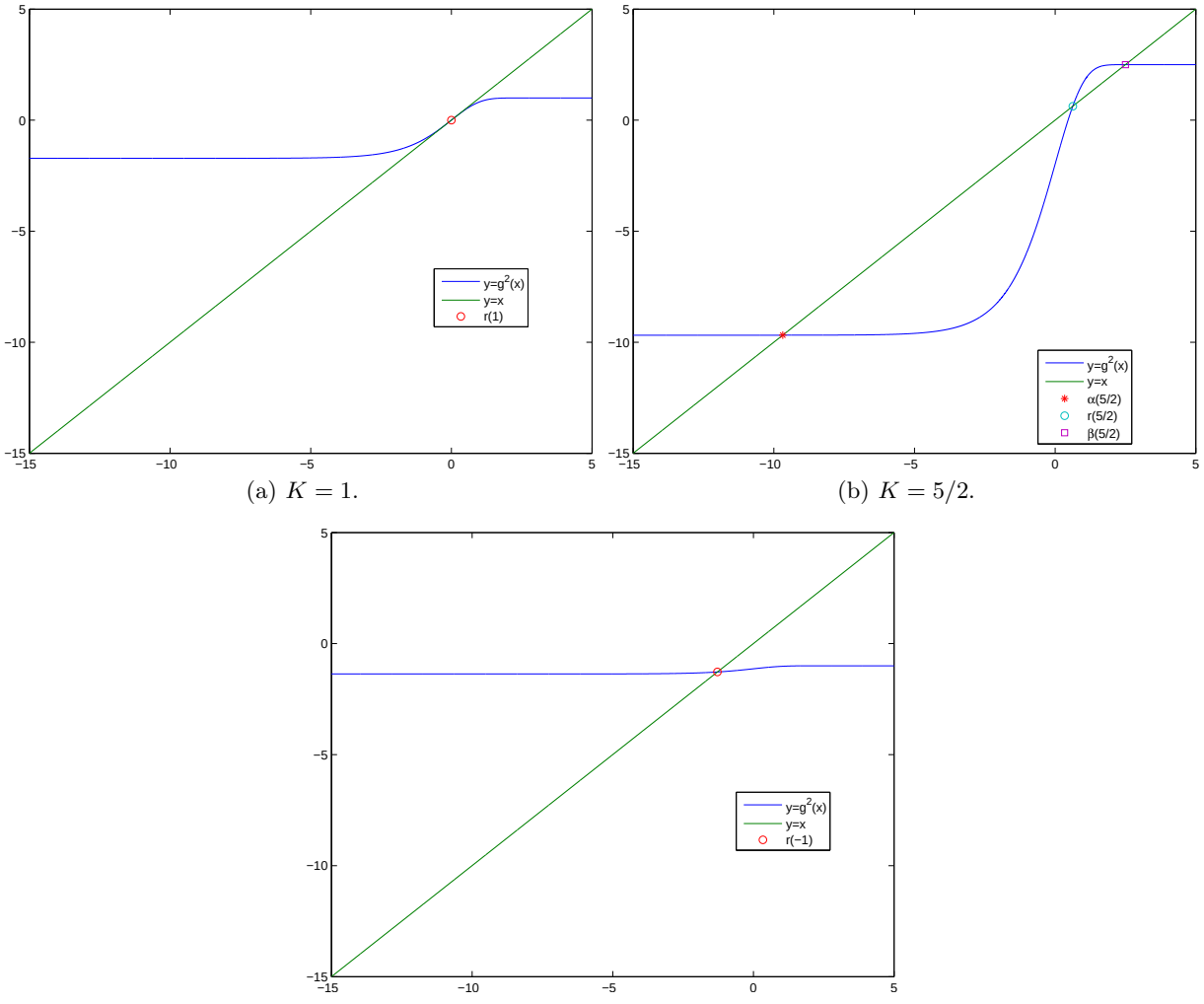
- Si $K > 1$,

$$I_3 =]-\infty, \alpha(K)[, \tag{P.61a}$$
$$I_4 =]\alpha(K), r(K)[, \tag{P.61b}$$
$$I_5 =]r(K), \beta(K)[, \tag{P.61c}$$
$$I_6 =]\beta(K), +\infty[, \tag{P.61d}$$

Remarquons d'abord que, d'après la croissance stricte de g^2,

$$\text{chacun des intervalles } J_k \text{ est } g^2\text{-stable.} \tag{P.62}$$

En effet, par exemple, $x \in I_3$, on a $x < \alpha(K)$ et donc $g^2(x) < g^2(\alpha(K)) = \alpha(K)$. Si par exemple, $x \in I_4$, on a $\alpha(K) < x < r(K)$ et donc $g^2(\alpha(K)) < g^2(x) < g^2(r(K))$, ce qui implique $\alpha(K) < g^2(x) < r(K)$.



(a) $K = 1$. (b) $K = 5/2$.

(c) $K = -1$.

FIGURE P.8. Les graphiques de la fonction g^2 sur l'intervalle $[-15, 5]$ pour différentes valeurs de K.

Remarquons aussi que si v_0 est égal à $r(K)$ si $K \leq 1$ ou si v_0 appartient à $\{\alpha(K), r(K), \beta(K)\}$ si $K > 1$, alors la suite v_n est constante et égale à v_0.

Supposons maintenant que v_0 appartient à l'un des intervalles I_k. Alors, pour tout n, v_n appartient à I_k, ce qui se montre par récurrence sur n en utilisant (P.62).

Enfin, v_0 appartient à l'un des intervalles I_k, alors la suite v_n est strictement monotone, à terme dans I_k. En effet, si par exemple v_0 appartient à I_4, alors, pour tout n est dans I_4 et d'après (P.55b) appliqué à $x = v_n$, on a $v_{n+1} = g^2(v_n) < v_n$. La suite v_n étant monotone et bornée (dans $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$), elle converge vers $l \in \overline{\mathbb{R}}$. De plus, d'après (P.40) et (P.55), si v_0 appartient à I_k, pour $k \in \{1, 3, 5\}$, la suite v_n est strictement croissante et majorée dans $\mathbb{R}$ et, pour $k \in \{2, 4, 6\}$, la suite v_n est strictement croissante et minorée dans $\mathbb{R}$. Dans tous les cas, la suite v_n converge donc vers l, qui est point fixe de g^2, puisque g^2 est continue (on passe à la limite $n \to \infty$ dans (P.59))

Or, les seuls points fixes de g^2 sont nécessairement $\alpha(K)$ ou $r(K)$ ou $\beta(K)$. La limite de v_n est donc nécessairement l'un de ces trois réels. Dans le cas où $K \leq 1$, le seul point fixe de g^2 est $r(K)$, qui est donc la valeur de l.

Si $K > 1$, montrons que l vaut $\alpha(K)$ ou $\beta(K)$.

— Si v_0 appartient à I_3, v_n est croissante et est à valeur dans I_3. La valeur de l est donc le seul point fixe de g^2 qui est dans l'adhérence de I_3, c'est-à-dire $]-\infty, \alpha(K)]$, qui ne peut être que $\alpha(K)$.
— Il en est de même si v_0 appartient à I_6 avec $l = \beta(K)$.
— Si v_0 appartient à I_4, v_n est décroissante et est à valeur dans I_4. La valeur de l est donc le seul point fixe de g^2 qui est dans l'adhérence de I_4, c'est-à-dire $[\alpha(K), r(K)]$. Puisque v_n est strictement décroissante, ce ne peut pas être $r(K)$ et on a donc $l = \alpha(K)$.
— Il en est de même si v_0 appartient à I_5 avec $l = \alpha(K)$.

Bref, on a montré que

$$\text{Si } K \leq 1, \text{ alors pour tout } v_0 \in \mathbb{R}, \text{ la suite } v_n \text{ converge vers } r(K). \tag{P.63a}$$
$$\text{Si } K > 1 \text{ et si } v_0 = r(K), \text{ alors la suite } v_n \text{ est constante et vaut } r(K). \tag{P.63b}$$
$$\text{Si } K > 1 \text{ et si } v_0 < r(K), \text{ alors la suite } v_n \text{ converge vers } \alpha(K). \tag{P.63c}$$
$$\text{Si } K > 1 \text{ et si } v_0 > r(K), \text{ alors la suite } v_n \text{ converge vers } \beta(K). \tag{P.63d}$$

Remarque P.6. On a étudié dans les remarques P.2, P.4 et P.5, les aspects attractifs de $r(K)$ et répulsifs de $\alpha(K)$ et $\beta(K)$ (voir remarque 4.15 page 85 du cours). Il est intéressant de constater que la méthode du point fixe ne converge que vers des points attractifs et non vers des point répulsifs. Attention, les aspects attractif ou répulsif ne suffisent pas à montrer la convergence ou la divergence de la méthode ; ces propriétés découlent ici de l'étude globale de la suite (voir propositions 4.11 et 4.12).

Remarque P.7. Ce résultat pouvait être démontré en utilisant la proposition N.1.

(c) Concluons par quelques simulations confirmant cela.

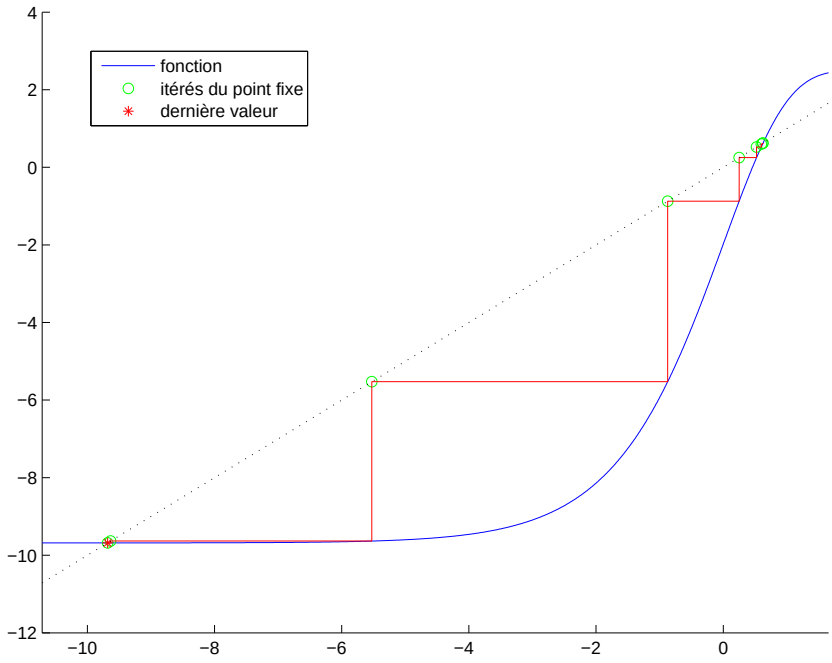


Figure P.9. Les 12 premières valeurs de v_n pour $K = 5/2$ et $x_0 = 5/8$.

Quelsques simulations ont été faites : Voir les figures P.9, P.10 et P.11, et les tableaux P.7 P.8 et P.9.
On peut évaluer $r(-1)$ et $\alpha(5/2)$ et $\beta(5/2)$: on obtient

$$r(-1) = -1.278464542761074, \tag{P.64a}$$
$$\alpha(5/2) = -9.681733635899509, \tag{P.64b}$$
$$\beta(5/2) = 2.499937586791849 \tag{P.64c}$$

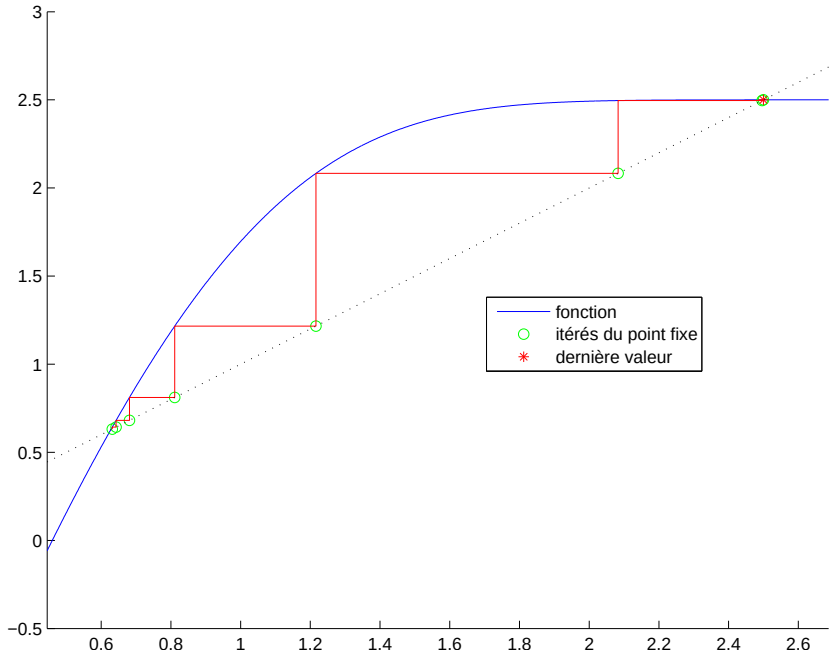


FIGURE P.10. Les 11 premières valeurs de v_n pour $K = 5/2$ et $x_0 = g\,(5/8)$.

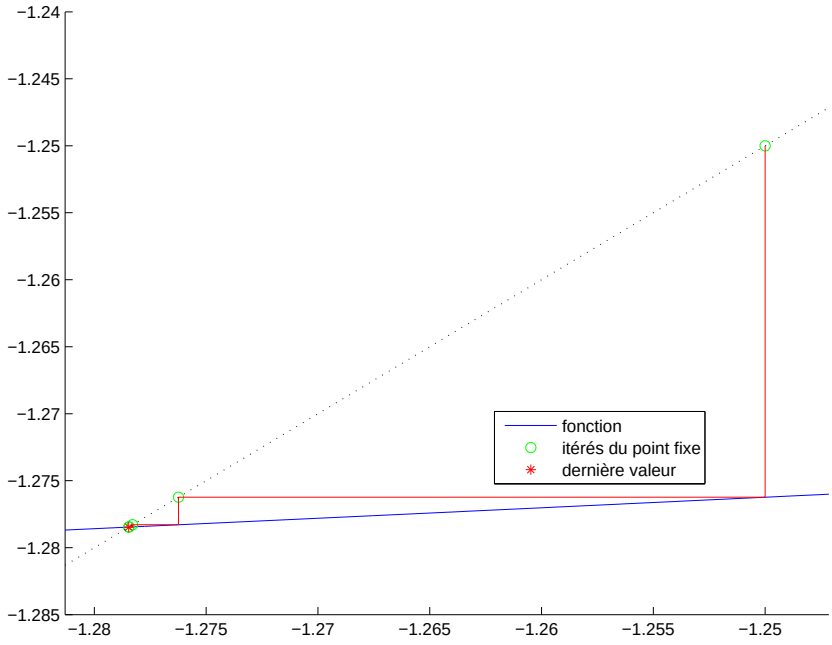


FIGURE P.11. Les 13 premières valeurs de v_n pour $K = -1$ et $x_0 = -5/4$.

et on obtient pour chacune des trois simulations :

$$|v_{13} - r(-1)| = 0, \tag{P.65a}$$

$$|v_{12} - \alpha(5/2)| = 0, \tag{P.65b}$$

$$|v_{11} - \beta(5/2)| = 0, \tag{P.65c}$$

n	v_n
0	0.62500000000000
1	0.61909312177470
2	0.59828327950539
3	0.52413439208549
4	0.24983444603184
5	−0.87429745515325
6	−5.52727415891270
7	−9.63414264798875
8	−9.68169657765000
9	−9.68173360772356
10	−9.68173363587809
11	−9.68173363589949
12	−9.68173363589951

TABLE P.7. Les 12 premières valeurs de v_n pour $K = 5/2$ et $x_0 = 5/8$.

n	v_n
0	0.63175404256778
1	0.64275701533088
2	0.68100658482373
3	0.81100379305089
4	1.21618714121982
5	2.08284501315852
6	2.49602318550744
7	2.49993454467100
8	2.49993758447888
9	2.49993758679009
10	2.49993758679185
11	2.49993758679185

TABLE P.8. Les 11 premières valeurs de v_n pour $K = 5/2$ et $x_0 = g(5/8)$.

ce qui confirme (P.63).

(d) Étudions maintenant la convergence de la suite x_n définie par (P.24).

(i) Dans le cas où $x_0 = r(K)$, d'après (P.17c), on a pour tout n, $x_n = r(K)$.

(ii) Si $x_0 < r(K)$, d'après (P.17b) appliqué à x_0, on a $x_1 = g(x_0) > r(K)$; puis, d'après (P.17a) appliqué à x_1, on a $x_2 = g(x_1) < r(K)$. On montre aisément par récurrence sur n que

$$x_0 < r(K) \Longrightarrow \forall n, \quad x_{2n} < r(K) \text{ et } x_{2n+1} > r(K). \tag{P.66}$$

De même, on montre aisément par récurrence sur n que

$$x_0 > r(K) \Longrightarrow \forall n, \quad x_{2n} > r(K) \text{ et } x_{2n+1} < r(K). \tag{P.67}$$

On considère ensuite les deux suites w_n et z_n des termes de rangs pairs et impairs définies par

$$\forall n \in \mathbb{N}, w_n = x_{2n}, \quad z_n = x_{2n+1}. \tag{P.68}$$

n	v_n
0	−1.25000000000000
1	−1.27623459377097
2	−1.27829148779366
3	−1.27845112280818
4	−1.27846350213932
5	−1.27846446206863
6	−1.27846453650398
7	−1.27846454227588
8	−1.27846454272345
9	−1.27846454275816
10	−1.27846454276085
11	−1.27846454276106
12	−1.27846454276107
13	−1.27846454276107

TABLE P.9. Les 13 premières valeurs de v_n pour $K = -1$ et $x_0 = -5/4$.

On alors

$$w_{n+1} = x_{2n+2} = g(x_{2n+1}) = g(g(x_{2n})) = g^2(w_n),$$
$$z_{n+1} = x_{2n+3} = g(x_{2n+2}) = g(g(x_{2n+1})) = g^2(z_n),$$

et autrement dit,

les deux suites w_n et z_n sont les deux suites du point fixe associées à la fonction g^2
par $x_{n+1} = g^2(x_n)$ de premiers termes respectifs x_0 et x_1. (P.69)

Ainsi, d'après (P.66) et (P.67), on a

$$x_0 < r(K) \Longrightarrow \forall n, \quad w_n < r(K) \text{ et } z_n > r(K), \tag{P.70a}$$
$$x_0 > r(K) \Longrightarrow \forall n, \quad w_n > r(K) \text{ et } z_n < r(K) \tag{P.70b}$$

Enfin, on n'a plus qu'à utiliser (P.63) :

Si $K \leq 1$, les deux suites w_n et z_n convergent toutes les deux vers $r(K)$, (P.71a)

et donc la suite x_n converge vers $r(K)$.

Si $K > 1$, les deux suites w_n et z_n convergent respectivement vers $\alpha(K)$ et $\beta(K)$ si $x_0 < r(K)$, (P.71b)

Si $K > 1$, les deux suites w_n et z_n convergent respectivement vers $\beta(K)$ et $\alpha(K)$ si $x_0 > r(K)$, (P.71c)

et donc, $K > 1$, puisque $\alpha(K) \neq \beta(K)$, la suite x_n ne converge donc pas.

Bref, pour récapituler, on a montré que, pour tout K dans $\mathbb{R}$.

Si $x_0 = r(K)$, la suite x_n est constante égale à $r(K)$. (P.72a)

Si $x_0 \neq r(K)$ et $K \leq 1$, la suite x_n converge vers $r(K)$. (P.72b)

Si $x_0 \neq r(K)$ et $K > 1$, la suite x_n est divergente. (P.72c)

REMARQUE P.8. Voir de nouveau la remarque P.6.

REMARQUE P.9. Ce résultat pouvait être démontré en utilisant la proposition N.2.

(e) Les simulations numériques et le corrigé des questions 2a et 2b, illustreront les résultats (P.72) établis.

◇

(a) On pose, dans cette question ;

$$a = -2, \tag{P.73a}$$

$$b = -1/2, \tag{P.73b}$$

$$K = -1. \tag{P.73c}$$

(i) Montrons, qu'avec ces valeurs, la méthode du point fixe converge pour tout $x_0 \in [a, b]$. Il suffit d'utiliser directement la proposition 4.19.

La fonction g étant décroissante, l'intervalle $[a, b]$ est g-stable si l'on a

$$a \leq g(b) \text{ et } g(a) \leq b, \tag{P.74}$$

ce qu'on vérifie numériquement.

La fonction g' donnée par (P.19) étant décroissante et négative, on a donc

$$\max_{x\in[a,b]} |g'(x)| = e^{\max\{a,b\}}, \tag{P.75}$$

qui est strictement plus petit que 1 si a et b sont strictement négatifs, ce qui est le cas ici. On a donc

$$k = \max_{x\in[a,b]} |g'(x)|\,, \tag{P.76}$$

où

$$k = e^{-1/2} \text{ avec } k < 1. \tag{P.77}$$

Les deux hypothèses de la proposition 4.19 sont vérifiées ce qui permet de conclure.

(ii) La valeur de n telle que

$$|x_n - r| \leq \varepsilon \tag{P.78}$$

avec

$$\varepsilon = 10^{-2}, \tag{P.79}$$

est donnnée par la proposition 4.21. Avec k donné par (P.77), on obtient numériquement

$$n = 11. \tag{P.80}$$

(iii)

On a affiché sur la figure P.12 page suivante, les 12 premières valeurs de x_n pour $K = -1$ et $x_0 = -5/4$. On a indiqué dans le tableau P.10 page suivante, les valeurs correspondantes (en allant plus loin, jusqu'à $n = 29$), en séparant les termes d'indices impairs et pairs.

Cela est conforme au résultat (P.72b), puisqu'ici, $K \leq 1$ et $r(-1)$ est donné par (P.64a). Cela est aussi conforme aux résultat (P.71a) : les deux suites de rangs pairs et impairs convergent toutes les deux vers $r(-1)$, en étant l'une inférieure, l'autre supérieure à $r(-1)$ et étant toutes les deux monotones.

Remarque P.10. Notons qu'il est possible de déterminer de façon explicite, grâce à la fonction W de Lambert la valeur de $r(K)$, pour tout K réel. Voir la section E.2. On a ici

$$a = 1, \tag{P.81a}$$

$$b = 1, \tag{P.81b}$$

$$c = -K \tag{P.81c}$$

et donc

$$\Delta = e^K \in \mathbb{R}_+^*, \tag{P.82}$$

et on a donc d'après (E.11), la solution unique $r(K)$,donnée par

$$r(K) = K - W_0(e^K). \tag{P.83}$$

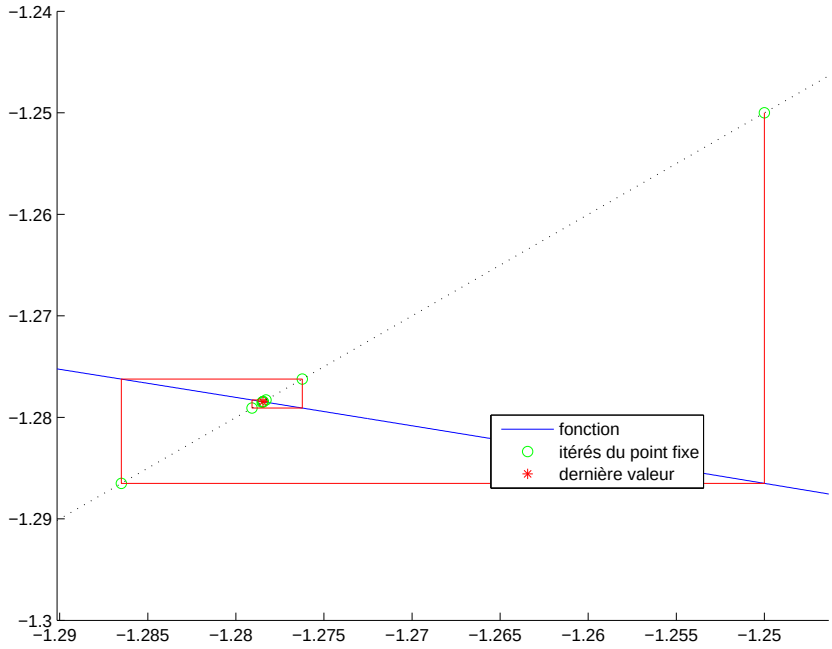


FIGURE P.12. Les 12 premières valeurs de x_n pour $K = -1$ et $x_0 = -5/4$.

n	x_{2n}	x_{2n+1}
0	−1.25000000000000	−1.28650479686019
1	−1.27623459377097	−1.27908619735840
2	−1.27829148779366	−1.27851273660342
3	−1.27845112280818	−1.27846827976719
4	−1.27846350213932	−1.27846483253749
5	−1.27846446206863	−1.27846456523106
6	−1.27846453650398	−1.27846454450345
7	−1.27846454227588	−1.27846454289618
8	−1.27846454272345	−1.27846454277155
9	−1.27846454275816	−1.27846454276189
10	−1.27846454276085	−1.27846454276114
11	−1.27846454276106	−1.27846454276108
12	−1.27846454276107	−1.27846454276107
13	−1.27846454276107	−1.27846454276107
14	−1.27846454276107	−1.27846454276107

TABLE P.10. Les 30 premières valeurs de x_n pour $x_0 = -5/4$ et $K = -1$.

Cela est confirmé par matlab :

```
solve('e^x=-x+K','x')
```

qui donne

$$\frac{-LambertW\left(\ln(e)\,e^{\ln(e)K}\right)+\ln(e)\,K}{\ln(e)}$$

la fonction `lambertw` est programmée sous matlab et on peut donc utiliser directement (P.83). Numériquement, on a

$$r(-1) = -1.2784645427610739, \tag{P.84}$$

ce qui confirme la valeur finale du tableau P.10. De plus, pour n donné par (P.80), on a

$$|x_n - r(-1)| = 2.24699\,10^{-8},$$

ce qui confirme *a posteriori* le choix de n défini par la majoration (P.78).

(b) On pose, dans cette question ;

$$a = 1/4, \tag{P.85a}$$
$$b = 1, \tag{P.85b}$$
$$K = 5/2. \tag{P.85c}$$

(i) Comme dans la question 2a, la fonction g' donnée par (P.19) est décroissante et négative et on a donc

$$\min_{x\in[a,b]} |g'(x)| = e^{\min\{a,b\}}, \tag{P.86}$$

qui est strictement plus grand que 1 si l'un des réels a et b est strictement positif, ce qui est le cas ici. On a donc

$$k = \min_{x\in[a,b]} |g'(x)|\,, \tag{P.87}$$

où

$$k = e^{1/4} \text{ avec } k > 1. \tag{P.88}$$

Par ailleurs, on a

$$\text{signe}(f(a)f(b)) = -1, \tag{P.89}$$

et donc l'intervalle $[a, b]$ contient l'unique racine $r(K)$ de f. Donc, de (P.87) et (P.88), on peut déduire que

$$|g'(r(K))| > 1 \tag{P.90}$$

et le point $r(K)$ est dit répulsif. Voir la remarque 4.15. *Attention*, l'égalité (P.90) ne permet pas *a priori* d'affirmer que la méthode du point fixe est divergente.

Pour cela, on peut utiliser par exemple la proposition 4.12. Ses trois hypothèses sont vérifiées. En effet, on pose $I = [a, b]$ et on successivement :

— g est définie sur $\mathbb{R}$ donc sur I.

— La propriété (4.32b) est vérifiée. En effet, On utilise la remarque 4.13.La fonction g étant décroissante, l'intervalle $\mathbb{R} \setminus [a, b]$ est g-stable si l'on a

$$g(b) \le a \text{ et } g(a) \ge b, \tag{P.91}$$

ce qu'on vérifie numériquement.

— Enfin, la propriété (4.32c) est vérifiée d'après (P.87) et (P.88).

On peut donc conclure que la suite x_n diverge pour tout $x_0 \in [a, b] \setminus \{r(K)\}$.

La preuve de ce résultat a été établie dans le cas général pour tout $K > 1$ et pour tout $x_0 \neq r(K)$. Voir (P.72c).

♢

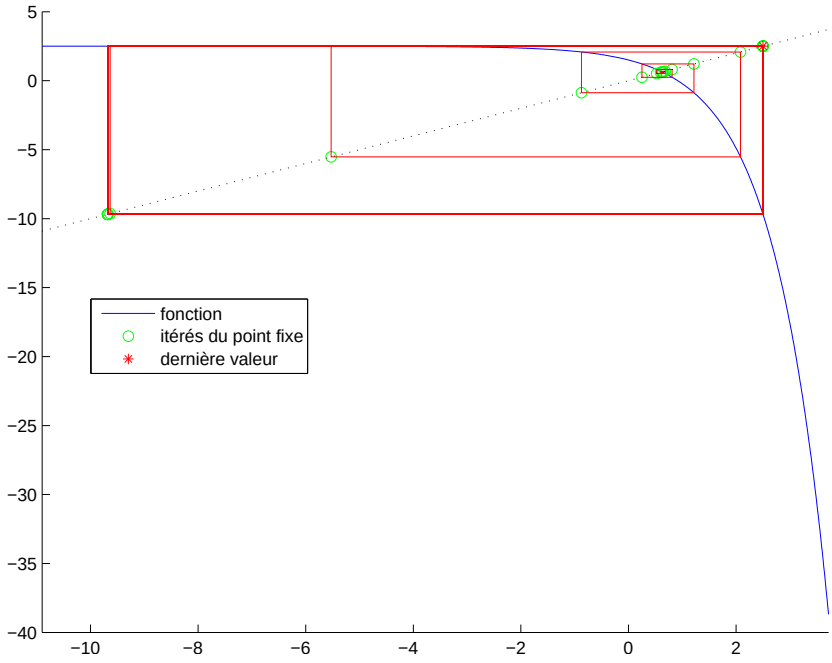
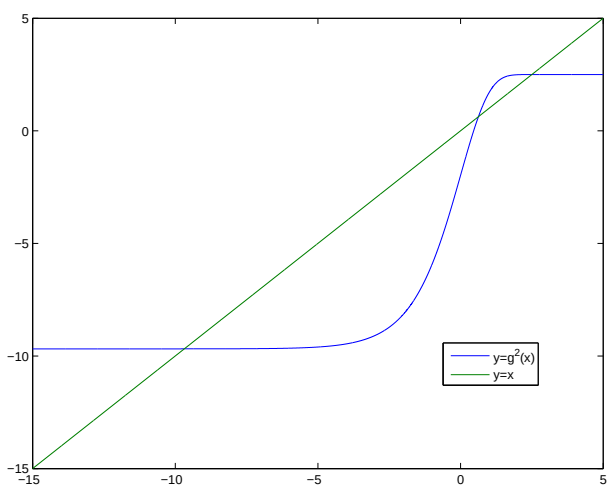


FIGURE P.13. Les 30 premières valeurs de x_n pour $x_0 = 5/8$ et $K = 5/2$.

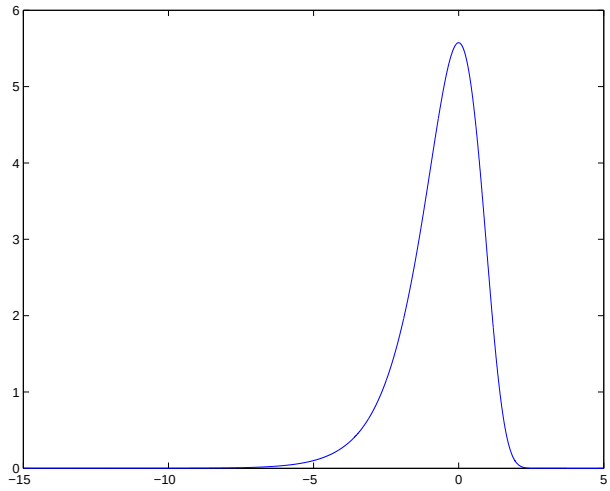
n	x_{2n}	x_{2n+1}
0	0.62500000000000	0.63175404256778
1	0.61909312177470	0.64275701533088
2	0.59828327950539	0.68100658482373
3	0.52413439208549	0.81100379305089
4	0.24983444603184	1.21618714121982
5	−0.87429745515325	2.08284501315852
6	−5.52727415891270	2.49602318550744
7	−9.63414264798875	2.49993454467100
8	−9.68169657765000	2.49993758447888
9	−9.68173360772356	2.49993758679009
10	−9.68173363587809	2.49993758679185
11	−9.68173363589949	2.49993758679185
12	−9.68173363589951	2.49993758679185
13	−9.68173363589951	2.49993758679185
14	−9.68173363589951	2.49993758679185

TABLE P.11. Les 30 premières valeurs de x_n pour $x_0 = 5/8$ et $K = 5/2$.

(ii)



(a) g^2



(b) $(g^2)'$

FIGURE P.14. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[-15, 5]$.

On a affiché sur la figure P.13 page précédente, les 30 premières valeurs calculées avec les paramètres donnés par (P.85) et x_0 donné par

$$x_0 = 5/8. \tag{P.92}$$

On a indiqué dans le tableau P.11 page précédente, les valeurs correspondantes, en séparant les termes d'indices impairs et pairs.

(A)

On constate sur la figure P.13 page précédente et surtout sur le tableau P.11 page précédente, que la suite x_n semble ne pas converger. Plus précisément, il semble apparaître que la suite des termes de rangs pairs convergerait en décroissant vers la valeur

$$l_p = -9.6817336358995085, \tag{P.93}$$

et que la suite des termes de rangs impairs convergerait en croissant vers la valeur

$$l_i = 2.4999375867918490, \tag{P.94}$$

(B)

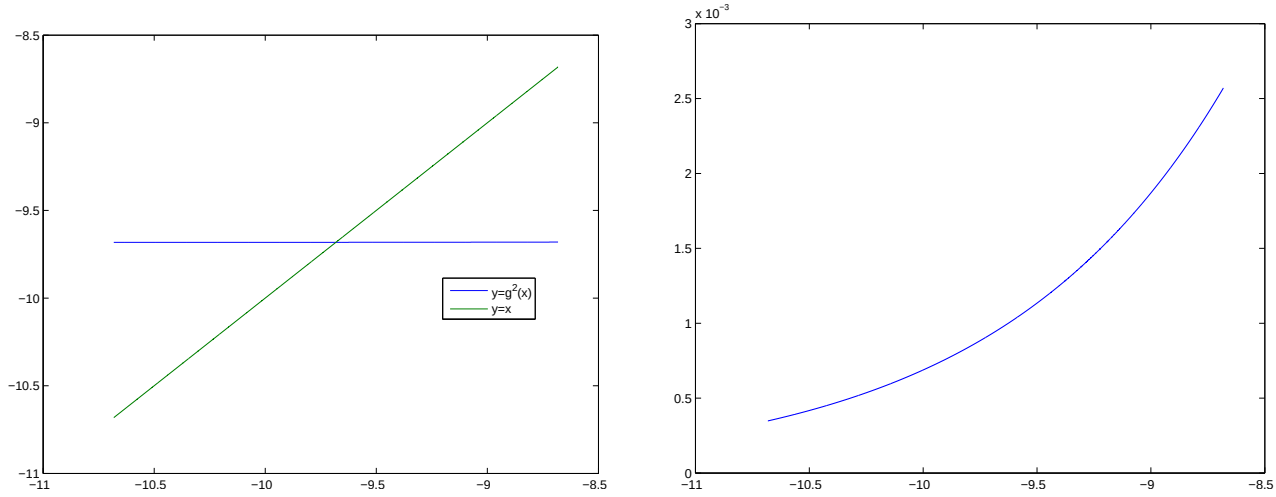
Tâchons d'expliquer de façon qualitative les observations faites dans la question 2(b)iiA grâce aux figures P.14, P.15 et P.16. En fait, la suite x_{2n} des termes de rangs pairs vérifie la relation de récurrence

$$x_{2n+2} = g(x_{2n+1}) = g(g(x_{2n})) = g^2(x_{2n}),$$

où g^2 est donnée par (P.25). De même, la suite x_{2n+1} des termes de rangs impairs vérifie la relation de récurrence

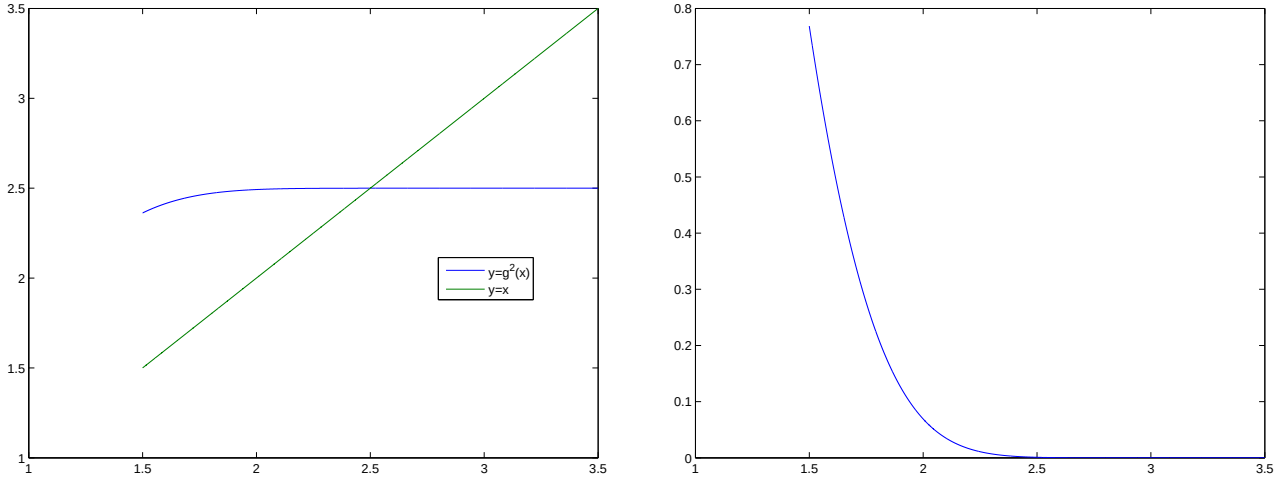
$$x_{2n+3} = g(x_{2n+2}) = g(g(x_{2n+1})) = g^2(x_{2n+1}).$$

Le comportement de la fonction g^2 permet donc de prédire la convergence des suites x_{2n} et x_{2n+1}. On constate sur la figure P.14, que la fonction g^2 semble avoir trois points fixes dont les valeurs numériques sont proches des trois valeurs données par (P.64). Il semblerait que le plus grand et le plus petit point fixe soient des points attractifs pour la fonction g^2 comme le montrent les figures P.15 et P.16, et que le point fixe intermédiaire soit répulsif



(a) g^2 (b) $(g^2)'$

FIGURE P.15. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[-10.68173, -8.68173]$.



(a) g^2 (b) $(g^2)'$

FIGURE P.16. Les graphiques des fonction g^2 et $(g^2)'$ sur l'intervalle $[1.49994, 3.49994]$.

comme le montre la figure (14(b)). On peut confirmer cela puisque que l'on peut vérifier numériquement qu'en ces trois points fixes, on a

$$g'(r(5/2)) = 2.567854985668169, \tag{P.95a}$$
$$g'(\alpha(5/2)) = 0.000760301077059, \tag{P.95b}$$
$$g'(\beta(5/2)) = 0.000760301077059. \tag{P.95c}$$

On pourrait aussi montrer que les hypothèses de la proposition 4.19 sont valables pour la fonction g^2 sur des intervalles autour des valeurs du premier et du dernier point fixe de g^2 données par (P.64b) et (P.64c). Ainsi, il y aurait bien convergence des suites x_{2n} et x_{2n+1} vers ces valeurs.

De façon plus rigoureuse, il suffit d'utiliser les résultats (P.71b) et (P.72c), qui prévoient la divergence de la suite x_n et la convergence de x_{2n} et x_{2n+1} vers $\alpha(K)$ et $\beta(K)$ (distincts), puisque $x_0 < r(K)$. Enfin, on confirme cela de façon numérique puisque pour $n = 15$, on a

$$|x_{2n} - l_p| = 0,$$
$$|x_{2n+1} - l_i| = 0.$$

(iii) Pour résoudre (P.9) dont on sait que la solution $r(5/2)$ existe et est unique, on ne peut donc utiliser la méthode du point fixe qui diverge. En revanche, on peut utiliser la méthode de la dichotomie sur l'intervalle $[a, b]$, qui converge ici vers $r(5/2)$ d'après (P.89). On choisit n pour que la méthode de dichotomie fournisse $r(5/2)$ avec une erreur inférieure à ε donné par

$$\varepsilon = 10^{-16},$$

en utilisant la proposition 4.5. On obtient pour le n-ième milieu

$$x_n = 0.627352959583406,$$

et, en utilisant (P.83), qui fournit,

$$r(5/2) = 0.627352959583406,$$

on a

$$|x_n - r(5/2)| = 0.$$

Annexe Q

Un théorème de point fixe

Attention : cette annexe, rédigée avant le cours, a des notations parfois légèrement différentes de celui-là. Rappelons par exemple le théorème [BM03, Théorème 4.8] :

THÉORÈME Q.1 (condition suffisante de convergence de la méthode du point fixe).
Voir désormais le théorème 4.19, fondé sur l'hypothèse il existe un réel k de $[0,1[$ tel que

$$\forall x \in [\alpha, \beta] \quad |g'(x)| \leq k. \tag{Q.1}$$

DÉMONSTRATION. Voir preuve dans [BM03]. □

De cette preuve, il ressort que, pour tout $n \in \mathbb{N}$, on a

$$|x_{n+1} - l| \leq k\,|x_n - l| \tag{Q.2}$$

dont on déduit par récurrence sur n que pour tout $n \in \mathbb{N}$, on a

$$|x_n - l| \leq k^n\,|x_0 - l|\,, \tag{Q.3}$$

ce qui fournit aussi

$$|x_n - l| \leq k^n\,|\beta - \alpha|\,, \tag{Q.4}$$

et donc de retrouver le fait que x_n tend vers l. De plus, cette inégalité permet de déterminer le rang n à partir duquel

$$|x_n - l| \leq \varepsilon, \tag{Q.5}$$

pour tout $\varepsilon > 0$.

Dans le cas où l'intervalle $[\alpha, \beta]$ est trop grand, l'estimation (Q.3) n'est pas utile et on lui préfère la propriété suivante :

THÉORÈME Q.2 (Seconde conséquence sur la convergence de la méthode du point fixe).
Sous les hypothèses du théorème Q.1, on a pour tout $n \in \mathbb{N}^*$,

$$|x_n - l| \leq \frac{k^n}{1-k}\,|x_1 - x_0| = \frac{k^n}{1-k}\,|g(x_0) - x_0|\,. \tag{Q.6}$$

DÉMONSTRATION. Voir preuve dans [RDO88, Théorème du point fixe p. 63, section 2.4.3]. Ce théorème de point fixe se démontre dans un cadre beaucoup plus général que le théorème Q.1. Il repose sur la majoration suivante : pour tout $n \geq 1$: il existe $\xi \in [x_{n-1}, x_n]$ tel que

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| = |g'(\xi)|\,|x_n - x_{n-1}|\,,$$

et donc

$$|x_{n+1} - x_n| = |g'(\xi)|\,|x_n - x_{n-1}|\,. \tag{Q.7}$$

D'après (Q.1), il vient donc

$$|x_{n+1} - x_n| \leq k\,|x_n - x_{n-1}| \tag{Q.8}$$

et donc par récurrence sur n

$$|x_{n+1} - x_n| \leq k^n\,|x_1 - x_0|\,, \tag{Q.9}$$

On considère ensuite deux entiers n et p en supposant par exemple $p < n$. On a par inégalité triangulaire :

$$|x_p - x_n| \leq \sum_{q=0}^{n-p-1} |x_{p+q} - x_{p+q+1}| \tag{Q.10}$$

et donc, d'après (Q.9),

$$|x_p - x_n| \leq |x_1 - x_0| \sum_{q=0}^{n_p-1} k^{p+q}.$$

En utilisant la somme des termes d'une suite géométrique, il vient compte tenu de $0 \leq k < 1$,

$$|x_p - x_n| \leq \frac{k^p}{1-k} |x_1 - x_0| . \tag{Q.11}$$

Cette inégalité assure d'une part que la suite x_n a une limite égale à l, l'unique point fixe de g et d'autre part, en y faisant $n \to +\infty$:

$$\forall p \geq 1, \quad |x_p - a| \leq \frac{k^p}{1-k} |x_1 - x_0| . \tag{Q.12}$$

ce qui est donc (Q.6). □

Donnons une dernière majoration, utile parfois en pratique :

Lemme Q.3 (Autre conséquence sur la convergence de la méthode du point fixe).
Sous les hypothèses du théorème Q.1, on a pour tout $n \in \mathbb{N}^*$,

$$|x_n - l| \leq \frac{k}{1-k} |x_n - x_{n-1}| . \tag{Q.13}$$

Démonstration. On a effet, d'après (Q.1), de façon analogue à (Q.7)

$$\begin{aligned} |x_n - l| &= |g(x_{n-1}) - g(l)|, \\ &= |g'(\xi)||x_{n-1} - l|, \\ &\leq k|x_{n-1} - l|, \\ &\leq k(|x_{n-1} - x_n| + |x_n - l|), \end{aligned}$$

et donc

$$(1-k)|x_n - l| \leq k|x_{n-1} - x_n|,$$

ce qui permet de conclure car $1 - k > 0$. □

Ce résultat fournit une preuve alternative du théorème Q.2.

Autre preuve du théorème Q.2. On réécrit (Q.9) sous la forme

$$|x_n - x_{n-1}| \leq k^{n-1} |x_1 - x_0| ,$$

ce qui permet de conclure directement en utilisant (Q.13). □

Annexe R

Majoration de l'erreur pour une méthode d'ordre p

Montrons dans cette annexe le résultat du lemme 4.32 page 90 et de la proposition 4.33 page 91.

Démonstration du lemme 4.32.
Par définition, il existe une constante $C > 0$ telle que

$$\forall n \in \mathbb{N}, \quad \frac{|e_{n+1}|}{|e_n|^p} \leq C. \tag{R.1}$$

D'après l'inégalité (R.1) :

$$|e_1| \leq C|e_0|^p,$$

Puis, de nouveau d'après l'inégalité (R.1) :

$$|e_2| \leq C|e_1|^p = C(C|e_0|^p)^p = C^{1+p}|e_0|^{(p^2)},$$

Puis, de nouveau d'après l'inégalité (R.1) :

$$|e_3| \leq C|e_2|^p = C\Big(C^{1+p}|e_0|^{p^2}\Big)^p = C^{1+p+p^2}|e_0|^{(p^3)}.$$

On a alors

$$\forall n \in \mathbb{N}, \quad |e_n| \leq C^{1+p+p^2+\ldots+p^{n-1}}|e_0|^{(p^n)}. \tag{R.2}$$

Cela se montre par récurrence sur n.
Démontrons cela par récurrence sur $n \in \mathbb{N}$. Pour $n = 0$, on a bien

$$|e_0| \leq |e_0| = C^{\sum \emptyset}|e_0|^{(p^0)}.$$

Supposons maintenant (R.2) vraie pour n. Démontrons-là pour $n+1$. On a d'après l'inégalité (R.1) :

$$|e_{n+1}| \leq C|e_n|^p,$$

et donc on a successivement :

$$\begin{aligned}
|e_{n+1}| &\leq C\Big(C^{1+p+p^2+\ldots+p^{n-1}}|e_0|^{(p^n)}\Big)^p, \\
&= CC^{(p(1+p+p^2+\ldots+p^{n-1}))}\Big(|e_0|^{(p^n)}\Big)^p, \\
&= CC^{(p+p^2+p^3+\ldots+p^n)}|e_0|^{(p\times p^n)}, \\
&= C^{(1+p+p^2+p^3+\ldots+p^n)}|e_0|^{(p^{n+1})},
\end{aligned}$$

ce qui est bien (R.2) au rang $n+1$. $\diamondsuit$

- Si $p = 1$, on a

$$1 + p + p^2 + \ldots + p^{n-1} = n \times 1 = n,$$

et donc d'après (R.2),

$$\forall n \in \mathbb{N}, \quad |e_n| \leq C^n|e_0|.$$

ce qui permet de conclure.

Remarque R.1. Ce calcul est en fait identique à celui fait dans la preuve de la proposition 4.20 page 88.

- Sinon, on a $p \neq 1$ et donc
$$1 + p + p^2 + ... + p^{n-1} = \frac{p^n - 1}{p - 1}$$
et donc (R.2) est équivalent successivement à
$$\begin{aligned}\forall n \in \mathbb{N}, \quad |e_n| &\leq C^{\left(\frac{p^n-1}{p-1}\right)} |e_0|^{(p^n)}, \\ &\leq C^{\left(\frac{p^n}{p-1}\right)} C^{\left(\frac{-1}{p-1}\right)} |e_0|^{(p^n)}, \\ &\leq C^{\left(\frac{1}{1-p}\right)} \left(C^{\left(\frac{1}{p-1}\right)}\right)^{(p^n)} |e_0|^{(p^n)}, \\ &\leq C^{\left(\frac{1}{1-p}\right)} \left(|e_0| C^{\left(\frac{1}{p-1}\right)}\right)^{(p^n)},\end{aligned}$$
ce qui permet de conclure en considérant les nombres γ et δ défini par
$$\gamma = C^{\left(\frac{1}{1-p}\right)}, \tag{R.3a}$$
$$\delta = |e_0| C^{\left(\frac{1}{p-1}\right)}, \tag{R.3b}$$

□

Démonstration de la proposition 4.33 page 91.

(1) Si $p = 1$, il suffit de raisonner comme dans la preuve de la proposition 4.21.

(2) Sinon, $p > 1$ et si $|e_0|$ est assez petit (pour que (4.52) ait lieu), d'après (4.50b), on a $\delta < 1$ et la limite de $|x_n - r|$ est bien nulle quand n tend vers l'infini !

Pour avoir
$$|x_n - r| \leq \varepsilon$$
il suffit donc que
$$\gamma \delta^{(p^n)} \leq \varepsilon,$$
ce qui est équivalent à
$$\delta^{(p^n)} \leq \frac{\varepsilon}{\gamma}.$$
Si on prend le logarithme de cette inégalité, on arrive à
$$p^n \geq \frac{\ln \frac{\varepsilon}{\gamma}}{\ln \delta}.$$
Si on prend de nouveau le logarithme de cette inégalité, on arrive à
$$n \ln p \geq \ln \left(\frac{\ln \frac{\varepsilon}{\gamma}}{\ln \delta} \right).$$
et donc
$$n \geq \frac{1}{\ln p} \ln \left(\frac{\ln \frac{\varepsilon}{\gamma}}{\ln \delta} \right),$$
ce qui permet de conclure.

□

Annexe S

Convergence des méthodes d'ordre p

Attention : cette annexe, rédigée avant le cours, a des notations parfois légèrement différentes de celui-là.

Reprenons les résultats de l'annexe Q et adaptons-les à une méthode d'ordre p. Dans toute cette annexe, la suite (x_n) est définie par $x_{n+1} = g(x_n)$ et on suppose que cette suite converge vers l, point fixe de g. On note

$$\forall n \in \mathbb{N}, \quad e_n = x_n - l. \tag{S.1}$$

Rappelons la [BM03, Définition 4.3] : voir la définition 4.30.

Rappelons la [BM03, Proposition D.1].

Proposition S.1.

- *Si l'ordre de convergence p est égal à un, il existe $\delta \in [0,1[$ et un entier N tels que*
$$\forall n \in \mathbb{N}, \quad |e_{n+N}| \leq \delta^n |e_N|. \tag{S.2}$$
- *Si l'ordre de convergence p est strictement supérieur à un, il existe $\delta \in [0,1[$, $\gamma \in \mathbb{R}_+$ et un entier N tels que*
$$\forall n \in \mathbb{N}, \quad |e_{n+N}| \leq \gamma \delta^{(p^n)}. \tag{S.3}$$

Démonstration.

- Si l'ordre de convergence est égal à un, il existe $\delta \in [0,1[$ et un entier N tels que, pour tout entier k :
$$|e_{N+k+1}| \leq \delta \, |e_{N+k}| \, .$$
On en déduit
$$|e_{N+n}| \leq \delta^n \, |e_N| \, .$$
- De même, si l'ordre de convergence est strictement supérieur à un, il existe $D \in \mathbb{R}_+$ tel que, pour tout entier k et pour tout entier N,
$$|e_{N+k+1}| \leq D \, |e_{N+k}|^p \, .$$
Ainsi, pour $k = 0$, on a
$$|e_{N+1}| \leq D \, |e_N|^p \, .$$
Pour $k = 1$, il vient donc
$$|e_{N+2}| \leq D \, |e_{N+1}|^p \leq D(D \, |e_N|^p)^p = D^{1+p} \, |e_N|^{\left(p^2\right)} \, .$$
On montre par une récurrence immédiate sur n que
$$\forall n \in \mathbb{N}, \quad |e_{n+N}| \leq D^{1+p+p^2+\ldots+p^{n-1}} \, |e_N|^{(p^n)} \, ,$$
ce qui implique, puisque $p > 1$:
$$|e_{n+N}| \leq D^{\frac{p^n-1}{p-1}} |e_N|^{(p^n)} = D^{\frac{-1}{p-1}} \left(D^{\frac{1}{p-1}} \, |e_N| \right)^{(p^n)} .$$
La majoration (S.2) est valable puisque la convergence est *a fortiori* d'ordre 1. Ainsi e_n tend vers zéro et en prenant N assez grand
$$D^{\frac{1}{p-1}} \, |e_N| < 1,$$
ce qui nous permet de conclure.

□

REMARQUE S.2. Dans le cas où $p \geq 2$, notons que cette preuve fournit les valeurs de γ et δ données par (avec N assez grand)

$$\gamma = D^{\frac{1}{1-p}}, \tag{S.4a}$$

$$\delta = D^{\frac{1}{p-1}}|e_N| < 1. \tag{S.4b}$$

Si on note $[\alpha, \beta]$ l'intervalle d'étude, si on choisit $N = 0$ et que l'on suppose

$$D^{\frac{1}{p-1}}|\beta - \alpha| < 1, \tag{S.5}$$

alors on a

$$\gamma = D^{\frac{1}{1-p}}, \tag{S.6a}$$

$$\delta = D^{\frac{1}{p-1}}(\beta - \alpha), \tag{S.6b}$$

$$\forall n \in \mathbb{N}, \quad |e_n| \leq \gamma\delta^{(p^n)}. \tag{S.6c}$$

Rappelons le [BM03, Corollaire D.4.].

COROLLAIRE S.3. *Si l'ordre de convergence p est strictement supérieur à un, on a à l'infini*

$$r_{n+1} \sim pr_n, \tag{S.7}$$

ce qui signifie, que pour n assez grand, à chaque étape, le nombre de décimales exactes est asymptotiquement multiplié par l'ordre de convergence de la méthode.

DÉMONSTRATION. Voir [BM03, Corollaire D.4.]. □

Nous concluons en donnant une généralisation du résultat (Q.6) pour des méthodes d'ordre p. L'inconvénient de la majoration (S.6c) n'est pertinente d'après (S.6b) que si $\beta - \alpha$ est assez petit pour assurer $D < 1$.

THÉORÈME S.4. *On suppose que la fonction g est de classe C^p sur $[\alpha, \beta]$, qu'elle laisse $I = [\alpha, \beta]$ stable, que sont vérifiées les hypothèses* (4.87) *et* (4.88) *avec $p \geq 2$. On considère M telle que*

$$\forall x \in [\alpha, \beta], \quad \left|g^{(p)}(x)\right| \leq M. \tag{S.8}$$

et on suppose que (Q.1) *a lieu*[1]*. Posons*

$$A = \frac{M}{(p-1)!}(\alpha - \beta)^{p-1}, \tag{S.9a}$$

$$B = k^{p-1} \in [0, 1[. \tag{S.9b}$$

On pose, pour tout $n \in \mathbb{N}^*$,

$$F(n) = A^n B^{\frac{(n)(n+1)}{2}} \sum_{q=0}^{\infty} A^q B^{\frac{q^2+2nq+q}{2}}. \tag{S.10}$$

Alors

$$\lim_{n\to\infty} F(n) = 0 \tag{S.11}$$

et

$$\forall n \in \mathbb{N}^*, \quad |x_n - l| \leq |x_1 - x_0|\, F(n). \tag{S.12}$$

1. ce qui est vrai d'après (4.87) si $\beta - \alpha$ est assez petit.

DÉMONSTRATION. Pour tout $x \in [\alpha, \beta]$, la formule de Taylor-Young appliquée à g' sur $[l, x]$ implique qu'il existe $\xi \in [l, x]$ tel que

$$g'(x) = g'(l) + (x-l)g''(l) + ... + \frac{1}{(p-2)!}g^{(p-1)}(l)(x-l)^{p-2} + \frac{1}{(p-1)!}g^{p}(l)(x-l)^{p-1},$$

soit, compte tenu de (4.87) et de (S.8),

$$\forall x \in [\alpha, \beta]\, |g'(x)| \leq \frac{M}{(p-1)!}|x-l|^{p-1}, \tag{S.13}$$

Si, de plus $x \in [x_n, x_{n+1}]$ ou $x \in [x_{n+1}, x_n]$, alors

$$|x-l| \leq \max\left(|x_n - l|, |x_{n+1} - l|\right),$$

et, d'après (Q.3), toujours valable

$$\leq |x_0 - l| \max(k^n, k^{n+1}),$$

et, puisque $k < 1$

$$= |x_0 - l|\, k^n.$$

On a donc, d'après (S.13),

$$\text{Si } x \in [x_n, x_{n+1}] \text{ ou } x \in [x_{n+1}, x_n], \quad |g'(x)| \leq \frac{M}{(p-1)!}(k^n\, |x_0 - l|)^{p-1}. \tag{S.14}$$

Par ailleurs, (Q.7) est toujours valable et fournit, grâce à (S.14),

$$\begin{aligned} |x_{n+1} - x_n| &= |g'(\xi)|\, |x_n - x_{n-1}|\,, \\ &\leq \frac{M}{(p-1)!}(k^n\, |x_0 - l|)^{p-1}|\, |x_n - x_{n-1}|\,, \\ &\leq \frac{M}{(p-1)!}(k^n\, |\alpha - \beta|)^{p-1}|\, |x_n - x_{n-1}|\,. \end{aligned}$$

Bref, en considérant A et B définis par (S.9), on a

$$\forall n \in \mathbb{N}^*, \quad |x_{n+1} - x_n| \leq AB^n|\, |x_n - x_{n-1}|\,. \tag{S.15}$$

En raisonnant de façon proche de l'établissement par récurrence de (Q.3), on a

$$\begin{aligned} |x_{n+1} - x_n| &\leq AB^n|\, |x_n - x_{n-1}|\,, \\ &\leq A^2B^{n+(n-1)}|\, |x_{n-1} - x_{n-2}|\,, \\ &\vdots \\ &\leq A^nB^{n+(n-1)+(n-2)+...+1}\, |x_1 - x_0|\,, \\ &\leq A^nB^{\frac{n(n+1)}{2}}\, |x_1 - x_0|\,, \end{aligned}$$

soit

$$\forall n \in \mathbb{N}^*, \quad |x_{n+1} - x_n| \leq A^nB^{\frac{n(n+1)}{2}}|\, |x_n - x_{n-1}|\,. \tag{S.16}$$

On considère ensuite deux entiers n et p en supposant par exemple $p < n$. On écrit de nouveau l'inégalité triangulaire (Q.10) et il vient

$$\begin{aligned}
|x_p - x_n| &\leq \sum_{q=0}^{n-p-1} |x_{p+q} - x_{p+q+1}|, \\
&\leq \sum_{q=0}^{n-p-1} A^{p+q} B^{\frac{(p+q)(p+q+1)}{2}} |x_1 - x_0|, \\
&\leq A^p B^{\frac{p(p+1)}{2}} |x_1 - x_0| \sum_{q=0}^{n-p-1} A^q B^{\frac{(p+q)(p+q+1)}{2} - \frac{p(p+1)}{2}}, \\
&\leq A^p B^{\frac{p(p+1)}{2}} |x_1 - x_0| \sum_{q=0}^{n-p-1} A^q B^{\frac{q^2+2pq+q}{2}},
\end{aligned}$$

Ainsi,

$$\forall n, p \in \mathbb{N}^*, \quad n > p \Longrightarrow |x_p - x_n| \leq A^p B^{\frac{p(p+1)}{2}} |x_1 - x_0| \sum_{q=0}^{n-p-1} A^q B^{\frac{q^2+2pq+q}{2}}. \tag{S.17}$$

À p fixé, la série de terme général $A^q B^{\frac{q^2+2pq+q}{2}}$ est convergente. En effet, on a

$$A^q B^{\frac{q^2+2pq+q}{2}} \leq A^q B^{\frac{q^2}{2}}$$

et donc

$$q^2 A^q B^{\frac{(q^2+2pq+q)(p+1)}{2}} = \exp\left(\frac{q^2}{2} \ln B + q \ln A + 2 \ln q\right),$$

qui tend vers zéro quand q tend vers l'infini et donc

$$A^q B^{\frac{q^2+2pq+q}{2}} = o\left(\frac{1}{q^2}\right).$$

Dans (S.17), on peut faire tendre n vers l'infini, ce qui donne

$$\forall p \in \mathbb{N}^*, \quad |x_p - l| \leq A^p B^{\frac{p(p+1)}{2}} |x_1 - x_0| \sum_{q=0}^{\infty} A^q B^{\frac{q^2+2pq+q}{2}}.$$

On considère F défini par (S.10). On a donc démontré (S.12). Pour conclure, montrons (S.11). Il suffit d'écrire

$$A^p B^{\frac{p(p+1)}{2}} \sum_{q=0}^{\infty} A^q B^{\frac{q^2+2pq+q}{2}} \leq A^p B^{\frac{p(p+1)}{2}} \sum_{q=0}^{\infty} A^q B^{\frac{q^2}{2}} = 0\left(A^p B^{\frac{p(p+1)}{2}}\right),$$

en remarquant comme précédemment que la série de terme général $A^q B^{\frac{q^2}{2}}$ est convergent. □

L'intérêt de ce théorème est qu'elle est valable pour tout A, le nombre B étant toujours plus petit que 1. Sa difficulté réside dans le calcul de la somme (S.10), qui peut s'approcher par troncature et contrôle du reste.

Exemple numérique issu des TD ?

Calcul explicite de la somme (S.10) *?*

Annexe T

Compléments sur la divergence de la méthode du point fixe (sous forme d'un exercice corrigé)

Afin d'illuster les propositions 4.11 et 4.12, nous donnons dans cette annexe, un exercice donné en examen en Informatique 3A, à l'automne 2018, puis nous concluons par quelques remarques.

Énoncé

Soit la fonction g définie par

$$\forall x \in I = [0,1], \quad g(x) = \begin{cases} 2x, & \text{si } x \in [0,1/2], \\ 2(1-x), & \text{si } x \in [1/2,1]. \end{cases} \tag{T.1}$$

(1) (a) Montrer que les points fixes de g sont 0 et $2/3$.

(b) Montrer que I est g-stable (i.e., si $x \in I$ alors $g(x) \in I$).

(c) Que peut-on en déduire sur la suite de la méthode du point fixe définie par $x_{n+1} = g(x_n)$?

(d) Quelles sont les seules limites possibles de cette suite ?

(2) (a) Est-ce que l'on peut affirmer que l'hypothèse (4.39b) de la proposition 4.19 page 87 est vérifiée ? est vérifiée ?

(b) Peut-on affirmer que la méthode du point fixe converge ?

(3) (a) Est-ce que l'on peut affirmer que l'hypothèse (4.32c) de la proposition 4.12 est vérifiée ?

(b) Peut-on affirmer que la méthode du point fixe ne converge pas ?

(4) (a) Calculer les 6 premières valeurs de la suite x_n du point fixe pour $x_0 = 1/24$. On s'efforcera de calculer les différentes valeurs de x_n sous forme de fraction. On pourra faire un petit graphique.
Est-ce que dans ce cas, la suite du point fixe converge ?

(b) Calculer les 5 premières valeurs de la suite x_n du point fixe pour $x_0 = 1/7$. On s'efforcera de calculer les différentes valeurs de x_n sous forme de fraction. On pourra faire un petit graphique.
Est-ce que dans ce cas, la suite du point fixe converge ?

(c) (i) Calculer les 9 premières valeurs de la suite x_n du point fixe pour $x_0 = 1/4\,\pi$. Avec du courage, on pourra déterminer x_n sous la forme $a_n + \pi b_n$ où a_n et b_n sont des entiers (pour $n \geq 2$).

(ii) Que remarquez-vous ?

Corrigé

Pour plus de détails, on pourra consulter [BM03, Exercice 4.9 et TP 4.N] ainsi que `https://www.univers-ti-nspire.fr/files/pdf/14-th_point_fixe-TNS21.pdf`.

Soit la fonction g définie par

$$\forall x \in I = [0,1], \quad g(x) = \begin{cases} 2x, & \text{si } x \in [0,1/2], \\ 2(1-x), & \text{si } x \in [1/2,1]. \end{cases} \tag{T.2}$$

(1) (a) Si on résoud $g(x) = x$, on a deux cas :
— Si $x \in [0, 1/2]$, cela est équivalent à $2x = x$ et donc $x = 0$ qui est bien dans $[0, 1/2]$.
— Si, au contraire, on a $x \in [1/2, 1]$, cela est équivalent à $2(1 - x) = x$ et donc $3x = 2$, soit encore $x = 2/3$ qui est bien dans $[0, 1/2]$.
Ainsi,

$$\text{les points fixes de } g \text{ sont } 0 \text{ et } 2/3. \tag{T.3}$$

(b) Soit $x \in [0, 1]$. On a deux cas :
— Si $x \in [0, 1/2]$, alors $g(x) = 2x \in [0, 2 \times 1/2] = [0, 1]$.
— Si $x \in [1/2, 1]$, alors $g(x) = 2(1 - x) \in [0, 2 \times 1/2] = [0, 1]$.
Ainsi

$$I \text{ est } g\text{-stable.} \tag{T.4}$$

(c) De cela, on peut déduire par récurrence sur n, que si $x_0 \in I$, alors, pour tout n, x_n est dans I, et c'est tout !

(d) Puisque g est continue, les seules limites possibles de x_n sont les points fixes, c'est-à-dire 0 et 2/3.

$$\text{Les seules limites possibles de } x_n \text{ sont } 0 \text{ et } 2/3. \tag{T.5}$$

(2) (a) g est dérivable sauf en 1/2. Il est clair que

$$\forall x \in I \setminus \{1/2\}, \quad |g'(x)| = 2, \tag{T.6}$$

et donc l'hypothèse (4.39b) de la proposition 4.19 page 87 n'est pas vérifiée.

(b) Ainsi, les hypothèses de la proposition 4.19 ne sont pas assurées. Celle-ci assurait que la convergence du point fixe avait lieu ; si les hypothèses ne sont pas vérifiées, on ne peut donc affirmer que la méthode converge (ou diverge d'ailleurs !).

(3) (a) D'après (T.6), l'hypothèse (4.32c) de la proposition 4.12 page 85 n'est pas vérifiée.

(b) Comme précédemment, on ne peut pas affirmer que la méthode du point fixe ne converge pas (ou converge !).

REMARQUE T.1. On peut en fait réduire I à $I_1 = [0, \varepsilon]$ ou $I_2 = [2/3 - \varepsilon, 2/3 + \varepsilon]$ (avec $0 < \varepsilon < 1/3$), ce qui rend vraie l'hypothèse (4.32c) de la proposition 4.12 page 85. Malheureusement, dans ce cas, l'hypothèse (4.32b) de la proposition 4.12 page 85, n'est pas valable. En effet, si par exemple, x n'appartient pas à I_1, on ne peut affirmer que $g(x)$ (qui existe) n'appartient pas à I_1. En effet, x peut appartenir à $[1 - \varepsilon/2, 1]$. Dans ce cas, $g(x)$ appartient à $[2(1 - 1), 2(1 - 1 + \varepsilon/2)] = [0, \varepsilon] = I_1$. De même, si x n'appartient pas à I_2, il peut appartenir à $[1/3 - \varepsilon/2, 1/3 + \varepsilon/2]$ et dans ce cas $g(x)$ appartient à $[2(1/3 - \varepsilon/2), 2(1/3 + \varepsilon/2)] = [2/3 - \varepsilon, 2/3 + \varepsilon] = I_2$.

(4) (a)

On obtient les valeurs données dans les tableaux T.1 ou T.2. Voir aussi la figure T.1. On constate qu'une fois les 4 premières valeurs de x_n passées ($\{1/24, 1/12, 1/6, 1/3\}$), la valeur de x_n est constamment égale à 2/3, point fixe de g.

Dans ce cas-là, la suite converge donc vers 2/3.

(b)

On obtient les valeurs données dans les tableaux T.3 ou T.4. Voir aussi la figure T.2.

On constate qu'une fois la première valeur de x_n passée ($\{1/7\}$), les valeurs de x_n présentent un cycle : $\{2/7, 4/7, 6/7\}$, puisque $g(6/7) = 2/7$.

Dans ce cas-là, la suite ne converge pas, puisqu'elle "oscille" en permanence.

(c)

(i)

n	x_n
0	1/24
1	1/12
2	1/6
3	1/3
4	2/3
5	2/3

TABLE T.1. Valeurs de x_n pour $x_0 = 1/24$.

n	x_n
0	0.041666666666667
1	0.083333333333333
2	0.166666666666667
3	0.333333333333333
4	0.666666666666667
5	0.666666666666667

TABLE T.2. Valeurs (numériques) de x_n pour $x_0 = 1/24$.

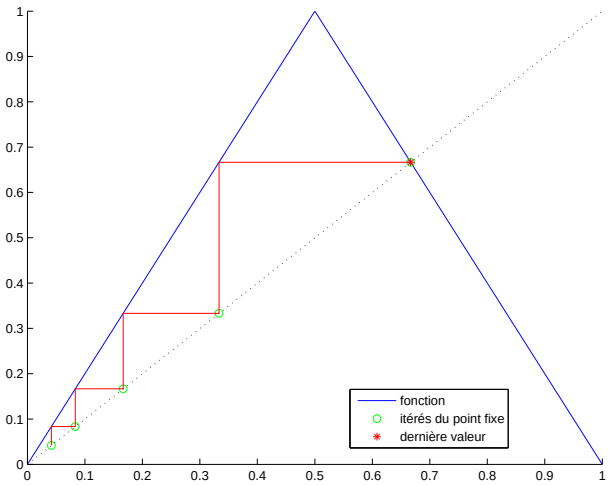


FIGURE T.1. méthode du point fixe pour $x_0 = 1/24$.

n	x_n
0	1/7
1	2/7
2	4/7
3	6/7
4	2/7

TABLE T.3. Valeurs de x_n pour $x_0 = 1/7$.

n	x_n
0	0.142857142857143
1	0.285714285714286
2	0.571428571428571
3	0.857142857142857
4	0.285714285714286

TABLE T.4. Valeurs (numériques) de x_n pour $x_0 = 1/7$.

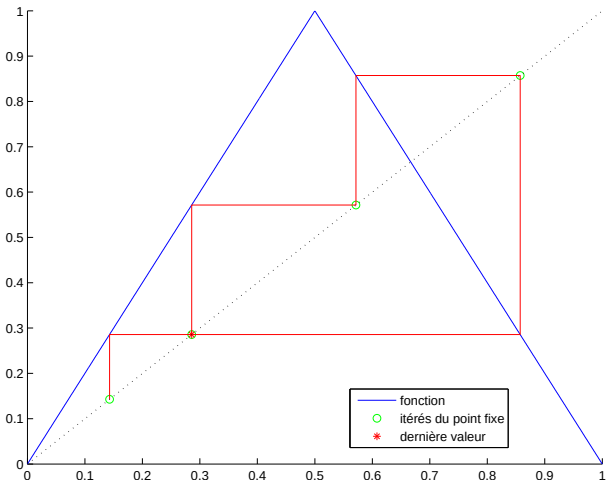


FIGURE T.2. méthode du point fixe pour $x_0 = 1/7$.

n	a_n	a_n
0	0	1/4
1	2	−1/2
2	4	−1
3	−6	2
4	−12	4
5	26	−8
6	−50	16
7	−100	32
8	202	−64
9	−402	128
10	−804	256
11	−1608	512
12	−3216	1024
13	6434	−2048
14	12868	−4096
15	25736	−8192
16	51472	−16384
17	102944	−32768
18	205888	−65536
19	−411774	131072
20	823550	−262144
21	1647100	−524288
22	−3294198	1048576
23	6588398	−2097152
24	−13176794	4194304
25	26353590	−8388608
26	−52707178	16777216
27	105414358	−33554432
28	−210828714	67108864
29	−421657428	134217728
30	−843314856	268435456
31	1686629714	−536870912
32	−3373259426	1073741824
33	−6746518852	2147483648
34	−13493037704	4294967296
35	26986075410	−8589934592
36	−53972150818	17179869184
37	−107944301636	34359738368
38	−215888603272	68719476736
39	−431777206544	137438953472
40	863554413090	−274877906944
41	−1727108826178	549755813888
42	3454217652358	−1099511627776
43	6908435304716	−2199023255552
44	−13816870609430	4398046511104
45	27633741218862	−8796093022208
46	−55267482437722	17592186044416
47	−110534964875444	35184372088832
48	−221069929750888	70368744177664
49	442139859501778	−140737488355328
50	884279719003556	−281474976710656
51	−1768559438007110	562949953421312
52	−3537118876014220	1125899906842624
53	−7074237752028440	2251799813685248
54	−14148475504056880	4503599627370496
55	28296951008113762	−9007199254740992
56	−56593902016227522	18014398509481984
57	−113187804032455044	36028797018963968
58	−226375608064910088	72057594037927936
59	452751216129820178	−144115188075855872
60	905502432259640356	−288230376151711744

TABLE T.5. Valeurs de $x_n = a_n + \pi b_n$ pour $x_0 = 1/4\,\pi$.

n	x_n
0	0.785398163397448
1	0.429203673205103
2	0.858407346410207
3	0.283185307179586
4	0.566370614359173
5	0.867258771281654
6	0.265482457436692
7	0.530964914873384
8	0.938070170253233
9	0.123859659493535
10	0.247719318987069
11	0.495438637974138
12	0.990877275948276
13	0.018245448103448
14	0.036490896206895
15	0.072981792413791
16	0.145963584827581
17	0.291927169655162
18	0.583854339310324
19	0.832291321379352
20	0.335417357241296
21	0.670834714482593
22	0.658330571034814
23	0.683338857930372
24	0.633322284139257
25	0.733355431721486
26	0.533289136557027
27	0.933421726885945
28	0.133156546228109
29	0.266313092456218
30	0.532626184912436
31	0.934747630175127
32	0.130504739649746
33	0.261009479299491
34	0.522018958598983
35	0.955962082802035
36	0.088075834395931
37	0.176151668791861
38	0.352303337583722
39	0.704606675167445
40	0.590786649665111
41	0.818426700669779
42	0.363146598660442
43	0.726293197320884
44	0.547413605358232
45	0.905172789283535
46	0.189654421432930
47	0.379308842865860
48	0.758617685731720
49	0.482764628536561
50	0.965529257073122
51	0.068941485853756
52	0.137882971707512
53	0.275765943415025
54	0.551531886830049
55	0.896936226339902
56	0.206127547320196
57	0.412255094640392
58	0.824510189280785
59	0.350979621438431
60	0.701959242876861

TABLE T.6. Valeurs (numériques) de x_n pour $x_0 = 1/7$.

On obtient les 60 premières valeurs données dans les tableaux T.5 ou T.6. Voir aussi la figure

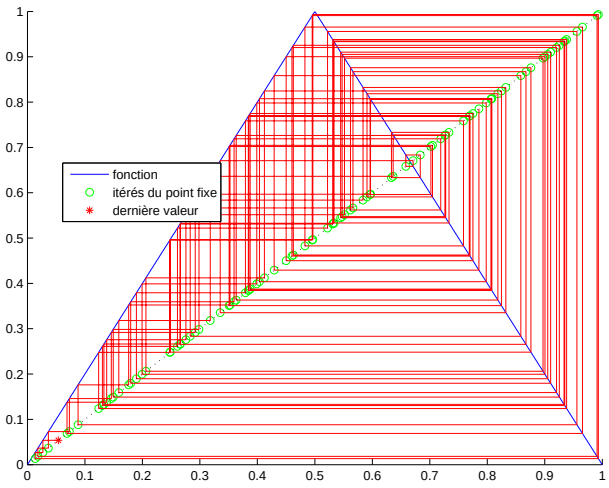


FIGURE T.3. méthode du point fixe pour $x_0 = 1/4\,\pi$.

T.3.

On peut montrer que, si $x_0 = 1/4\,\pi$, alors

$$\forall n \in \mathbb{N} \setminus \{0,1\}, \quad \exists!(a_n, b_n) \in \mathbb{Z}, \quad x_n = a_n + \pi b_n. \tag{T.7}$$

En effet, c'est vrai pour $n = 0$ avec $a_0 = 0$ et $b_0 = 1/4$ (qui n'est pas entier !). On démontre ensuite que (T.7) est vrai par récurrence sur n avec a_n et b_n dans $\mathbb{Q}$. Pour l'existence, on suppose que cela est vrai pour n. Pour $n+1$, on a, si $x_n \in [0, 1/2]$,

$$x_{n+1} = g(x_n) = 2x_n = 2a_n + \pi(2b_n),$$

et donc

$$\begin{aligned} a_{n+1} &= 2a_n, \\ b_{n+1} &= 2b_n. \end{aligned}$$

Au contraire, si $x_n \in [1/2, 1]$, on a

$$x_{n+1} = g(x_n) = 2(1 - x_n) = 2(1 - a_n) + \pi(-2b_n),$$

et donc

$$\begin{aligned} a_{n+1} &= 2(1 - a_n), \\ b_{n+1} &= -2b_n. \end{aligned}$$

On a donc les formules de récurrence :

$$\forall n \in \mathbb{N}, \quad \begin{cases} a_{n+1} = 2a_n, \quad b_{n+1} = 2b_n, & \text{si } a_n + \pi b_n \in [0, 1/2], \\ a_{n+1} = 2(1 - a_n), \quad b_{n+1} = -2b_n, & \text{si } a_n + \pi b_n \in [1/2, 1]. \end{cases} \tag{T.8}$$

Il est clair que, si a_n et b_n sont rationnels, il en est de même pour a_{n+1} et b_{n+1}. L'unicité vient du fait que, si

$$a_n + \pi b_n = a'_n + \pi b'_n,$$

alors

$$(a_n - a'_n) + \pi(b_n - b'_n) = 0,$$

et l'irrationnalité de π entraîne la nullité de $a_n - a'_n$ et de $b_n - b'_n$. Enfin, on constate dans le tableau T.5, que, pour $n = 2$, a_n et b_n sont entiers et (T.8) entraîne par récurrence sur n que cela est vrai pour tout $n \geq 2$.

(ii) On constate que les premières valeurs de x_n sont toutes deux à deux distinctes. On peut montrer qu'il en est de même pour toutes les valeurs de x_n.

REMARQUE T.2. Sur le tableau T.5, on constate que les valeurs de $|a_n|$ et de $|b_n|$ sont de plus en plus grandes. Plus précisément, montrons que

$$\lim_{n\to+\infty} |a_n| = +\infty, \tag{T.9a}$$

$$\lim_{n\to+\infty} |b_n| = +\infty. \tag{T.9b}$$

En fait, montrons tout d'abord par récurrence sur n que

$$\forall n \geq 2, \quad |b_n| = 2^{n-2}. \tag{T.10}$$

ce qui implique (T.9b). Dans le tableau T.5, on constate que (T.10) est vrai pour $n = 2$. Supposons maintenant (T.10) vraie pour $n \geq 2$. Montrons-la pour $n + 1$. D'après (T.8) et la récurrence, on a donc

$$|b_{n+1}| = 2|b_n| = 2 \times 2^{n-2} = 2^{n-1},$$

ce qui est bien (T.10) à l'ordre $n + 1$. Montrons maintenant que

$$\forall n \geq 2, \quad |a_n| \geq 2^{n-2}(|a_2| - 2) + 2. \tag{T.11}$$

ce qui implique, d'après la valeur de $|a_2|$ donnée dans le tableau T.5, (T.9a). Montrons d'abord

$$\forall n \geq 2, \quad |a_{n+1}| \geq 2|a_n| - 2. \tag{T.12}$$

On a si $a_n + \pi b_n \in [0, 1/2]$

$$|a_{n+1}| = 2|a_n| \geq 2|a_n| - 2.$$

Si $a_n + \pi b_n \in [1/2, 1]$, on a alors, d'après une inégalité triangulaire,

$$|a_{n+1}| = |2(1 - a_n)| = 2|1 - a_n| = 2|a_n - 1| \geq 2(|a_n| - 1) = 2|a_n| - 2.$$

Concluons enfin en montrant (T.11). On a pour $\alpha = 2$,

$$\alpha = 2\alpha - 2, \tag{T.13}$$

et donc en soustrayant (T.12) à (T.13), on a

$$\forall n \geq 2, \quad |a_{n+1}| - \alpha \geq 2|a_n| - 2\alpha,$$

soit

$$\forall n \geq 2, \quad |a_{n+1}| - \alpha \geq 2(|a_n| - \alpha),$$

et par une récurrence immédiate

$$\forall n \geq 2, \quad |a_n| - \alpha \geq 2^{n-2}(|a_2| - \alpha),$$

et, puisque $\alpha = 2$, (T.11) en découle.

Quelques remarques

(1) Dans cet exercice, on étudie une méthode de point fixe dont la convergence ne peut être *a priori* affirmée (la proposition 4.19 page 87 du cours ne s'applique pas.) et la divergence ne peut être *a priori* affirmée (la proposition 4.12 page 85 du cours ne s'applique pas.) En fait, pour tout x_0 appartenant à $I = [0, 1]$, la suite est soit convergente, soit divergente. Plus précisément, on montre que, pour tout $x_0 \in I$, seuls les trois cas exclusifs suivants peuvent se présenter :
   (a) La suite (x_n) possède une infinité de valeurs distinctes ;
   (b) il existe $p \geq 0$ tel que $x_0, x_1, ..., x_p$ sont deux à deux distincts, puis la suite est stationnaire, c'est-à-dire
   $$\forall q \geq p + 1, \quad x_q = x_p. \tag{T.14}$$
   (c) il existe $p \geq 1$ tel que $x_0, x_1, ..., x_p$ sont deux à deux distincts, puis la suite est cyclique mais non stationnaire, c'est-à-dire : il existe $r \in \{0, ..., p - 1\}$ tel que
   $$x_{p+1} = x_r, \quad x_{p+2} = x_{r+1}, \quad ... \quad x_{p+k} = x_{p-1} \text{ avec } k = p - 1 - r. \tag{T.15}$$
   Dans ce cas, on dit que la suite est cyclique d'ordre $k + 1$. Naturellement, les valeurs suivantes des itérés se déduisent par périodicité.

   Dans le cas 1a, on peut appliquer alors la proposition 4.11 page 84 du cours et il y a divergence. Dans le cas 1c, il y a aussi divergence. Au contraire, dans le cas 1b, il y a convergence. On peut même expliciter la partition U, V et W de I telle que le 1a a lieu si $x_0 \in U$, le 1b a lieu si $x_0 \in V$, le 1c a lieu si $x_0 \in W$. On a : $(\mathbb{R} \setminus \mathbb{Q}) \cup I \subset U$ et en particulier, $\pi/4$ appartient à U. Par exemple, $1/24$ appartiennt à V. Par exemple, $1/7$ ou $1/84$ appartiennent à W. Les ensembles V et W sont dénombrables. On peut donc affirmer que, pour presque tout $x_0 \in I$, la suite du point fixe est divergente. L'ensemble des x_0 où il y a convergence est cependant dense dans I.

(2) Sur le plan numérique, les arrondis de calculs rendent les choses plus complexes.

   La divergence a en théorie lieu par exemple avec $x_0 = \pi/4$. Si on fait les calculs en symboliques, on obtient bien toutes les valeurs de x_n différentes comme le montre la figure 4(a), ce qui n'est plus vrai pour la figure 4(b). Dans ce dernier cas, pour $n = 51$, on a $x_n = 0$, c'est-à-dire une convergence vers 0.

   Si on prend $P = 318310$ points équirépartis dans $[0, 1]$ définis par $x_i = ih$ où i est un entier et $h = \pi/1000000$, on observe dans tous les cas, une convergence vers zéro, alors que, dans ces cas, on devrait observer une divergence !

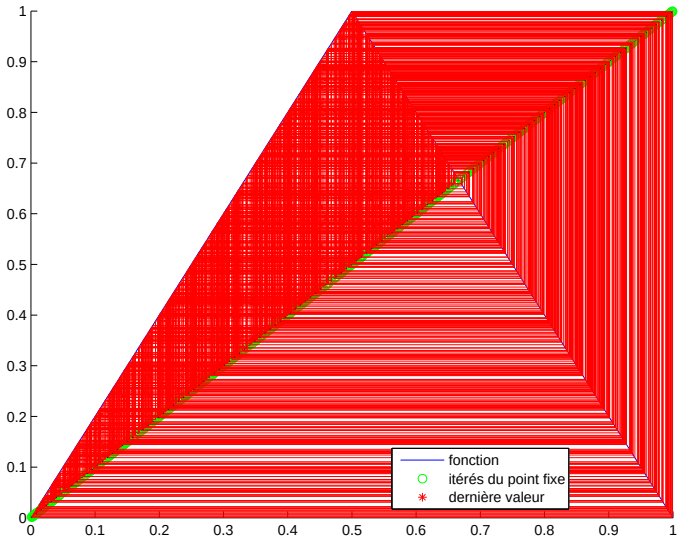
   Si on prend cette fois-ci $P = 1000000$ points aléatoirement répartis dans $[0, 1]$, on observe dans tous les cas, une convergence vers zéro !

(3) Pour éviter d'avoir ce problème d'arrondis ou d'avoir des temps de calculs trop long dûs aux symbolique, on peut utiliser une autre fonction g définie par
   $$\forall x \in [0, 1], \quad g(x) = -4x^2 + 4x. \tag{T.16}$$
   Comme précédemment, la fonction g définit une suite (x_n), qui en général ne converge pas. On peut aussi définir la partition U, V et W de I qui traduit les trois cas précédemment évoqués. Cependant la détermination analytique de U et V n'est pas possible ici. Les calculs se passent beaucoup mieux comme le montre la suite.

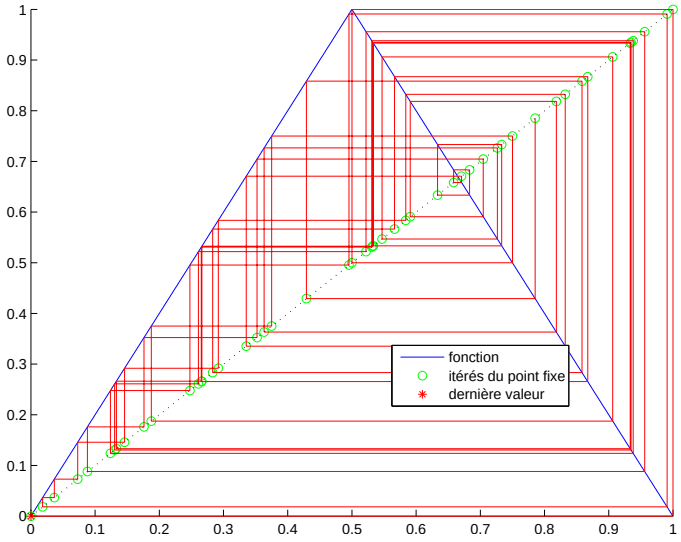
   La divergence a de nouveau en théorie lieu par exemple avec $x_0 = \pi/4$, ce qui est bien observé cette fois-ci sur la figure T.5 page 259.

   Si on prend $P = 319$ points équirépartis dans $[0, 1]$ définis par $x_i = ih$ où i est un entier et $h = \pi/1000$, on n'observe que dans 0.31 % des cas, une convergence vers zéro, alors que, dans ces cas, on devrait observer une divergence !

   Si on prend cette fois-ci $P = 1000$ points aléatoirement répartis dans $[0, 1]$, on observe dans tous les cas, une divergence (aucune valeur égale).



(a) en symbolique



(b) en numérique

FIGURE T.4. calculs pour $x_0 = \pi/4$

(4) Pour les deux fonctions g définies précédemment, on parle de comportement chaotique de la suite x_n. En effet, elle peut converger ou pas. En cas de divergence, les valeurs qu'elles prend, si elles sont en nombre infini, occupent "tout" $[0, 1]$ de façon imprévisible.

(a)

Plus précisément, on considère les deux valeurs initiales très proches $u_0 = \pi/4$ et $u_0 = \pi/4 + \varepsilon$ où

$$\varepsilon = 1.\,10^{-12},$$

et on trace les valeurs $|u_n(u_0) - u_n(u_0 + \varepsilon)|$ sur la figure T.6. Cet écart apparaît imprévisible et ne reste jamais très longtemps proche de zéro.

(b)

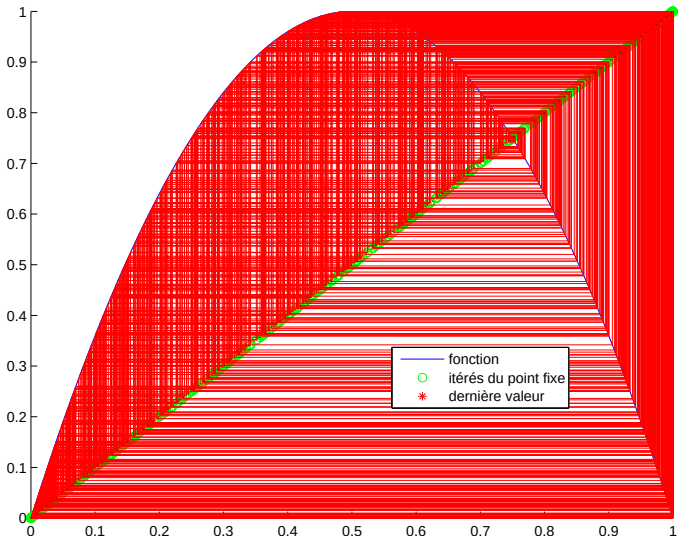


FIGURE T.5. calculs pour $x_0 = \pi/4$ et g définie par (T.16).

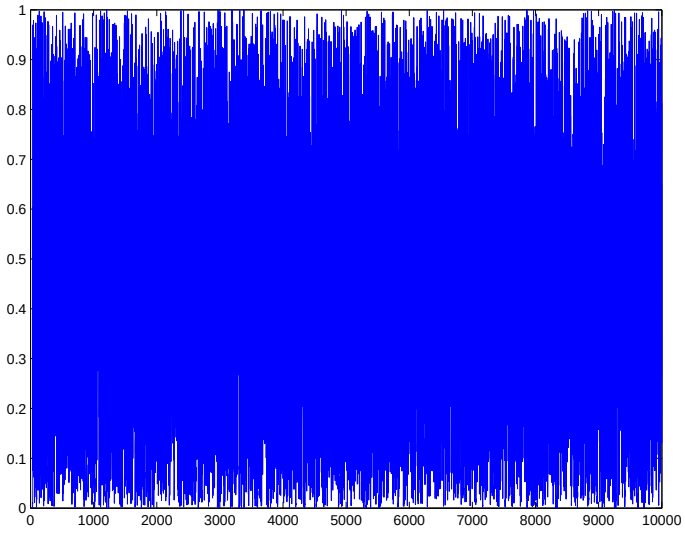


FIGURE T.6. Écarts entre x_n pour $x_0 = \pi/4$ et $x_0 = \pi/4 + \varepsilon$ et g définie par (T.16).

Faisons varier n dans $\{1000, 10000, 100000\}$. Par exemple, pour $n = 10000$, les valeurs (toutes différentes) prises par x_n sont représentées en figure T.7. Si on calcule les différentes valeurs de ρ, défini comme le rapport de l'écart maximal divisé par $h = 1/(n+1)$, on obtient

$$\text{pour } n = 1000, \quad \rho = 10.66,$$
$$\text{pour } n = 10000, \quad \rho = 11.32,$$
$$\text{pour } n = 100000, \quad \rho = 15.41.$$

Cela signifie que, si n augmente, les valeurs que prend x_n, si elles sont en nombre infini, occupent "de plus en plus" $[0, 1]$ et cela, de façon imprévisible, comme le montre par exemple la figure T.8.

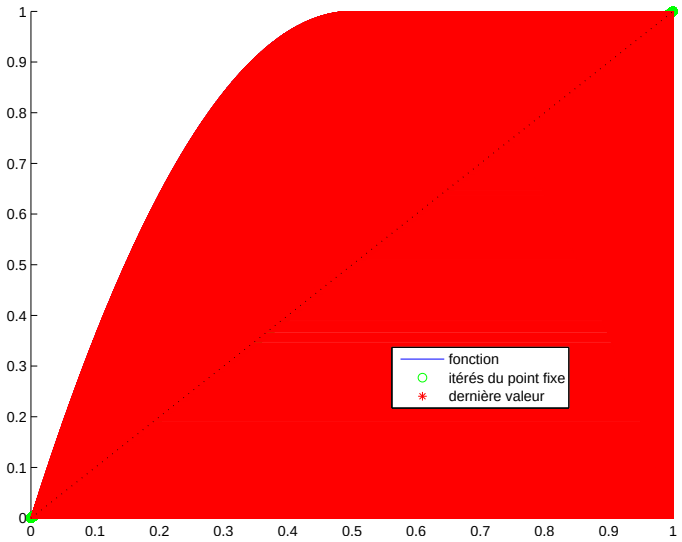


FIGURE T.7. 10000 premières valeurs de x_n pour $x_0 = \pi/4$ et g définie par (T.16).

FIGURE T.8. Zoom des valeurs entre 0.50000 et 0.50154 de x_n pour $x_0 = \pi/4$ et g définie par (T.16).

Annexe U

Dégénérescence de la méthode de Newton et méthode de Newton modifiée

Justifions tout d'abord la remarque 4.57 page 100.

Les résultats de cette annexe utilisent les notions de racines multiples de fonctions, qui généralisent celles de polynômes, le tout étant rappelé dans l'annexe X page 283.

PROPOSITION U.1. *Soit $m \in \mathbb{N}^*$. Si f est de classe $\mathcal{C}^{m+2}$ sur un intervalle du type $[r-\varepsilon, r+\varepsilon]$, si r est un zéro d'ordre*[1] *m alors,*

(1) Il existe h de classe $\mathcal{C}^2$ sur $[r-\varepsilon, r+\varepsilon]$ telle que

$$\forall x \in [r-\varepsilon, r+\varepsilon], \quad f(x) = (x-r)^m h(x), \tag{U.1a}$$

avec

$$h(r) \neq 0. \tag{U.1b}$$

(2) Quitte à supposer ε assez petit,

$$f' \text{ est non nulle sur } [r-\varepsilon, r+\varepsilon] \setminus \{r\}. \tag{U.2}$$

(3) (a) La méthode de Newton est exactement linéaire.

(b) Si, de plus, f est de classe $\mathcal{C}^{m+3}$ sur $[r-\varepsilon, r+\varepsilon]$, alors, la méthode de Newton modifiée définie par $x_{n+1} = g(x_n)$ où

$$\forall x \in [r-\varepsilon, r+\varepsilon], \quad g(x) = \begin{cases} x - m\dfrac{f(x)}{f'(x)}, & \text{si } x \neq r, \\ r, & \text{si } x = r. \end{cases} \tag{U.3}$$

est au moins quadratique. Si

$$h'(r) \neq 0, \tag{U.4}$$

elle est exactement quadratique

On pourra aussi consulter
`https://math-linux.com/mathematiques/resolution-numerique-des-equations-non-lineaires/article/methode-de-newton`

DÉMONSTRATION DE LA PROPOSITION U.1.

(1) Les résultats du point 1 proviennent de la section X.2 de l'annexe X, notamment le lemme X.3 avec $p = m+2$.

(2) On déduit de (U.1a)

$$\forall x \in [r-\varepsilon, r+\varepsilon], \quad f'(x) = m(x-r)^{m-1}h(x) + (x-r)^m h'(x). \tag{U.5}$$

De (U.1b), on déduit donc (U.2).

1. Voir annexe X et particulièrement la section X.2 et la définition X.6.

(3) Considérons la fonction d'itération ϕ_A définie par, pour tout $A \in \mathbb{R}^*$:

$$\forall x \in [r-\varepsilon, r+\varepsilon] \setminus \{r\}, \quad \phi_A(x) = x - A\frac{f(x)}{f'(x)} \tag{U.6}$$

On a alors, grâce à (U.1a)

$$\forall x \in [r-\varepsilon, r+\varepsilon] \setminus \{r\}, \quad \phi_A(x) = x - A\frac{(x-r)^m h(x)}{m(x-r)^{m-1}h(x) + (x-r)^m h'(x)}, \tag{U.7}$$

et donc

$$\forall x \in [r-\varepsilon, r+\varepsilon] \setminus \{r\}, \quad \phi_A(x) = x - A\frac{(x-r)h(x)}{mh(x) + (x-r)h'(x)}, \tag{U.8}$$

et donc, par prolongement par continuité

$$\forall x \in [r-\varepsilon, r+\varepsilon], \quad \phi_A(x) = x - A\frac{h(x)(x-r)}{mh(x) + (x-r)h'(x)}. \tag{U.9}$$

D'après (U.1b), ϕ_A est de classe $\mathcal{C}^1$ puisque son dénominateur est non nul en r. Ainsi,

$$\phi'_A(x) = 1 - A\frac{\left(h(x) + (x-r)h'(x)\right)\left(mh(x) + (x-r)h'(x)\right) - \left(h(x)(x-r)\right)\left(mh'(x) + h'(x) + (x-r)h''(x)\right)}{\left(mh(x) + (x-r)h'(x)\right)^2} \tag{U.10}$$

D'après (U.9)

$$\phi_A(x) = r \tag{U.11}$$

et d'après (U.10)

$$\phi'_A(x) = r - \frac{Amh^2(r)}{m^2h^2(r)}$$

et donc

$$\phi'_A(x) = 1 - \frac{A}{m}. \tag{U.12}$$

On peut montrer sous matlab que

$$\phi_A(r) = r - A \times 0 = r, \tag{U.13a}$$

$$\phi'_A(r) = 1 - Am^{-1}. \tag{U.13b}$$

On retrouve donc les résultats (U.11) et (U.12). De plus, si f est de classe $\mathcal{C}^{m+3}$, alors h est de classe $\mathcal{C}^3$ et ϕ_A est de classe $\mathcal{C}^2$. On peut montrer grâce à matlab que

$$\forall x \in [r-\varepsilon, r+\varepsilon] \setminus \{r\}, \quad \phi''_A(x) = A\frac{N(x)}{D(x)},$$

avec

$$D(x) = \left(mh\left(x\right) + \left(\frac{d}{dx}h\left(x\right)\right)x - \left(\frac{d}{dx}h\left(x\right)\right)r\right)^3,$$

et $N(x)$ une expression, complexe et non affichée, dépendant de r, x, m et des trois premières dérivées de h et que

$$\phi''_A(r) = \frac{2Ah'(r)}{m^2h(r)}. \tag{U.14}$$

Remarque U.2. Informatiquement, on ne peut poser $\phi_A(r) = r$, car r est, en principe, inconnu ! De, l'expression théorique donnée par (U.8) n'est pas accessible. Seule, l'expression donnée par (4.91) est accessible. L'usage de cette expression peut aussi, compte tenu des arrondis de calculs, donner des résultats moins précis que ce que prévoit la théorie. D'après (U.2), cette expression est légitime. Cependant, si $x_{n+1} = x_n$ et si on se trouve dans le cas où $x_{n+1} = x_n$, alors on a

$$A\frac{f(x_n)}{f'(x_n)} = 0$$

et donc $f(x_n) = 0$ et on s'arrête là !

On peut conclure et démontrer maintenant les résultats des cas 3a et 3b.

(a) La méthode de Newton classique correspond à $A = 1$ et $\phi_A = g$, définie par (4.91). g est de classe $\mathcal{C}^1$. On a, d'après (U.12)
$$|g'(r)| = 1 - \frac{1}{m},$$
et on a $m > 1$ donc $1/m < 1$ et $1 - 1/m > 0$. On a aussi $1 - 1/m < 1$. Ainsi
$$|g'(r)| \in]0, 1[.$$
D'après la proposition 4.34, la méthode de Newton est donc exactement linéaire.

(b) La méthode de Newton modifiée correspond à $A = m$ et $\phi_A = g$, définie par (U.3). Si f est de classe $\mathcal{C}^{m+3}$, g est de classe $\mathcal{C}^3$. On alors, grâce à (U.12), $g'(r) = 1 - m/m = 0$ et d'après la proposition 4.38, cette méthode de Newton modifiée est au moins quadratique. De plus, dans ce cas, grâce à (U.14), on a
$$g''(r) = \frac{2h'(r)}{mh(r)}, \tag{U.15}$$
ce qui permet de conclure, grâce à (U.4).

□

On pourra consulter les deux exemples ci-dessous.

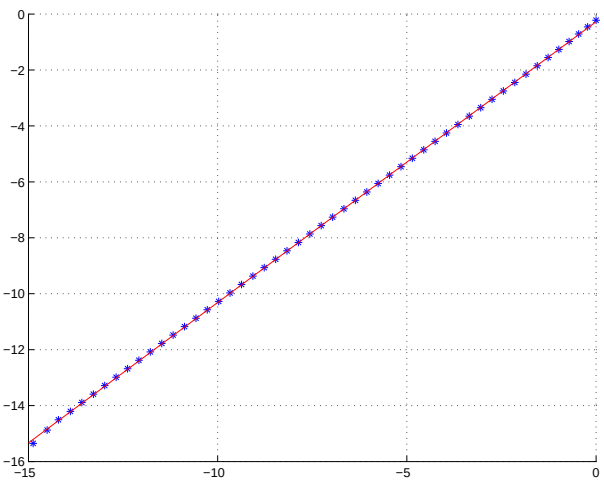
Exemple U.3. Considérons la fonction f définie par
$$f(x) = (x - \alpha)^p (x - \beta), \tag{U.16}$$
où β et α sont deux réels distincts et $p \in \mathbb{N}^*$. Si $p \geq 2$, il est clair que α est un zéro de multiplicité $p \geq 2$ de f.

On choisit les valeurs suivantes :
$$\alpha = 2,$$
$$\beta = 1,$$
$$p = 2.$$

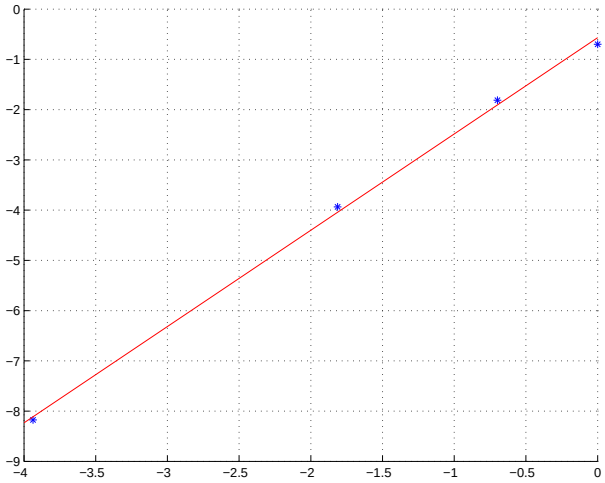
Pour g définie par (U.3), on obtient sous matlb
$$\lim_{x \to \alpha} g(x) = 2,$$
$$\lim_{x \to \alpha} g'(x) = 0,$$
$$\lim_{x \to \alpha} g''(x) = 1,$$
ce qui confirme que l'ordre de la méthode de Newton modifiée est exactement 2.

On étudie f au voisinage de α. On obtient les itérés de la méthode de Newton dans le tableau U.1 et ceux de la méthode modifiée dans le tableau U.2, qui semble mieux se comporter. Afin de comparer le comportement des deux méthodes, on procède exactement comme dans le corrigé de l'exercice de TD 4.4. On obtient alors les deux nuages $\ln - \ln$ des points représentés sur les figures U.1. Enfin, les deux mesures de pentes de ces nuages fournissent les corrélations 0.99998221 et 0.99937031 ainsi que les deux pentes 1.00377621 et 1.91625254 ce qui confirme l'ordre 1 de la méthode de Newton et l'ordre 2 de la méthode modifiée.

Exemple U.4. On pourra consulter l'exercice 4.4 des TD où la méthode de newton modifiée est utilisée et comparée à la méthode de Newton habituelle dans la remarque 4.8 du corrigé.



(a) Méthode de Newton habituelle



(b) Méthode de Newton modifiée

FIGURE U.1. Le nuage de points ln − ln.

n	x_n
0	3.000000000000000
1	2.600000000000000
2	2.347368421052632
3	2.193516663631397
4	2.104014284855781
5	2.054346842021410
6	2.027856158851746
7	2.014114290149136
8	2.007105915824401
9	2.003565448288779
10	2.001785885343129
11	2.000893737887931
12	2.000447068368469
13	2.000223584118280
14	2.000111804552414
15	2.000055905400748
16	2.000027953481662
17	2.000013976936172
18	2.000006988516924
19	2.000003494270672
20	2.000001747138388
21	2.000000873569957
22	2.000000436785169
23	2.000000218392632
24	2.000000109196328
25	2.000000054598167
26	2.000000027299084
27	2.000000013649542
28	2.000000006824771
29	2.000000003412386
30	2.000000001706193
31	2.000000000853097
32	2.000000000426549
44	2.000000000000104
45	2.000000000000052
46	2.000000000000026
47	2.000000000000013
48	2.000000000000007
49	2.000000000000004
50	2.000000000000002
51	2.000000000000001
52	2.000000000000000
53	2.000000000000000

TABLE U.1. Itérations de la méthode de Newton

n	x_n
0	3.000000000000000
1	2.200000000000000
2	2.015384615384615
3	2.000115673799884
4	2.000000006689053
5	2.000000000000000

TABLE U.2. Itérations de la méthode de Newton modifiée

Annexe V

Étude et calcul de l tel que $l = \cos l$ sous la forme d'un problème corrigé

Ce problème a été donné à l'examen d'Automne 2020.

Énoncé

(1) On considère la suite $(u_n)_{n\in\mathbb{N}}$ définie par $u_0 \in \mathbb{R}$ et, pour tout $n \geq 0$, $u_{n+1} = \cos(u_n)$.

(a) Montrer que l'intervalle $I = [\cos(1), 1]$ est stable par la fonction g donnée par $g(x) = \cos(x)$. Par la suite, on pourra donc supposer (quitte à poser $u_0 = u_2$) que u_2 appartient à I.

(b) Étudier la convergence de la suite $(u_n)_{n\in\mathbb{N}}$.

(c) (i) On pose

$$\varepsilon_1 = 10^{-1} \text{ et } \varepsilon_2 = 10^{-15}. \tag{V.1}$$

Notons α la limite de la suite. Déterminer deux entiers n_1 et n_2 tels que

$$|u_{n_1} - \alpha| \leq \varepsilon_1 \text{ et } |u_{n_2} - \alpha| \leq \varepsilon_2. \tag{V.2}$$

(ii) Pour $u_0 = 10$, calculer numériquement les $n_1 + 1$ premières valeurs de u_n. Déterminer l'écart entre u_{n_1} et α, solution de $x = \cos(x)$ déterminé de façon exacte donné par

$$\alpha = 0.7390851332151606 \tag{V.3}$$

et conclure.

(2) On veut maintenant déterminer de façon numérique la valeur de α donnée par (V.3) mais plus rapidement que dans la question 1.

(a) Quelle méthode itérative [1] pourriez-vous utiliser pour répondre à cette question ?

(b) On définit la fonction h par

$$\forall x \in \mathbb{R}, \quad h(x) = x - \cos(x). \tag{V.4}$$

On *admet* que h admet un unique zéro sur $\mathbb{R}$. On note (w_n) la suite associée à cette méthode. Précisez la fonction G telle que

$$\forall x \in \mathbb{R}, \quad w_{n+1} = G(w_n). \tag{V.5}$$

(c) (i) Pourquoi la méthode étudiée est-elle quadratique au voisinage de l'unique zéro de h sur l'intervalle I, défini par

$$I = [0, \pi/2] \,? \tag{V.6}$$

1. Il n'y a pas d'ambiguïté sur cette question, car on n'a vu qu'une seule méthode répondant à cela !

(ii) Démontrer que l'inégalité

$$\forall n \in \mathbb{N}, \quad |w_{n+1} - l| \leq D\,|w_n - l|^2 \tag{V.7}$$

est satisfaite pour

$$D = \frac{1}{2} \max_{x \in I} |G''(x)|\,, \tag{V.8}$$

en admettant que la méthode étudiée converge pour tout x_0 appartenant à l'intervalle I vers l'unique zéro de h.

(iii) En fait, l'intervalle I est un peu trop grand. On se place désormais sur l'intervalle J défini par

$$J = [\nu, \pi/2 - \nu], \text{ où } \nu = 0.3000, \tag{V.9}$$

sur lequel tout ce qui a été fait précédemment est encore valable. On fournit alors la majoration suivante de la constante D, valable sur l'intervalle J :

$$D \leq 0.73741536. \tag{V.10}$$

On considère les nombres ε_1 et ε_2 définis par (V.1). Déterminer deux entiers m_1 et m_2 tels que

$$|u_{m_1} - \alpha| \leq \varepsilon_1 \text{ et } |u_{m_2} - \alpha| \leq \varepsilon_2, \tag{V.11}$$

et comparer avec les résultats de la question 1(c)i.

Commentez !

(iv) Déterminez les $m_1 + 1$ premières valeurs de la suite w_n pour $w_0 = 0.3000$ et concluez grâce à la valeur de α donnée par (V.3).

Corrigé

(1) (a) Définissons la fonction g par

$$\forall x \in \mathbb{R}, \quad g(x) = \cos(x). \tag{V.12}$$

Posons

$$I = [\cos(1), 1] \tag{V.13}$$

qui est g-stable. Voir la définition 4.18 du cours. En effet, $\cos(1) \approx 0.5403023$ appartient à $[0, \pi/2]$ et 1 appartient à $[0, \pi/2]$ et puisque cos est décroissant sur $[0, \pi/2]$, pour tout $x \in [\cos(1), 1]$, on a $\cos(x) \in [\cos(1), \cos(\cos(1)] \approx [\cos(1), 0.8575532] \subset [\cos(1), 1]$.

Remarque V.1. On peut aussi montrer cela sans calculer les valeurs approchées. En effet, $1 \in]0, \pi/2[$ et donc $\cos(1) \in]0, 1[$. Ainsi, $[\cos(1), 1] \subset [0, 1] \subset [0, \pi/2]$ et par décroissance de cos sur $[0, \pi/2]$, on a $g([\cos(1), 1]) = [\cos(1), \cos(\cos(1))] \subset [\cos(1), 1]$ puisque $\cos(\cos(1)) \leq 1$ et $\cos(\cos(1)) \neq 1$, sinon $\cos(1) = 0$.

Si $u_0 \in \mathbb{R}$, on a $u_1 = \cos(u_0) \in [-1, 1]$ et u_2 appartient donc à $\cos([-1, 1] = \cos([0, 1]) = [\cos(1), \cos(0)] = I$. Donc,

$$n \geq 2, \quad u_n \in I, \tag{V.14}$$

ce que l'on montre par récurrence, puisque I est g-stable. Si la suite u_n converge, la fonction g étant continue, c'est nécessairement vers un point fixe de g. Définissons la fonction h par

$$\forall x \in \mathbb{R}, \quad h(x) = x - \cos(x). \tag{V.15}$$

Puisque $h(\cos(1))h(1) \approx -0.1458395 < 0$, la fonction h étant continue, cela implique que h admet au moins une racine sur $[\cos(1), 1]$ et donc g admet au moins un point fixe, noté α et on vérifie que

$$\alpha \in I \tag{V.16}$$

On montrera plus bas que α est unique.

Remarque V.2. Un autre choix consiste à choisir

$$I = [1/2, 1]. \tag{V.17}$$

Si $u_0 \in \mathbb{R}$, on a $u_1 = \cos(u_0) \in [-1, 1]$ et u_2 appartient donc à $\cos([-1, 1] = \cos([0, 1]) = [\cos(1), \cos(0)]$, qui est inclus dans I. En effet, on a $0 < 1 < \pi/3 < \pi/2$ et donc $\cos(1) > \cos(\pi/3) = 1/2$. De plus, I est g-stable. En effet, si x est dans $[1/2, 1]$, on a $\cos(x)$ dans $[\cos(1), \cos(1/2)]$ inclus dans $[\cos(\pi/3), \cos(0)] = I$. Ainsi, (V.14) reste vraie. Enfin, (V.16) reste vraie puisque $h(1/2)h(1) \approx -0.1735738 < 0$.

Remarque V.3. Un autre choix consiste à choisir

$$I = [0, 1]. \tag{V.18}$$

Si $u_0 \in \mathbb{R}$, on a $u_1 = \cos(u_0) \in [-1, 1]$. Si $u_1 \in [-1, 0[$, on peut remplacer u_1 par $-u_1 \in [0, 1]$ puisque $u_2 = \cos(u_1) = \cos(-u_1)$. On peut donc supposer que $u_1 \in [0, 1]$. On a donc I est g-stable. En effet, on a $0 < 1 < \pi/2$ et donc $\cos([0, 1]) = [\cos(1), \cos(0)]) = [\cos(1), 1] \subset [0, 1]$. Ainsi, (V.14) reste vraie.

Par la suite, on pourra donc supposer (quitte à appliquer la règle (V.22)) que u_2 (ou que u_1 dans le cas de la remarque V.3) appartient à I.

(b) On peut donc maintenant conclure sur la convergence de la suite $(u_n)_{n\in\mathbb{N}}$.
On a, pour tout $x \in I$, $|g'(x)| = |\sin(x)|$. La fonction sin est croissante sur $I = [\cos(1), 1]$ et est comprise, sur cet intervalle, entre $\sin(\cos(1)) \approx 0.5143953$ et $\sin(1) \approx 0.8414710$. L'inégalité de l'hypothèse (4.39b) du cours est donc vraie avec

$$k = \sin(1). \tag{V.19}$$

Puisque $0 < 1 < \pi/2$, on a $\sin(1) \in]0, 1[$. On laisse au lecteur le soin de vérifier que ce raisonnement est encore valable si l'on choisit I donné dans la remarque V.2 ou V.3, puisque

$$0 \leq \sin(0) \leq \sin(1/2) \leq \sin(1).$$

Il suffit donc d'appliquer la proposition 4.19 du cours qui implique à la fois la convergence de la suite et l'existence et l'unicité du point fixe α de g.

Remarque V.4. Le choix de I donné par (V.17) ou (V.18) serait aussi valable.

Remarque V.5. On aurait pu se passer de l'utilisation de la proposition 4.19 du cours et montrer la convergence de la suite à la main, en raisonnant comme suit :

(i)
Étudions donc tout d'abord la fonction g, définie par (V.12). Si on définit la fonction h par (V.15), alors, on a

$$\forall x \in \mathbb{R}, \quad h'(x) = 1 + \sin(x), \tag{V.20}$$

qui est positive sur $\mathbb{R}$ et nulle pour $x = \pi/2 + 2k\pi$ où k entier. De plus $h(\pm\infty) = \pm\infty$. Ainsi h est strictement croissante sur $\mathbb{R}$ et n'admet qu'un seul zéro, noté α et qui est donc l'unique point fixe de g.

(ii)
Voir par exemple les figures V.1 et V.2
On a, pour tout $x \in I$,

$$x > \alpha \Longrightarrow g(x) < \alpha, \tag{V.21a}$$

$$x < \alpha \Longrightarrow g(x) > \alpha. \tag{V.21b}$$

En effet, si $x \in]\alpha, 1[\subset [0, \pi/2]$, puisque $g = \cos$ est décroissante sur $[0, \pi/2]$, on a $g(x) \in]\cos(1), \cos\alpha]$ et donc $g(x) < \alpha$. Il en est de même si $x < \alpha$. On peut supposer sans perte de généralité que $u_0 \in I$, quitte à appliquer la régle suivante :

$$\text{on remplace } u_0 \text{ par } u_2. \tag{V.22}$$

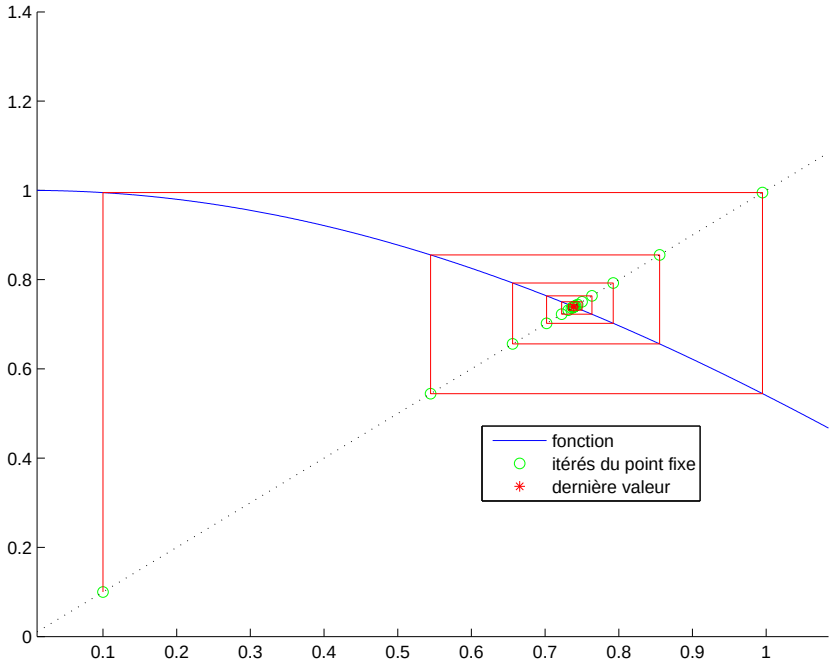


FIGURE V.1. Le tracé graphique des valeurs de u_n pour $u_0 = 0.100$.

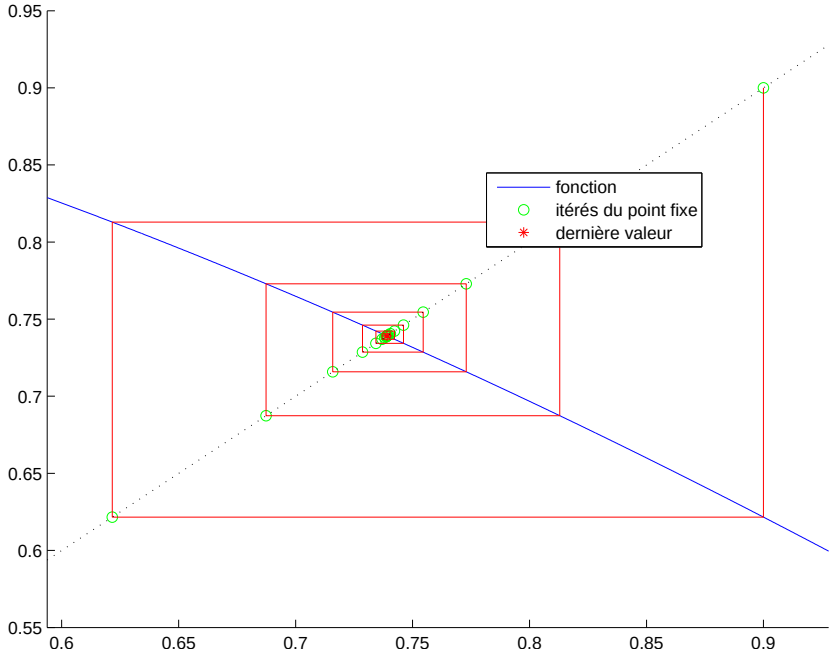


FIGURE V.2. Le tracé graphique des valeurs de u_n pour $u_0 = 0.900$.

Ainsi, si $u_0 = \alpha$, alors la suite u_n est constante et égale à α. Sinon, si $u_0 < \alpha$, alors d'après (V.21), $u_1 = g(u_0) > \alpha$ et $u_2 < \alpha$ et par récurrence

$$\forall n \in \mathbb{N}, \quad u_{2n} \in [\cos(1), \alpha[, \quad u_{2n+1} \in]\alpha, 1]. \tag{V.23}$$

De même, on montrer que, si $u_0 > \alpha$,

$$\forall n \in \mathbb{N}, \quad u_{2n+1} \in [\cos(1), \alpha[, \quad u_{2n} \in]\alpha, 1]. \tag{V.24}$$

Enfin, si on pose

$$f(x) = g(g(x)) - x, \tag{V.25}$$

on a

$$f'(x) = g'(g(x))g'(x) - 1$$

et donc

$$f'(x) = \sin(\cos(x))\sin(x) - 1. \tag{V.26}$$

Pour $x \in I$, $\cos(x) \in I$, $\sin(x) \in [\sin(\cos(1)), \sin(1)]$ et $\sin(\cos(x)) \in [\sin(\cos(1)), \sin(1)] \approx [0.5143953, 0.8414710]$. On a donc $\sin(\cos(x))\sin(x) < \sin^2(1)$ et $f'(x) \leq \sin^2(1) - 1 \approx -0.2919266 < 0$ et donc f est strictement décroissante sur I. Or $f(\alpha) = g(g\alpha)) - \alpha = 0$, donc pour tout $x \in I$

$$x > \alpha \Longrightarrow g(g(x)) < x, \tag{V.27a}$$

$$x < \alpha \Longrightarrow g(g(x)) > x. \tag{V.27b}$$

Remarque V.6. Comme précédemment, on peut montrer cela sans les valeurs numériques. En effet, si $x \in [\cos(1), 1]$, on a $\cos(x) \in [\cos(1), 1] \subset]0, 1[$ et donc $\sin(\cos(x)) \in]0, 1[$. Il en est de même pour $\sin(x)$. Ainsi, en particulier $0 < \sin(\cos(x))\sin(x) < 1$ et $f'(x) < 0$.

Remarque V.7. Ce raisonnement est encore valable pour le choix (V.17) ou (V.18).

Soit $n \in \mathbb{N}$. Si $u_{2n} < \alpha$, alors, d'après (V.27), on a $u_{2n+2} > u_{2n}$. En combinant cela, (V.23) et (V.24), on montre donc par récurrence que si $u_0 < \alpha$,

$$\forall n \in \mathbb{N}, \quad u_{2n} \in [\cos(1), \alpha[\text{ et la suite } u_{2n} \text{ est croissante}\,; \tag{V.28a}$$

$$\forall n \in \mathbb{N}, \quad u_{2n+1} \in]\alpha, 1] \text{ et la suite } u_{2n+1} \text{ est décroissante}\,; \tag{V.28b}$$

De même, si $u_0 > \alpha$,

$$\forall n \in \mathbb{N}, \quad u_{2n+1} \in [\cos(1), \alpha[\text{ et la suite } u_{2n+1} \text{ est croissante}\,; \tag{V.29a}$$

$$\forall n \in \mathbb{N}, \quad u_{2n} \in]\alpha, 1] \text{ et la suite } u_{2n} \text{ est décroissante}\,; \tag{V.29b}$$

Dans tous les cas, les deux suites u_{2n} et u_{2n+1} sont monotones, bornées (car dans I) et donc convergentes, chacune vers un zéro de f, puisque à la limite $f(l) = g(g(l)) - l = 0$. Puisque f est strictement décroissante, f n'admet au plus qu'un seul zéro ; or α est un zéro de f donc les deux suites u_{2n} et u_{2n+1} convergent toutes les deux vers α et u_n aussi.

Remarque V.8. Ce résultat pouvait être démontré en utilisant la proposition N.2.

♢

(c) (i) On pose

$$\varepsilon_1 = 10^{-1} \text{ et } \varepsilon_2 = 10^{-15}. \tag{V.30}$$

On considère la valeur de k donnée par (V.19) et a et b définis par

$$a = \cos(1) \text{ et } b = 1. \tag{V.31}$$

Utilisons la proposition 4.21 du cours. Ainsi, les deux entiers n_1 et n_2 tels que $|u_n - l| \leq \varepsilon_1$ et $|u_n - l| \leq \varepsilon_2$ sont donnés respectivement par

$$n_1 = 9 \text{ et } n_2 = 196. \tag{V.32}$$

Attention, puisque que u_0, n'est pas dans l'intervalle $I = [a, b]$ et puisque l'on a appliqué la règle (V.22), il faut rajouter 2 à chacun des entiers définis précédemment de sorte que

$$n_1 = 11 \text{ et } n_2 = 198. \tag{V.33}$$

n	u_n
0	10.000000000000000
1	−0.839071529076452
2	0.668153917531387
3	0.784966720933852
4	0.707411791257438
5	0.760046415906649
6	0.724804033046625
7	0.748629360225378
8	0.732622462482921
9	0.743423001354606
10	0.736156148510940
11	0.741054959438722

TABLE V.1. 12 premières valeurs de la méthode du point fixe pour $g(x) = \cos(x)$.

(ii)

Donnons dans le tableau (V.1), les $n_1 + 1$ premières valeurs de u_n.

Rappelons que α, unique solution de $x = \cos(x)$ déterminé de façon exacte [2], est donné par

$$\alpha = 0.7390851332151606$$

On a

$$|\alpha - u_{n_1}| \approx 1.9698\,10^{-3},$$

ce qui est bien inférieur à la valeur de $\varepsilon_1 = 10^{-1}$ donnée dans l'énoncé.

On pourrait remarquer que pour $n = 198$, on a

$$|\alpha - u_n| \approx 3.9540\,10^{-16},$$

ce qui est strictement inférieur à la valeur de $\varepsilon_2 = 10^{-15}$ donnée dans l'énoncé.

(2) (a) La seule méthode connue d'ordre strictement plus grand que 1 est, dans ce cours, la méthode de Newton, censée être quadratique, donc convergeant plus rapidement que la méthode du point fixe précédemment et qui n'est que linéaire.

(b) On rappelle que la fonction h est définie par (V.15). La valeur de α définie dans la question (1) est donc un zéro de la fonction h qui est unique sur $\mathbb{R}$ d'après l'étude faite dans la remarque V.5 page 269 et qui comme précédemment est noté α. Pour l'intervalle I définie par

$$I = [0, \pi/2], \tag{V.34}$$

on a

$$\text{signe}(h(0)h(\pi/2)) = -1, \tag{V.35}$$

et l'unique zéro de h est donc dans l'intérieur de l'intervalle I. La relation liant w_{n+1} à w_n est donnée par la définition (4.91) appliquée à la fonction h. On a donc ici d'après (V.15) et (V.20),

$$x - \frac{h(x)}{h'(x)} = x - \frac{x - \cos(x)}{1 + \sin(x)} = \frac{x + x\sin(x) - x + \cos(x)}{1 + \sin(x)},$$

2. en fait très précise grâce à l'une des fonction `solve` ou `fzero` de matlab.

et donc

$$w_{n+1} = G(w_n), \tag{V.36a}$$

où

$$G(x) = \frac{x\sin(x) + \cos(x)}{1 + \sin(x)}. \tag{V.36b}$$

(c) Cette correction est très proche de la la correction de l'exercice 4.3 page 45 de TD et notamment de la question 2 page 45 auxquelles on renvoie pour plus de détail.

(i) Voir la proposition 4.54 du cours. On y a vu que la la méthode de Newton est quadratique ssi $h'(\alpha) \neq 0$ et si $h''(\alpha) \neq 0$. On sait que $h'(\alpha)$ est non nul, (d'après la question 1 puisque $h' > 0$), et le cours assure que $G'(\alpha) = 0$. Ainsi, la méthode est au moins quadratique.

De plus, elle est exactement quadratique si et seulement si $G''(\alpha) \neq 0$, ce qui est équivalent à $h''(\alpha) \neq 0$.

D'après (V.20), h' est on nul à l'intérieur de l'intervalle I auquel on sait que la racine α appartient. Ainsi, d'après la proposition 4.54, la méthode de Newton est au moins quadratique. Enfin, d'après (V.20), on a

$$\forall x \in \mathbb{R}, \quad h''(x) = \cos(x), \tag{V.37}$$

qui est non nul à l'intérieur de l'intervalle I à laquelle on sait que la racine α appartient.

Ainsi, d'après la proposition 4.54, la méthode de Newton est exactement quadratique.

(ii) Voir les propositions 4.38 et 4.54 du cours.

On y a vu que le développement de Taylor de la fonction g sur $[x^*, x_n]$ permet de montrer que

$$|x_{n+1} - x^*| \leq D\,|x_n - x^*|^2\,, \tag{V.38}$$

avec

$$D = \frac{1}{2}\max_{x\in I} |G''(x)| \tag{V.39}$$

en admettant d'abord *a priori* que la suite x_n converge et que les x_n sont dans l'intervalle I.

La constante D est donnée par la formule (4.97) du cours qui peut aussi être remplacée avantageusement par (4.101).

Remarque V.9. Justifions maintenant rigoureusement la convergence de la suite w_n pour tout x_0 appartenant à I. Il suffit pour cela d'invoquer la proposition Y.1,appliquée à la fonction h, dont les différentes hypothèses sont vérifiées sur l'intervalle I. En effet, :

(A) l'hypothèse 1) est clairement vérifiée ;

(B) l'hypothèse 2) est clairement vérifiée d'après (V.35) ;

(C) l'hypothèse 3) est vérifiée d'après (V.20), puisque le sinus est positif ou nul sur I ;

(D) l'hypothèse 4) est vérifiée d'après (V.37). Attention, en $\pi/2$, le cosinus s'annule mais on peut montrer que si on part de $x_0 = \pi/2$ ou de $x_0 < \pi/2$, on reste dans tous les cas dans un intervalle strictement inclus dans $[0, \pi/2[$.

(E) l'hypothèse 5) est vérifiée d'après les valeurs numériques suivantes :

$$\left(\frac{\pi}{2}\right)^{-1} \frac{|h(0)|}{|h'(0)|} \approx 0.63661977236758 < 1,$$
$$\left(\frac{\pi}{2}\right)^{-1} \frac{\left|h\left(\frac{\pi}{2}\right)\right|}{\left|h'\left(\frac{\pi}{2}\right)\right|} \approx 0.50000000000000 < 1.$$

♢

(iii) Il suffit de raisonner comme dans la question 4 page 47 de la correction de l'exercice de TD 4.3 ou dans la proposition 4.33.

La constante D est donnée par la formule (4.101) du cours qui fournit donc sur l'intervalle I :

$$D = \max_{x\in I} \frac{|\cos(x)|}{|1+\sin(x)|}.$$

On a donc, pour tout $x \in I = [0, \pi/2]$:

$$\frac{|\cos(x)|}{|1+\sin(x)|} = \frac{\cos(x)}{1+\sin(x)} \leq \frac{\cos(x)}{1} \leq \cos x \leq 1.$$

Dans ce cas, il est nécessaire pour appliquer la proposition 4.33 que la majoration (4.52) soit vérifiée ; or, ici, en prenant $e_0 = \pi/2$ et $p = 2$ (méthode quadratique) , on a

$$|e_0| D^{\left(\frac{1}{p-1}\right)} = \frac{\pi}{2},$$

qui n'est pas majoré strictement par 1. Si au contraire, on se place sur l'intervalle $J = [\nu, \pi/2-\nu]$ avec $0 < \nu < \pi/4$, on a clairement pour tout $x \in J$:

$$\frac{|\cos(x)|}{|1+\sin(x)|} = \frac{\cos(x)}{1+\sin(x)}$$

et donc

$$D \leq \frac{\cos(\nu)}{1+\sin\nu}. \tag{V.40}$$

Avec le choix de ν donné dans l'énoncé, on vérifie d'une part que la proposition Y.1,s'applique pour la fonction h, comme fait dans la remarque V.9, et d'autre part que l'on a bien, d'après (V.40), le résultat donné dans l'énoncé :

$$D \leq 0.73741536. \tag{V.41}$$

♢

Dans ce cas, la majoration (4.52) est vérifiée puisque numériquement en prenant $p = 2$ (méthode quadratique), on a

$$\left(\frac{\pi}{2} - 2\nu\right) D^{\left(\frac{1}{p-1}\right)} = 0.715880114974; \tag{V.42}$$

et on peut donc appliquer la la proposition 4.33. Il ne reste plus qu'à appliquer l'équation (4.54) rappelée ici :

$$n = \left\lceil \frac{1}{\ln p} \ln\left(\frac{\ln\frac{\varepsilon}{\gamma}}{\ln\delta}\right)\right\rceil, \tag{V.43}$$

avec γ et δ définis par l'équation (4.50) c'est-à-dire

$$\gamma = C^{\left(\frac{1}{1-p}\right)}, \tag{V.44a}$$

$$\delta = |e_0| C^{\left(\frac{1}{p-1}\right)}, \tag{V.44b}$$

Numériquement, on a pour $p = 2$, on a pour $|e_0| = \frac{\pi}{2} - 2\nu$, on a

$$\gamma = 0.737415351928,$$
$$\delta = 0.715880114974.$$

puis pour les deux entiers m_1 et m_2, associé à ε_1 et ε_2, on a

$$m_1 = 3,$$
$$m_2 = 7,$$

qui sont des valeurs beaucoup plus faibles que celles données par (V.32) ! Cela nous montre *a posteriori* l'efficacité accrue de la méthode de Newton !

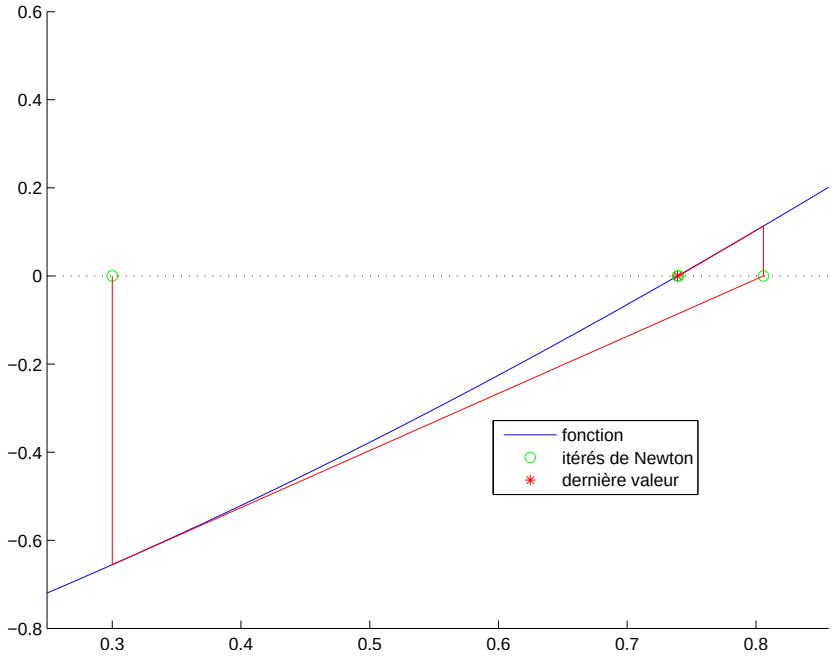


FIGURE V.3. Le tracé graphique des valeurs de w_n de la méthode de Newton pour $w_0 = 0.3000$.

n	w_n
0	0.300000000000000
1	0.805848141739496
2	0.740002131837112
3	0.739085318708854
4	0.739085133215168
5	0.739085133215161

TABLE V.2. 6 premières valeurs de la méthode de Newton.

(iv) Concluons par des simulations numériques.
Voir la figure V.3 et le tableau V.2
Enfin, on a

$$|\alpha - u_{m_1}| \approx 9.1699\,10^{-4},$$

ce qui est bien inférieur à la valeur de $\varepsilon_1 = 10^{-1}$ donnée dans l'énoncé.
On pourrait remarquer que pour $n = 5$, on a

$$|\alpha - u_n| \approx 3.0638\,10^{-17},$$

ce qui est strictement inférieur à la valeur de $\varepsilon_2 = 10^{-15}$ donnée dans l'énoncé.

Annexe W

Exemple d'une méthode de Newton divergente partout

Ce problème a été donné à l'examen d'Automne 2022.

Énoncé

(1) Pour cette question et la question 2, on considère f, définie sur $\mathbb{R}$ par

$$\forall x \in \mathbb{R}, \quad f(x) = \begin{cases} \sqrt{x} & \text{si } x \geq 0, \\ -\sqrt{-x} & \text{si } x < 0. \end{cases} \tag{W.1}$$

Étudier sommairement et tracer la fonction f. On montrera en particulier que

$$\forall x \in \mathbb{R}, \quad f'(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & \text{si } x > 0, \\ \\ \dfrac{1}{2\sqrt{-x}} & \text{si } x < 0. \end{cases} \tag{W.2}$$

(2) (a) (i) Pour $u_0 \in \{1, 4\}$, déterminer les premières valeurs de la méthode de Newton associée à la recherche de l'unique zéro de f. On pourra illustrer les valeurs obtenues sur le graphe de la question 1. Que remarquez-vous ?

(ii) Pour tout $x \in \mathbb{R}$, on note u_n les valeurs de la méthode de Newton définie par $u_0 = x$. Montrer que pour tout $x > 0$, on a

$$\forall n \in \mathbb{N}, \quad u_n = \begin{cases} x & \text{si } n \text{ est pair}, \\ -x & \text{si } n \text{ est impair}. \end{cases} \tag{W.3}$$

(iii) Qu'en est-il si $x < 0$ et $x = 0$?

(iv) Conclure sur la convergence de la méthode de Newton pour la fonction f.

(b) Pourriez-vous, en utilisant l'un des méthodes vues en cours, approcher numériquement l'unique zéro de f. Faites quelques simulations numériques pour cette méthode choisie.

(3) *Question facultative*

On cherche maintenant s'il existe d'autres fonctions que f présentant le comportement observé pour la méthode de Newton dans les question 1 et 2.

(a) Soit f de $\mathbb{R}$ dans $\mathbb{R}$, définie sur $\mathbb{R}$, que l'on suppose nulle en zéro seulement, impaire et dérivable sur $\mathbb{R}^*$. Montrer que pour $u_0 > 0$, on a

$$u_1 = -u_0 \iff 2u_0 - \frac{f(u_0)}{f'(u_0)} = 0. \tag{W.4}$$

et que si (W.4) est valable alors

$$u_2 = u_0. \tag{W.5}$$

(b) Que peut-on en déduire sur u_n pour $n \in \mathbb{N}$?

(c) On cherche f telle que (W.4) ait lieu pour tout $u_0 > 0$, ce qui revient à écrire

$$\forall x > 0, \quad \frac{f(x)}{f'(x)} = 2x.$$

Montrer que cela est équivalent à

$$\forall x > 0, \quad \frac{f'(x)}{f(x)} = \frac{1}{2x}$$

et en déduire f sur $\mathbb{R}_+^*$ en supposant que f'/f est de signe constant sur $\mathbb{R}_+^*$. En déduire l'expression de f sur $\mathbb{R}$ et conclure.

Corrigé

(1) La fonction f, définie sur $\mathbb{R}$ par

$$\forall x \in \mathbb{R}, \quad f(x) = \begin{cases} \sqrt{x} & \text{si } x \geq 0, \\ -\sqrt{-x} & \text{si } x < 0, \end{cases} \tag{W.6}$$

est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^*$ et il est clair que, si $x > 0$,

$$f'(x) = \frac{1}{2\sqrt{x}}.$$

Si $x < 0$,

$$\begin{aligned} f'(x) &= -\left(\sqrt{-x}\right)', \\ &= \frac{1}{2\sqrt{-x}}. \end{aligned}$$

On a donc

$$\forall x \in \mathbb{R}, \quad f'(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & \text{si } x > 0, \\ \\ \dfrac{1}{2\sqrt{-x}} & \text{si } x < 0. \end{cases} \tag{W.7}$$

Il est évident que f est impaire, que $f(0) = 0$ et que f est strictement croissante sur $\mathbb{R}$. Voir son graphe et celui de sa dérivée sur la figure W.1.

(2) (a) (i) À la main, ou en utilisant par exemple la fonction `newton.m` disponible sur le site à l'adresse habituelle, on obtient les valeurs suivantes de u_n pour $u_0 = 1$

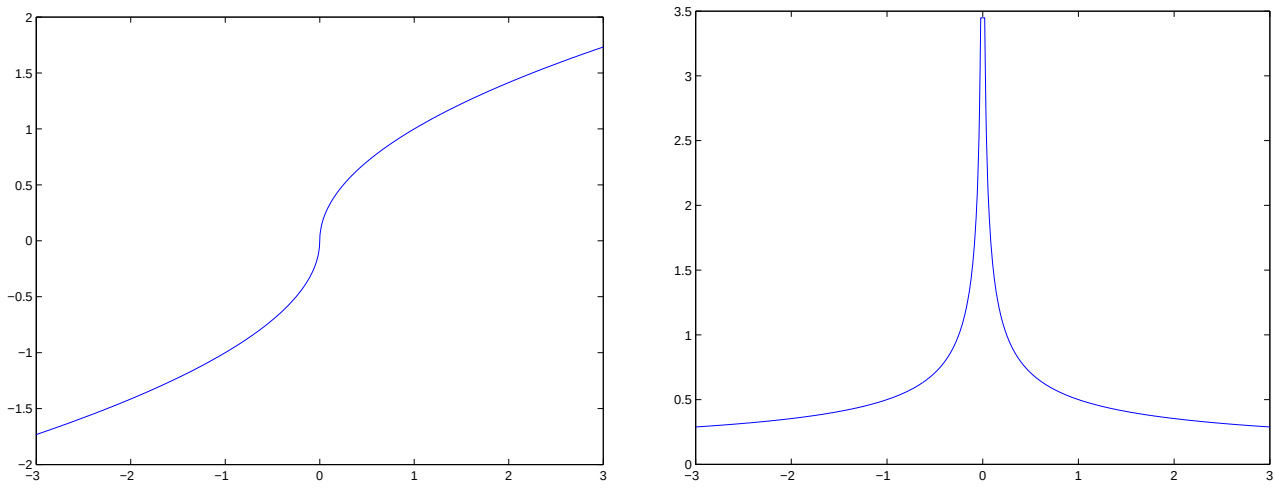
$$1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ ....$$

et pour $u_0 = 4$, on obtient

$$4,\ -4,\ 4,\ -4,\ 4,\ -4,\ 4,\ -4,\ 4,\ -4,\ 4,\ -4,\ ....$$

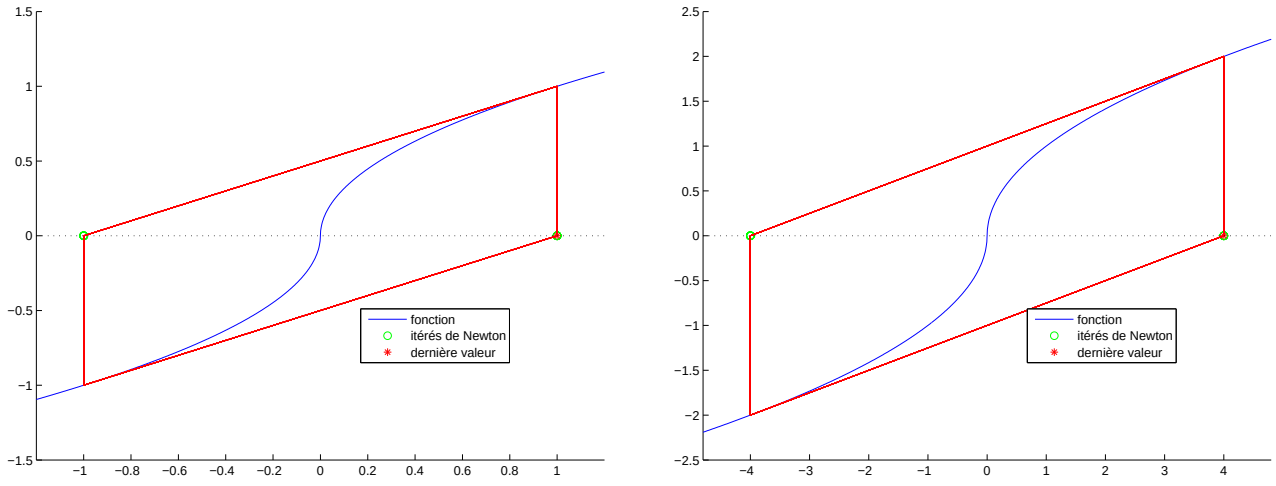
On pourra taper, sous matlab, grâce à la fonction `newton.m` disponible sur le site à l'adresse habituelle, les commandes suivantes

```
fun=@(x)(x>=0).*sqrt(x)-(x<0).*sqrt(-x);
dfun=@(x)(1/2)*((x>=0)./sqrt(x)+(x<0)./sqrt(-x));
[xvecta,xdif,fx,nit]=newton(1,-1,20,fun,dfun,1);
figure;
[xvectb,xdif,fx,nit]=newton(4,-1,20,fun,dfun,1);
disp(xvecta);
disp(xvectb);
```



(a) f (b) f'

FIGURE W.1. Graphes de f et de f'.



(a) $u_0 = 1$ (b) $u_0 = 4$

FIGURE W.2. Quelques valeurs de la méthode de Newton

ce qui donnera les graphiques de la figure W.2. On constate que les valeurs sont cycliques avec, pour $x = u_0 \in \{1, 4\}$,

$$\forall n \in \mathbb{N}, \quad u_n = \begin{cases} x & \text{si } n \text{ est pair,} \\ -x & \text{si } n \text{ est impair.} \end{cases} \tag{W.8}$$

(ii) De façon générale, on note u_n les valeurs de la méthode de Newton définie par $u_0 = x$. Montrons que pour tout $x > 0$, on a

$$\forall n \in \mathbb{N}, \quad u_n = \begin{cases} x & \text{si } n \text{ est pair,} \\ -x & \text{si } n \text{ est impair.} \end{cases} \tag{W.9}$$

On considère g donnée par

$$\forall x \in \mathbb{R}^*, \quad g(x) = x - \frac{f(x)}{f'(x)}. \tag{W.10}$$

On a, d'après (W.6)-(W.7) si $x > 0$,

$$\begin{aligned} g(x) &= x - \frac{\sqrt{x}}{\frac{1}{2\sqrt{x}}}, \\ &= x - 2\sqrt{x}\sqrt{x}, \\ &= x - 2\big(\sqrt{x}\big)^2, \\ &= x - 2x, \\ &= -x. \end{aligned}$$

On vérifie aussi, en même temps que, si $x < 0$

$$\begin{aligned} g(x) &= x + \frac{\sqrt{-x}}{\frac{1}{2\sqrt{-x}}}, \\ &= x + 2\sqrt{-x}\sqrt{-x}, \\ &= x + 2\big(\sqrt{-x}\big)^2, \\ &= x - 2x, \\ &= -x. \end{aligned}$$

et donc,

$$\forall x \in \mathbb{R}^*, \quad g(x) = -x. \tag{W.11}$$

Puisque la méthode de Newton est définie par $u_{n+1} = g(u_n)$, on a, si $u_0 = x > 0$

$$u_1 = g(u_0) = -u_0 < 0$$

et donc

$$u_2 = g(u_1) = -u_1 = u_0,$$

et par récurrence, on montre que

$$\forall n \in \mathbb{N}, \quad u_n = \begin{cases} x & \text{si } n \text{ est pair,} \\ -x & \text{si } n \text{ est impair.} \end{cases} \tag{W.12}$$

Cela est vrai aussi si $x < 0$. On vérifie aussi que, puisque f n'est pas dérivable en zéro.

$$\text{pour } x = 0, \text{ la méthode de Newton n'est pas définie.} \tag{W.13}$$

Néanmoins, on peut considérer que g est définie par continuité en zéro, d'après (W.11), en posant

$$g(0) = 0. \tag{W.14}$$

On a donc, immédiatement, avec cette convention, si $x = u_0 = 0$

$$\forall n \in \mathbb{N}, \quad u_n = 0. \tag{W.15}$$

(iii) Voir question 2(a)ii

(iv) D'après (W.12), vraie aussi pour $x < 0$,

$$\text{la suite de la méthode de Newton est cyclique et ne converge pas si } u_0 \neq 0. \tag{W.16}$$

et, en adoptant la convention (W.14), on a, d'après (W.15)

$$\text{la suite de la méthode de Newton est constante et converge vers zéro si } u_0 = 0. \tag{W.17}$$

REMARQUE W.1. Cet exemple est un peu pathologique, pour ne pas dire capilotracté ! Les hypothèses habituellement faites sur la méthode de Newton exigent que f est dérivable en la racine, mieux à dérivée non nulle (voir les propositions du cours 4.54 et 4.62).

(b) Les hypothèses de la méthode du point fixe (voir proposition 4.11 du cours) ne sont pas assurées, puisque f n'est pas dérivable, et sa dérivée sur l'intervalle $[-1,1]\setminus\{0\}$ n'est pas bornée, puisqu'elle tend vers $+\infty$ quand x tend vers zéro ! La seule méthode qui fonctionne ici est la méthode de la dichotomie. À la main, ou en utilisant par exemple la fonction `bisection.m` disponible sur le site à l'adresse habituelle, et en tapant

```
[xvect,xdif,fx,nit]=bisection(-1,1,0,150,fun)
```

en obtient le zéro exact en une seule itération, ce qui est normal, par symétrie. Si on tape par exemple

```
[xvect,xdif,fx,nit]=bisection(-1,1.1,eps,150,fun)
```

on obtient la valeur suivante $1.033563\,10^{-16}$ en 54 itérations.

(3) (a) Soit f, définie sur $\mathbb{R}$, que l'on suppose nulle en zéro seulement, impaire et dérivable sur $\mathbb{R}^*$. Soit $u_0 > 0$. On a par définition de la méthode de Newton

$$u_1 = u_0 - \frac{f(u_0)}{f'(u_0)}.$$

Ainsi,

$$u_1 = -u_0 \tag{W.18}$$

est équivalent à

$$2u_0 - \frac{f(u_0)}{f'(u_0)} = 0 \tag{W.19}$$

Sous cette hypothèse, on a alors, grâce à (W.18),

$$u_1 \neq 0, \tag{W.20}$$

on a, par imparité (si f est impaire, sa dérivée est paire)

$$\begin{aligned} u_2 &= u_1 - \frac{f(u_1)}{f'(u_1)}, \\ &= -u_0 - \frac{f(-u_0)}{f'(-u_0)}, \\ &= -u_0 + \frac{f(u_0)}{f'(u_0)}, \\ &= -u_0 + 2u_0 \end{aligned}$$

et donc

$$u_2 = u_0. \tag{W.21}$$

(b) On en déduit (W.8).

(c) On cherche f telle que (W.19) ait lieu pour tout $u_0 > 0$, ce qui revient à écrire

$$\forall x > 0, \quad \frac{f(x)}{f'(x)} = 2x, \tag{W.22}$$

ce qui est équivalent, par hypothèse sur f à

$$\forall x > 0, \quad \frac{f'(x)}{f(x)} = \frac{1}{2x}. \tag{W.23}$$

On résout cette équation différentielle de la façon suivante : Considérons l'équation différentielle

$$\forall t \in [\tau, +\infty[, \quad 2ty'(t) - y(t) = 0. \tag{W.24}$$

On a deux façons de procéder.

(i) Soit, on raisonne comme dans [Bas22a, Chapitre "Équations différentielles (ordinaires)", section "Équations différentielles d'ordre un"] disponible sur `http://utbmjb.chez-alice.fr/Polytech/MFI/coursMFI.pdf`, en écrivant et en supposant y non nul (et par exemple strictement positif)

$$\begin{aligned}
2ty(t)' - y(t) = 0 &\iff \frac{y'(t)}{y(t)} = \frac{1}{2t}, \\
&\iff (\ln|y|)' = \left(\frac{1}{2}\ln|t|\right)', \\
&\iff (\ln|y|)' = \left(\ln\sqrt{t}\right)', \\
&\iff \ln(y) = c + \ln\sqrt{t}, \\
&\iff y = e^{c+\ln\sqrt{t}}, \\
&\iff y = e^c e^{\ln\sqrt{t}}, \\
&\iff y = c\sqrt{t}
\end{aligned}$$

où $K = e^c$. On a donc

$$y(t) = c\sqrt{t}. \tag{W.25}$$

(ii) Soit on applique directement le résultat de [Bas22a, Section "Équations différentielles d'ordre 1" de l'annexe "Théorie des équations différentielles linéaires à coefficients constants d'ordre 1 et 2"] disponible sur `http://utbmjb.chez-alice.fr/Polytech/MFI/coursMFI.pdf`. On sait, d'après le cours, en notant $a(t) = 2t$ et $b(t) = -1$, que y est solution de (W.24) ssi

$$y(t) = ce^{-\alpha(t)}$$

où c est un réel et α une primitive de $b/a = -1/(2t)$, soit

$$\alpha(t) = -1/2\ln|t| = -1/2\ln t = -\ln\sqrt{t}.$$

On a donc

$$\begin{aligned}
y(t) &= ce^{-\alpha(t)}, \\
&= ce^{\ln\sqrt{t}},
\end{aligned}$$

et donc

$$y(t) = c\sqrt{t}. \tag{W.26}$$

où c est un réel quelconque.

Finalement, on obtient, en prolongeant f par continuité en zéro, et en utilisant les hypothèses faites sur f :

$$\exists c \in \mathbb{R}, \quad \forall x \in \mathbb{R}, \quad f(x) = \begin{cases} c\sqrt{x} & \text{si } x \geq 0, \\ -c\sqrt{-x} & \text{si } x < 0. \end{cases} \tag{W.27}$$

On retrouve bien, à une constante multiplicative près, la fonction f donné par l'équation (W.6).

REMARQUE W.2. On réécrit (W.23) sous la forme

$$\forall x > 0, \quad f'(x) = \frac{1}{2}\frac{f(x) - f(0)}{x}.$$

Si f est dérivable en zéro, on obtient quand x tend vers zéro :

$$\lim_{x \to 0} \quad f'(x) = \frac{1}{2} f'(0),$$

et d'après le théorème de la limite de la dérivée, on a donc (dans $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$)

$$f'(0) = \frac{1}{2} f'(0),$$

ce qui implique que, soit $f'(0) = 0$ soit $f'(0) = +\infty$, ce qui est le cas pour a fonction f donné par l'équation (W.6).

REMARQUE W.3. Si on se passe de l'hypothèse d'imparité faite sur f, on laisse au lecteur le soin de vérifier que l'on obtient l'équation différentielle, dite non résolue, beaucoup plus difficile (probablement impossible) à résoudre

$$\forall x > 0, \quad \frac{f(x)}{f'(x)} + \frac{f\left(x - \frac{f(x)}{f'(x)}\right)}{f'\left(x - \frac{f(x)}{f'(x)}\right)} = 0. \tag{W.28}$$

Annexe X

Racines multiples

Dans cette annexe, nous rappelons tout d'abord la notion de racines multiples usuelles pour un polynôme puis nous l'étendons à une fonction quelconque

X.1. Racines multiples d'un polynôme

On pourra consulter par exemple `https://fr.wikipedia.org/wiki/Racine_d%27un_polyn%C3%B4me` ou [RDO93, section 6.4.3 2)].

On a la définition-proposition suivante

Définition X.1 (Ordre de multiplicité, racine simple, racine multiple). Soit P un polynôme, à coefficients réels ou complexes et a un nombre réel ou complexe. Alors,

(1) le plus grand entier $m \in \mathbb{N}^*$ tel que P soit divisible par $(X-a)^m$ est appelé l'ordre, ou la multiplicité, de la racine a relativement à P ;

(2) cet entier m est caractérisé par l'existence d'un polynôme Q tel que $P = (X-a)^m Q$ et $Q(a) \neq 0$;

(3) cet entier m est caractérisé par

$$P(a) = P'(a) = ... = P^{(m-1)}(a) = 0, \tag{X.1a}$$

$$P^{(m)}(a) \neq 0. \tag{X.1b}$$

(4) on dit que a est racine simple de P si $m = 1$ et racine multiple si $m > 1$.

Démonstration de l'équivalence des cas 1,2 et 3.

(1) Remarquons que l'énoncé du cas 1 est équivalent à

$$P \text{ est divisible par } (X-a)^m \text{ et } P \text{ n'est pas divisible par } (X-a)^{m+1}. \tag{X.2}$$

Démontrons donc l'équivalence du cas 2 et de (X.2), ce qui est fait dans [RDO93, section 6.4.3 2) Théorème II et définition], que l'on rappelle ici :

(a) Démontrons que le cas 2 implique (X.2).
Il est clair, dans ce cas, que P est divisible par $(X-a)^m$
Montrons maintenant que P n'est pas divisible par $(X-a)^{m+1}$. Faisons la division Euclidienne de Q par $X-a$: il existe donc un polynôme S et un réel (ou un complexe) λ tel

$$Q = (X-a)S + \lambda.$$

Si on choisit $X = a$, on a $\lambda = Q(a)$ et donc

$$Q = (X-a)S + Q(a), \tag{X.3}$$

dont on déduit que

$$Q(X-a)^m = (X-a)^{m+1}S + Q(a)(X-a)^m.$$

Du cas 2, on déduit donc que

$$P = (X-a)^m Q = (X-a)^{m+1}S + Q(a)(X-a)^m$$

et donc que le reste de la division de P par $(X-a)^{m+1}$ est le polynôme $Q(a)(X-a)^m$ qui est non nul puisque $Q(a) \neq 0$. Ainsi, P n'est pas divisible par $(X-a)^{m+1}$.

(b) Montrons maintenant que (X.2) implique le cas 2. L'hypothèse se traduit par $P = Q(X-a)^m$ avec Q non divisible par $X-a$. Or, d'après (X.3) dire que Q est divisible par $X-a$ revient à dire que $Q(a) = 0$. Ainsi, $Q(a) \neq a$.

(2) Démontrons que le cas 2 est équivalent au cas 3, ce qui est fait dans [RDO93, section 6.4.3 2) Théorème III], que l'on rappelle ici :

Rappelons la formule de Taylor à l'ordre p appliquée au polynôme P (de degré p). Voir le Théorème 4.36. Elle donne ici : il existe ξ tel que

$$P = \sum_{n=0}^{p} \frac{P^{(n)}(a)}{n!}(X-a)^n + \frac{P^{(p+1)}(\xi)}{(p+1)!}(X-a)^{p+1}.$$

et donc, puisque $P^{(p+1)} = 0$:

$$\begin{aligned} P &= \sum_{n=0}^{p} \frac{P^{(n)}(a)}{n!}(X-a)^n, \\ &= \sum_{n=0}^{m-1} \frac{P^{(n)}(a)}{n!}(X-a)^n + \frac{P^{(m)}(a)}{m!}(X-a)^m + \sum_{n=m+1}^{p} \frac{P^{(n)}(a)}{n!}(X-a)^n, \end{aligned}$$

et donc

$$P = \sum_{n=0}^{m-1} \frac{P^{(n)}(a)}{n!}(X-a)^n + (X-a)^m \left(\frac{P^{(m)}(a)}{m!} + \sum_{n=m+1}^{p} \frac{P^{(n)}(a)}{n!}(X-a)^{n-m} \right)$$

Cette expression met en valeur le quotient

$$Q = \frac{P^{(m)}(a)}{m!} + \sum_{n=m+1}^{p} \frac{P^{(n)}(a)}{n!}(X-a)^{n-m},$$

et le reste

$$R = \sum_{n=0}^{m-1} \frac{P^{(n)}(a)}{n!}(X-a)^n$$

de la division de P par $(X-a)^m$. En particulier, il vient

$$Q(a) = \frac{P^{(m)}(a)}{m!}$$

D'après le cas 2, a est racine d'ordre m ssi $P = (X-a)^m Q$ et $Q(a) \neq 0$, cela est aussi équivalent à $R = 0$ et $Q(a) \neq 0$ et donc en comparant avec l'expression du reste R et de $Q(a)$ à (X.1).

□

On déduit de la définition (X.1), le lemme suivant, laissé à la sagacité du lecteur :

LEMME X.2. *Le nombre a est racine simple de P ssi $P(a) = 0$ et $P'(a) \neq 0$.*

X.2. Racines multiples d'une fonction quelconque

Soient $a \in \mathbb{R}$, $\varepsilon > 0$ et On considère désormais une fonction f définie sur un intervalle $I_\varepsilon = [a-\varepsilon, a+\varepsilon]$. On donne alors le résultat suivant :

LEMME X.3. *Soient deux entiers m, $p \in \mathbb{N}^*$ tels que $m \leq p$ et f une fonction de I_ε dans $\mathbb{R}$, de classe $\mathcal{C}^p$, vérifiant*

$$f(a) = f'(a) = ... = f^{(m-1)}(a) = 0. \tag{X.4}$$

Alors la fonction g définie par

$$\forall t \in I_\varepsilon, \quad g(t) = \begin{cases} \dfrac{f(t)}{(t-a)^m} & \text{si } t \neq a, \\ \dfrac{f^{(m)}(a)}{m!} & \text{si } t = a, \end{cases} \tag{X.5}$$

est de classe $\mathcal{C}^{p-m}$.

DÉMONSTRATION. Voir [RDO88, section 6.7.2 6)]. □

On en déduit le résultat suivant :

LEMME X.4. *On suppose que f vérifie de plus*

$$f^{(m)}(a) \neq 0, \tag{X.6}$$

alors,

$$\forall t \in I_\varepsilon, \quad f(t) = (t-a)^m g(t), \tag{X.7}$$

avec

$$g(a) \neq 0. \tag{X.8}$$

DÉMONSTRATION. Il suffit de remarquer que $g(a) = \frac{f^{(m)}(a)}{m!} \neq 0$. □

Enfin, on a le lemme suivant :

LEMME X.5. *Soient un entier $m \in \mathbb{N}^*$, f et g deux fonctions de I_ε dans $\mathbb{R}$, de classe $\mathcal{C}^p$, vérifiant* (X.7) *et* (X.8)*. Alors, f vérifie* (X.4) *et* (X.6)*.*

DÉMONSTRATION. Soit $k \leq m-1$. Pour dériver k fois (X.7) il suffit d'appliquer la formule de Leibniz :

$$(uv)^{(k)} = \sum_{l=0}^{k} \binom{l}{k} u^{(l)} v^{(k-l)}. \tag{X.9}$$

Cette formule, tout à fait identique à la formule du binôme de Newton, est démontrée par exemple dans `https://fr.wikipedia.org/wiki/Formule_de_Leibniz` et `https://fr.wikipedia.org/wiki/R%C3%A8gle_du_produit`. On applique cette formule à (X.7) avec $u = (t-\alpha)^m$ et $v = g$. Chacune des dérivées l-ième pour $l \leq k \leq m-1$ de u est donc nulle en a et on a donc

$$f^{(k)}(a) = 0,$$

et on a donc (X.4). Si on prend $k = m$, alors toutes les dérivées en a sont nulles sauf la m-ième dérivée (qui corespond à $l = m$ dans (X.9)) qui vaut $m!$. On a donc

$$f^{(m)}(a) = m! g(a),$$

qui est non nul d'après (X.8). □

Autrement dit, les résultats du lemmes X.3, X.4 et X.5 généralisent l'équivalence des cas 2 et 3 de la définition X.1. Ces lemmes correspondent donc au cas où f présente une racine dite d'ordre m en a.

On peut donc donner la définition suivante d'une racine d'ordre m d'une fonction f.

Définition X.6. Soient f une fonction suffisamment régulière sur I_ε Les deux assertions suivantes sont équivalentes

(1) Il existe une fonction g suffisamment régulière sur I_ε vérifiant (X.7) et (X.8).

(2) On a (X.4) et (X.6).

Dans ce cas, on dit que a est racine simple de f si $m = 1$ et racine multiple si $m > 1$.

Annexe Y

Convergence globale de la méthode de Newton

Complétons le résultat de la section 4.5.2 page 100.

THÉORÈME Y.1.
Sous les hypothèses suivantes :

(1) f *de classe* C^2 *sur* $[a,b]$,

(2) $f(a)f(b)<0$,

(3) $\forall x\in[a,b]\quad f'(x)\neq 0$,

(4) f'' *de signe constant (non nul) sur* $[a,b]$,

(5) $|f(a)|/|f'(a)|<b-a$ *et* $|f(b)|/|f'(b)|<b-a$,

la suite (x_n) *des itérés de Newton relative à l'équation* (4.16) *et définie par* (4.91) *converge pour tout choix de* x_0 *dans* $[a,b]$ *vers l'unique l'unique solution* r *de* (4.16) *dans* $[a,b]$. *Si de plus,* f *de classe* C^3 *sur* $[a,b]$, *la méthode est quadratique.*

DÉMONSTRATION. Voir [BM03, Théorème 4.31] et théorème 4.54. □

REMARQUE Y.2. On peut ne supposer f que de classe $\mathcal{C}^2$, en utilisant la proposition 4.62 page 101.

EXEMPLE Y.3 (Exemple d'application du théorème Y.1).
Cet exemple est issu de [Bré19a]. On considère f définie sur $\mathbb{R}$ par

$$\forall x\in\mathbb{R},\quad f(x)=x-e^{-x}.$$

f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ et on a

$$\forall x\in\mathbb{R},\quad f'(x)=1+e^{-x},\tag{Y.1}$$

qui est srictement positive sur $\mathbb{R}$. Puisque $f(-\infty)=-\infty$ et que $f(\infty)=\infty$, f admet une unique racine sur $\mathbb{R}$. De plus, $f(0)=-1$ et $f(1)=1-1/e>0$ donc cette racine est dans $[0,1]$. On a

$$\forall x\in[0,1],\quad f''(x)=-e^{-x},\tag{Y.2}$$

qui est strictement négative. On a enfin

$$\frac{f(0)}{f'(0)}=-1/2,$$
$$\frac{f(1)}{f'(1)}=\frac{1-e^{-1}}{1+e^{-1}}\approx 0.46212.$$

Les hypothèses du théorème Y.1 sont vérifiées sur $[0,1]$. Si on choisit $x_0=0$, on obtient pour $n=6$ itérations, $x_n\approx 0.56714329040978$ et $f(x_n)\approx 1.1102\,10^{-16}$.

THÉORÈME Y.4.
Soit f *de* $[a,b]$ *dans* $\mathbb{R}$.

(1) On suppose qu'il existe $\varepsilon \in \{1,-1\}$ *et* $r \in [a,b[$ *tels que*

$$f(r) = 0, \tag{Y.3a}$$

$$\varepsilon f'(r) > 0, \tag{Y.3b}$$

$$f \text{ est de classe } \mathcal{C}^3 \text{ sur } [r,b], \tag{Y.3c}$$

$$\varepsilon f'' > 0 \text{ sur }]r,b]. \tag{Y.3d}$$

Alors la suite (x_n) *des itérés de Newton relative à l'équation* (4.16) *et définie par* (4.91) *converge pour tout choix de* x_0 *dans* $]r,b]$ *vers* r *et la méthode est au moins quadratique. Si, de plus,*

$$\varepsilon f''(r) > 0, \tag{Y.3e}$$

la méthode est quadratique.

(2) On suppose qu'il existe $\varepsilon \in \{1,-1\}$, $r \in]a,b]$ *vérifiant* (Y.3a) *et tels que*

$$\varepsilon f'(r) > 0, \tag{Y.4a}$$

$$f \text{ est de classe } \mathcal{C}^3 \text{ sur } [a,r], \tag{Y.4b}$$

$$\varepsilon f'' < 0 \text{ sur } [a,r[. \tag{Y.4c}$$

Alors la suite (x_n) *des itérés de Newton relative à l'équation* (4.16) *et définie par* (4.91) *converge pour tout choix de* x_0 *dans* $[a,r[$ *vers* r *et la méthode est au moins quadratique. Si, de plus,*

$$\varepsilon f''(r) > 0, \tag{Y.4d}$$

la méthode est quadratique.

Démonstration.

Ce théorème est adapté de [Bré19b].

Remarquons tout d'abord que si $x_0 = r$, alors la suite est constante et égale à r.

(1) Démontrons le point 1 avec $\varepsilon = 1$. Remarquons que l'hypothèse (Y.3d) implique que f' est strictement croissante sur $[r,b]$. D'après (Y.3b), on a donc

$$f' > 0 \text{ sur } [r,b]. \tag{Y.5}$$

f est donc strictement croissante sur $[r,b]$ et d'après (Y.3a), on a donc

$$f > 0 \text{ sur }]r,b]. \tag{Y.6}$$

Montrons, par récurrence sur n, que

$$\forall n \in \mathbb{N}, \quad r < x_n \leq b. \tag{Y.7}$$

Cela implique que x_{n+1} est bien défini. D'après le choix de x_0, c'est vrai pour $n = 0$. Supposons maintenant (Y.7) vraie pour un certain n. En considérant g définie par (4.92), on a

$$x_{n+1} - r = g(x_n) - g(r). \tag{Y.8}$$

Or, d'après (4.93), (Y.6) et (Y.3d), on a $g' > 0$ sur $]r,b]$ et donc g est strictement croissante sur $[r,b]$. D'après l'hypothèse de récurrence, on a

$$r < x_n \leq b, \tag{Y.9}$$

et donc $g(x_n) > g(r)$ et, d'après (Y.8) on a donc bien $x_{n+1} > r$. Par ailleurs, d'après (4.91), on a

$$x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)}. \tag{Y.10}$$

D'après (Y.5), (Y.6) et (Y.9), on a donc

$$x_{n+1} - x_n < 0, \tag{Y.11}$$

et donc, d'après (Y.9), $x_{n+1} < x_n \leq b$. Ainsi, la récurrence est établie. Enfin, en réutilisant (Y.10) et (Y.11), on constate que la suite (x_n) est décroissante et minorée par r et converge vers $l \in [r, b]$, qui est un zéro de f qui ne peut être que r. Sinon, on aurait $l > r$ et, d'après (Y.6), on aurait $f(l) > 0$. 1D'après (Y.3b) et la proposition 4.54 adaptée, la méthode est au moins quadratique. Si, de plus, on a (Y.3e), alors la proposition 4.54 implique que la méthode est quadratique.

(2) Le point 1 avec $\varepsilon = -1$. se montre en appliquant le point 1 avec $\varepsilon = 1$ et appliqué à $\widetilde{f} = -f$.

(3) Le point 2 est laissé au lecteur. On reprend la preuve du point 1 en adpatant et en montrant que, cette fois-ci, la suite est croissante, à valeurs dans dans $[a, r[$.

□

Remarque Y.5. On peut ne supposer f que de classe $\mathcal{C}^2$, en utilisant la proposition 4.62 page 101.

Exemple Y.6 (Exemple d'application du théorème Y.4). Soit $A > 0$. On considère f définie sur $\mathbb{R}$ par

$$\forall x \in \mathbb{R}, \quad f(x) = e^x - A. \tag{Y.12}$$

f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$. Elle n'a qu'un seul zéro, égal à $\ln(A)$. On peut appliquer le théorème Y.4 (cas 1 avec $\varepsilon = 1$) à l'intervalle $[\ln(A), b]$ pour tout $b > \log(A)$. Si on choisit $A = 10$ et $x_0 = b = 4$, on obtient pour $n = 8$ itérations, $x_n \approx 2.30258509299405$ et $\ln(A) \approx 2.30258509299405$.

Exemple Y.7 (Exemple d'application du théorème Y.4). Soit f un polynôme de degré $n \geq 3$, ayant n racines réelles distinctes et soit r, la plus grande racine de f. Le théorème (cas 1) s'applique sur tout intervalle $[r, b]$. En effet, d'après le théorème de Rolle, entre deux racines de f se trouve une racine de f'. Ainsi, f' a au moins $n-1$ racines, deux à deux distinctes. Puisque f' est polynomiale de degré $n-1$, elle a au plus $n-1$ racines. Ainsi, f' a exactement $n-1$ racines réelles, chacune d'elles étant située entre deux racines de f. De même, on montre que f'' a exactement $n-2$ racines réelles, chacune d'elles étant située entre deux racines de f'. Ainsi, si r désigne la plus grande racine de f, alors, pour tout $b > r$, f est de signe constant sur $[r, b]$. Notons $\varepsilon \in \{-1, 1\}$ ce signe. D'après ce qui précède, f'' est de signe constant sur $[r, +\infty[$. Or, puisque f'' y est du signe de son coefficient dominant et c'est aussi le signe du coefficient dominant de f, qui est donc $\varepsilon \in \{-1, 1\}$. De même, f' est du signe de ε sur $]r, +\infty[$ et puisque $f'(r) \neq 0$, f' est du signe de ε sur $[r, +\infty[$. Chacune des hypothèses (Y.3) est donc vérifiée.

Si on choisit f donnée par

$$f(x) = \prod_{k=0}^{N} (x - k),$$

dont toutes les racines simples sont k, pour $0 \leq k \leq N$ et donc dont la plus grande des racines est N, on obtient pour $N = 5$, $x_0 = 7$, pour $p = 10$ itérations, $x_p \approx 4.99999999999999$ et $f(x_p) \approx 2.13162\,10^{-12}$.

Annexe Z

Étude de $x^{x^{x^{x^{\cdots}}}}$ (sous la forme d'un problème corrigé)

Ce problème a été donné à l'examen à l'Automne 2019.

Énoncé

(1) On considère la fonction g définie par

$$\forall x \in \mathbb{R}, \quad g(x) = \sqrt{x}. \tag{Z.1}$$

et on pose

$$a = 2/5, \quad b = 1. \tag{Z.2}$$

(a) Montrer que g a un unique point fixe r sur $[a, b]$ et que la méthode du point fixe est convergente pour tout point de $x_0 \in [a, b]$ vers $r = 1$.

(b) Soit $(x_n)_{n\in\mathbb{N}}$, la suite associée à la méthode du point fixe. Déterminer l'entier n à partir duquel $|x_n - r| \leq \varepsilon$ où $\varepsilon = 1.\,10^{-1}$.

(c) Calculer les termes de la suite correspondant en choisissant $x_0 = 1/2$.

(2) On considère la fonction g définie par

$$\forall x \in \mathbb{R}, \quad g(x) = (1/2)^x. \tag{Z.3}$$

et on pose

$$a = 0, \quad b = 1. \tag{Z.4}$$

(a) Montrer que g a un unique point fixe r sur $[a, b]$ et que la méthode du point fixe est convergente pour tout point de $x_0 \in [a, b]$ vers $r = \frac{LambertW(\ln(2))}{\ln(2)}$.

(b) Soit $(x_n)_{n\in\mathbb{N}}$, la suite associée à la méthode du point fixe. Déterminer l'entier n à partir duquel $|x_n - r| \leq \varepsilon$ où $\varepsilon = 1.\,10^{-1}$.

(c) Calculer les termes de la suite correspondant en choisissant $x_0 = 1/2$.

(3) *Les questions suivantes sont facultatives*

On cherche à donner un sens à l'expression

$$y = x^{x^{x^{x^{\cdots}}}}, \tag{Z.5}$$

où la puissance est prise "un nombre infini de fois".

Cette notation est ambigüe puisqu'elle peut correspondre aux deux définitions suivantes : on prend la puissance à chaque fois par "au-dessus" :

$$y = \left(\left(\left((x^x)^x\right)^x\right)^x\right)^{x^{\cdots}}, \tag{Z.6}$$

ou, au contraire, par "en-dessous"

$$y = \cdot^{\cdot^{\cdot^{x^{\left(x^{\left(x^{\left(x^{x}\right)}\right)}\right)}}}}. \tag{Z.7}$$

Dans le premier cas, on définira donc une suite, à x fixé, définie par :

$$h_0(x) = x, \tag{Z.8a}$$
$$h_1(x) = x^x = (h_0(x))^x, \tag{Z.8b}$$
$$h_2(x) = (x^x)^x = (h_1(x))^x, \tag{Z.8c}$$

et ainsi de suite ... On a donc la relation de récurrence suivante :

$$\forall x, \quad \forall n \in \mathbb{N}, \quad h_{n+1}(x) = (h_n(x))^x, \tag{Z.9a}$$

avec l'initialisation suivante

$$\forall x, \quad h_0(x) = x. \tag{Z.9b}$$

Dans le second cas, on définira donc une suite, à x fixé, définie par :

$$f_0(x) = x, \tag{Z.10a}$$
$$f_1(x) = x^x = x^{f_0(x)}, \tag{Z.10b}$$
$$f_2(x) = x^{(x^x)} = x^{f_1(x)}, \tag{Z.10c}$$
$$f_3(x) = x^{\left(x^{(x^x)}\right)} = x^{f_2(x)}, \tag{Z.10d}$$

et ainsi de suite ... On a donc la relation de récurrence suivante :

$$\forall x, \quad \forall n \in \mathbb{N}, \quad f_{n+1}(x) = x^{f_n(x)}, \tag{Z.11a}$$

avec l'initialisation suivante

$$\forall x, \quad f_0(x) = x. \tag{Z.11b}$$

(a) Pour x fixé, on pose

$$\forall n \in \mathbb{N}, \quad u_n = h_n(x). \tag{Z.12}$$

(i)(A) Montrer que la suite $(u_n)_{n\in\mathbb{N}}$ est définie par

$$\begin{cases} \forall n \in \mathbb{N}, \quad u_{n+1} = G_x(u_n), \\ u_0 \text{ est donné.} \end{cases} \tag{Z.13}$$

On précisera (à x fixé) u_0 et la fonction (de $\mathbb{R}$ dans $\mathbb{R}$) G_x.

(B) Étudier la fonction G_x. On montrera notamment que G_x laisse stable les intervalles $[0, 1]$ et $]1, +\infty[$.

(C) Quels sont les points fixes de G_x ?

(D) En déduire les limites possibles de la suite $(u_n)_{n\in\mathbb{N}}$.

(E) Montrer que si $x \in [0, 1]$, la suite u_n tend vers 1 et que si $x \in]1, +\infty[$, la suite u_n tend vers $+\infty$.

(F) Quel est le lien avec la question 1 ?

(G) En déduire finalement l'expression de y défini par (Z.6).

(ii) Les résultats de la question 3(a)i, peuvent être en fait établis beaucoup plus rapidement !

(A) Montrer que l'on a

$$\forall x \in \mathbb{R}_+, \quad \forall n \in \mathbb{N}, \quad h_n(x) = x^{(x^n)}. \tag{Z.14}$$

(B) Conclure alors sur la convergence de la suite $h_n(x)$.

(b) Pour x fixé, on pose

$$\forall n \in \mathbb{N}, \quad v_n = f_n(x). \tag{Z.15}$$

(i) Montrer que la suite $(v_n)_{n\in\mathbb{N}}$ est définie par

$$\begin{cases} \forall n \in \mathbb{N}, \quad u_{n+1} = G_x(v_n), \\ u_0 \text{ est donné.} \end{cases} \tag{Z.16}$$

On précisera (à x fixé) u_0 et la fonction (de $\mathbb{R}$ dans $\mathbb{R}$) G_x.

(ii) Étudier la fonction G_x.

(iii) Quels sont les points fixes de G_x ?

(iv) En déduire les limites possibles de la suite $(v_n)_{n\in\mathbb{N}}$.

(v) Étudier la convergence de la suite v_n.

(vi) Quel est le lien avec la question 2 ?

(vii) En déduire finalement l'expression de y défini par (Z.7).

Corrigé

Sur cette question, on pourra le papier très complet [Cam10], disponible à l'url `http://citron.9grid.fr/docs/tetration.pdf` ou les pages suivantes : `http://mathvault.ca/derivative-tetration-hyperexponentiation/`, `http://en.wikipedia.org/wiki/Tetration` ou `https://fr.wikipedia.org/wiki/T%C3%A9tration`.

Il existe d'autres pages moins détaillées comme

`http://math.stackexchange.com/questions/1634746/solving-equation-with-infinite-exponent-tower`
`http://mathforum.org/library/drmath/view/70270.html`
`http://math.stackexchange.com/questions/1318481/is-this-a-valid-proof-for-xxxxx-dots-y?noredirect=1&lq=1`
`https://math.stackexchange.com/questions/1317314/derivative-of-xxx-to-infinity`

Attention, cette dernière est erronnée puis reprise !

Plus de détails plus fins pourront être trouvés dans [Kno81].

On pourra télécharger le zip de fonction matlab à l'adresse `http://utbmjb.chez-alice.fr/Polytech/MNBif/examcorMNBifA19.zip` qui contient des fonctions matlab à faire tourner pour la résolution des questions 3a et 3b.

(1)

(a) (i)(A) On a

$$g'(x) = 1/2\,\frac{1}{\sqrt{x}}. \tag{Z.17}$$

(B) Sur la figure 1(a), on constate que la fonction g semble avoir un point fixe, correspondant à la valeur

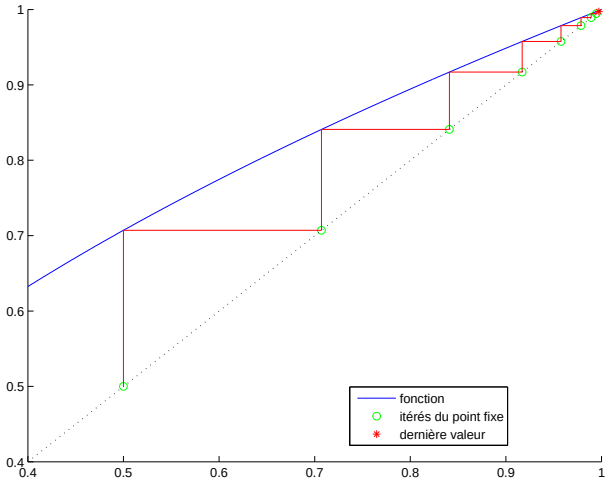
$$r = 1.$$

(C) Sur la figure 1(b), on constate que les valeurs de la fonction $|g'|$ sont inférieures à 0.790569. Démontrons cela rigoureusement. La dérivée de la fonction g est monotone et g' prend donc ses valeurs entre $g'(a)$ et $g'(b)$. Sa valeur maximale est donc donnée par

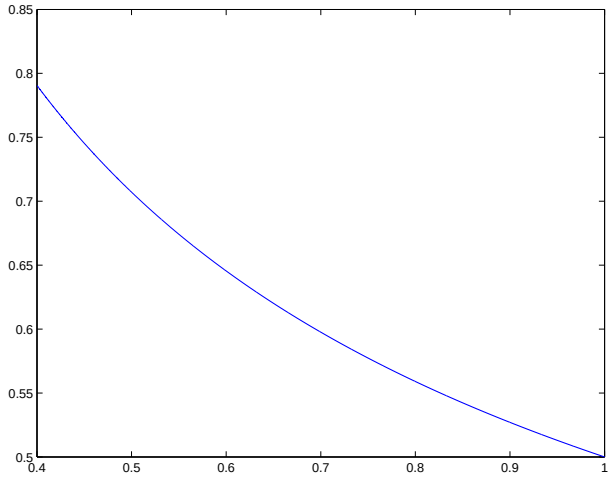
$$\alpha = 0.7905694150421$$

(D) Sur la figure 1(a), on constate que l'intervalle $[a, b]$ est g-stable.

Démontrons cela rigoureusement. La fonction g est croissante ; ainsi, sur l'intervalle $[a, b]$, elle prend les valeurs $[g(a), g(b)]$. On vérifie que $g(a) = 0.6324555320337$ et $g(b) = 1$ sont bien dans l'intervalle $[a, b]$.



(a) Le graphe de la fonction g et les premières valeurs de la suite x_n.



(b) Le graphe de la fonction $|g'|$.

FIGURE Z.1. Les graphes des fonctions g et $|g'|$.

(ii) D'après les points 1(a)iC et 1(a)iD, les deux hypothèses de la proposition 4.19 sont vérifiées et donc g admet un point fixe unique r dans $I = [a, b]$ et, pour tout x_0 de I, la suite (x_n) est définie et converge vers r. Cette valeur est nécessairement celle donnée dans l'énoncé, par unicité de celle-ci !

(b) Appliquons le résutat de la proposition 4.21 ; on choisit n définit par (4.45), où la valeur de k a été donnée plus haut, ce qui donne numériquement

$$n = 8. \tag{Z.18}$$

(c) On obtient alors progressivement :

$$\begin{aligned}
x_0 &= 0.5000000000000\ ;\\
x_1 &= g(x_0) = 0.7071067811865\ ;\\
x_2 &= g(x_1) = 0.8408964152537\ ;\\
x_3 &= g(x_2) = 0.9170040432047\ ;\\
x_4 &= g(x_3) = 0.9576032806986\ ;\\
x_5 &= g(x_4) = 0.9785720620877\ ;\\
x_6 &= g(x_5) = 0.9892280131940\ ;\\
x_7 &= g(x_6) = 0.9945994234836\ ;\\
x_8 &= g(x_7) = 0.9972960560855.
\end{aligned}$$

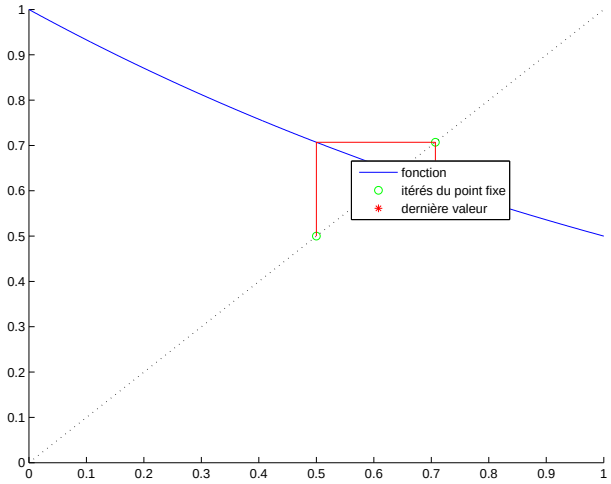
REMARQUE Z.1. Si on calcule l'erreur réellement commise, en utilisant la valeur de x_n déterminée ci-dessous et la valeur de r donnée dans l'énoncé, on a

$$|x_n - r| = |0.9972960560855 - 1| = 0.0027039439145,$$

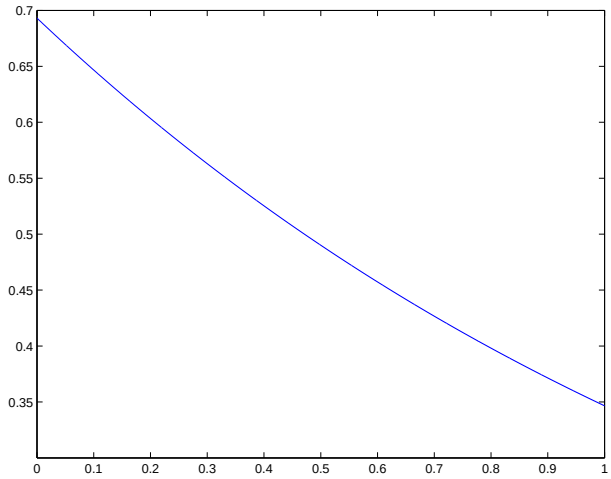
ce qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

REMARQUE Z.2. Si on utilise la majoration donnée par (Q.13), on obtient

$$|x_n - r| \leq 0.0101793883594,$$



(a) Le graphe de la fonction g et les premières valeurs de la suite x_n.



(b) Le graphe de la fonction $|g'|$.

FIGURE Z.2. Les graphes des fonctions g et $|g'|$.

qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

(2)

(a) (i)(A) On a

$$g'(x) = -(1/2)^x \ln(2). \tag{Z.19}$$

(B) Sur la figure 2(a), on constate que la fonction g semble avoir un point fixe, correspondant à la valeur

$$r = \frac{LambertW(\ln(2))}{\ln(2)}.$$

(C) Sur la figure 2(b), on constate que les valeurs de la fonction $|g'|$ sont inférieures à 0.693147. Démontrons cela rigoureusement. La dérivée de la fonction g est monotone et g' prend donc ses valeurs entre $g'(a)$ et $g'(b)$. Sa valeur maximale est donc donnée par

$$\alpha = 0.6931471805599$$

(D) Sur la figure 2(a), on constate que l'intervalle $[a, b]$ est g-stable.
Démontrons cela rigoureusement. La fonction g est croissante ; ainsi, sur l'intervalle $[a, b]$, elle prend les valeurs $[g(a), g(b)]$. On vérifie que $g(a) = 1$ et $g(b) = 0.5000000000000$ sont bien dans l'intervalle $[a, b]$.

(ii) D'après les points 2(a)iC et 2(a)iD, les deux hypothèses de la proposition 4.19 sont vérifiées et donc g admet un point fixe unique r dans $I = [a, b]$ et, pour tout x_0 de I, la suite (x_n) est définie et converge vers r. Cette valeur est nécessairement celle donnée dans l'énoncé, par unicité de celle-ci !

(b) Appliquons le résutat de la proposition 4.21 ; on choisit n définit par (4.45), où la valeur de k a été donnée plus haut, ce qui donne numériquement

$$n = 7. \tag{Z.20}$$

(c) On obtient alors progressivement :

$$\begin{aligned}
x_0 &= 0.5000000000000\,;\\
x_1 &= g(x_0) = 0.7071067811865\,;\\
x_2 &= g(x_1) = 0.6125473265361\,;\\
x_3 &= g(x_2) = 0.6540408600421\,;\\
x_4 &= g(x_3) = 0.6354978458134\,;\\
x_5 &= g(x_4) = 0.6437186417229\,;\\
x_6 &= g(x_5) = 0.6400610211772\,;\\
x_7 &= g(x_6) = 0.6416858070430.
\end{aligned}$$

REMARQUE Z.3. Si on calcule l'erreur réellement commise, en utilisant la valeur de x_n déterminée ci-dessous et la valeur de r donnée dans l'énoncé, on a

$$|x_n - r| = \left| 0.6416858070430 - \frac{LambertW\,(\ln(2))}{\ln(2)} \right| = 0.0005000625380,$$

ce qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

REMARQUE Z.4. Si on utilise la majoration donnée par (Q.13), on obtient

$$|x_n - r| \leq 0.0036702147431,$$

qui est bien inférieur à la valeur de ε donnée dans l'énoncé.

(3) (a) On pourra faire tourner la fonction matlab `etudefonctionh.m` pour obtenir les différentes courbes des corrigés de cette question.

Pour x fixé, on pose

$$\forall n \in \mathbb{N}, \quad u_n = h_n(x). \tag{Z.21}$$

(i)(A) La suite $(u_n)_{n\in\mathbb{N}}$ est donc définie par

$$\begin{cases} \forall n \in \mathbb{N}, \quad u_{n+1} = G_x(u_n), \\ u_0 \text{ est donné.} \end{cases} \tag{Z.22}$$

avec

$$\forall y, \quad G_x(y) = y^x = e^{x \ln y}, \tag{Z.23a}$$

$$u_0 = x. \tag{Z.23b}$$

(B) À x fixé, la fonction G_x est définie sur $\mathbb{R}_+^*$. Puisque $u_1 = G_x(x)$, il est nécessaire que

$$x \in \mathbb{R}_+^*, \tag{Z.24}$$

hypothèse que l'on fera pour toute la suite. Étudions donc G_x sur $\mathbb{R}_+^*$, qui y est de classe $\mathcal{C}^\infty$.

On a

$$\forall y \in \mathbb{R}_+^*, \quad G_x'(y) = \frac{x}{y} e^{x \ln y}, \tag{Z.25}$$

qui est strictement positive. De plus, on a

$$\lim_{y \to 0^+} G_x(y) = 0,$$

$$\lim_{y \to +\infty} G_x(y) = +\infty.$$

On peut donc prolonger G_x sur $\mathbb{R}_+$ en posant

$$\forall x \in \mathbb{R}_+^*, \quad G_x(0) = 0. \tag{Z.26}$$

y	0	1	$+\infty$
Variations de G_x	0 ↗	1 ↘	↗ $+\infty$
Signe de G'_x	$+$		

TABLE Z.1. Tableau de variation de G_x

Voir le tableau de variation Z.1. Puisque G_x est strictement croissante, on déduit de ce tableau que

$$G_x(]0,1[) =]0,1[, \tag{Z.27a}$$
$$G_x([0,1]) = [0,1], \tag{Z.27b}$$
$$G_x(]1,+\infty[) =]1,+\infty[, \tag{Z.27c}$$

et donc en particulier que

La fonction G_x laisse stable chacun des intervalles $]0,1[$, $[0,1]$ et $]1,+\infty[$. (Z.28)

(C) Pour $x \geq 0$, résolvons l'équation, en y, sur $\mathbb{R}_+$:

$$G_x(y) = y. \tag{Z.29}$$

Si $x > 0$, c'est donc équivalent à

$$e^{x\ln(y)} = y \tag{Z.30}$$

ce qui est équivalent (car $y > 0$) à

$$x\ln(y) = \ln(y). \tag{Z.31}$$

Si $y \neq 1$, cela implique $x = 1$. Donc, par contraposition, $x \neq 1$ implique $y = 1$. Dans ce cas (Z.29) est vérifée. Si $y = 1$, cela est vrai pour tout valeur de x.

Nous reviendrons plus tard sur le cas $x = 0$.

pour tout $x \in \mathbb{R}_+^* \setminus \{1\}$, l'unique point fixe de G_x est $y = 1$; (Z.32a)

si $x = 1$, tout réel $y > 0$ est point fixe de G_y. (Z.32b)

(D) • Si $x = 1$, d'après (Z.22), on a

$$\forall n \in \mathbb{N}, \quad u_{n+1} = G_1(u_n) = u_n \text{ et } u_0 = x = 1,$$

on a donc

$$\text{pour } x = 1, \ \lim_{n\to+\infty} u_n = 1. \tag{Z.33}$$

- Si $x \neq 1$, d'après (Z.22) et (Z.32a) et la continuité de G_x, si u_n converge,

$$\text{la seule limite possible de } u_n \text{ est } 1. \tag{Z.34}$$

- Nous reviendrons plus tard sur le cas $x = 0$.

(E) Étudions maintenant la convergence effective de la suite u_n.

Nous avons deux méthodes.

- Le cas $x = 1$ est déjà traité (voir (Z.33)) .
  Nous raisonnons à la main, en étudiant la fonction $y \mapsto G_x(y)/y$ définie sur $\mathbb{R}_+^*$. Supposons donc $x \in \mathbb{R}_+^* \setminus \{1\}$. On a,

$$\forall y \in \mathbb{R}_+^*, \quad \frac{G_x(y)}{y} = e^{x\ln(y)-\ln(y)} = e^{\ln(y)(x-1)} > 0,$$

  et donc

$$\forall (x,y) \in (]0,1[\times]0,1[)^2 \cup (]1,+\infty[\times]1,+\infty[)^2, \quad G_x(y) > y. \tag{Z.35}$$

  De cela, de (Z.28) et de (Z.23), nous obtenons par une récurrence immédiate que,

$$\forall x > 1, \quad u_n > 1 \text{ et } u_{n+1} > u_n, \tag{Z.36a}$$
$$\forall x < 1, \quad 0 < u_n < 1 \text{ et } \mathrm{u}_{n+1} > u_n. \tag{Z.36b}$$

  Dans le second cas, la suite u_n est croissante et majorée et converge donc vers $l > 0$. D'après (Z.32a), cette limite l ne peut être que 1. On a donc

$$\text{pour } x \in]0,1[, \ \lim_{n\to+\infty} u_n = 1. \tag{Z.37}$$

  Dans le premier cas, la suite u_n est croissante et ne peut pas être majorée. Si c'était le cas, elle convergerait vers $l > 1$, qui serait point fixe de G_x sur $]1,+\infty[$ ce qui est impossible d'après (Z.32a). On a donc

$$\text{pour } x \in]1,+\infty[, \ \lim_{n\to+\infty} u_n = +\infty. \tag{Z.38}$$

  Concluons par le cas le cas $x = 0$. La suite u_n n'est pas définie en toute rigueur dans ce cas. Conformément à matlab ou à la propriété

$$\lim_{x\to 0^+} x^x = \lim_{x\to 0^+} e^{x\ln(x)} = 1,$$

  on peut poser

$$0^0 = 1. \tag{Z.39}$$

  Ainsi, par définition u_n vaut 1 et

$$\text{pour } x = 0, \ \lim_{n\to+\infty} u_n = 1. \tag{Z.40}$$

- On peut aussi utiliser la proposition 4.19 pour montrer la convergence de la suite u_n pour $x \in]0,1[$.
  D'après (Z.25), on a

$$\forall y \in]0,1], \quad G'_x(y) = \frac{x}{y} e^{x\ln y} = x e^{-\ln(y)} e^{x\ln y},$$

  et donc

$$\forall y \in]0,1], \quad G'_x(y) = x e^{\ln(y)(x-1)}. \tag{Z.41}$$

Puisque $x \in]0,1[$, la fonction $y \mapsto \ln(y)(x-1)$ est décroissante sur $]0,1]$ et donc G'_x est décroissante sur $]0,1]$. Compte tenu de

$$\lim_{y\to 0^+} G'_x(y) = +\infty, \tag{Z.42a}$$

$$G'_x(1) = 0. \tag{Z.42b}$$

on peut obtenir le tableau de variation de G'_x.

y	0	ξ_x	1
Variations de G'_x	$+\infty$ ↘	1	↘ 0

TABLE Z.2. Tableau de variation de G'_x sur $[0,1]$.

Voir le tableau Z.2. Il nous montre qu'il existe un unique réel noté ξ_x tel que $G'_x(\xi_x) = 1$. Il est donné par

$$x e^{\ln(\xi_x)(x-1)} = 1,$$

ce qui est successivement équivalent à

$$\begin{aligned} e^{\ln(\xi_x)(x-1)} = \frac{1}{x} &\iff \ln(\xi_x)(x-1) = -\ln(x), \\ &\iff \ln(\xi_x) = \frac{\ln(x)}{1-x}, \\ &\iff \xi_x = e^{\frac{\ln(x)}{1-x}} \end{aligned}$$

et donc $\xi_x = H(x)$ où

$$\forall x \in]0,1[, \quad H(x) = e^{\frac{\ln(x)}{1-x}}. \tag{Z.43}$$

On a donc montré que

$$\forall x \in]0,1[, \quad \forall y \in [H(x),1], \quad 0 \le G'_x(y) \le 1. \tag{Z.44}$$

Remarquons aussi que G_x laisse l'intervalle $]H(x),1[$ stable. En effet, compte tenu de la monotonie de G_y, c'est équivalent à

$$\forall x \in]0,1[, \quad G_x(H(x)) > H(x). \tag{Z.45}$$

Cette propriété est admise.
On consultera la figure Z.3 pour s'en convaincre.
Si $x_0 \in [H(x),1]$, la suite u_n est donc à valeur dans $]H(x),1[$. Puisqu'elle est strictement croissante, elle est dans un intervalle du type $[H(x)+\eta,1]$ où $\eta > 0$. D'après la monotonie de G'_x et (Z.44), on a donc

$$\forall y \in [H(x)+\eta,1], \quad 0 \le G'_x(y) \le k \text{ où } k \in [0,1[. \tag{Z.46}$$

Ainsi, la proposition 4.19 s'appliquent : la convergence de la suite u_n pour $x \in]0,1[$ a lieu vers l'unique point fixe de G_x qui est donc ici égal à 1.

(F) Nous avons déjà montré cette convergence dans le cas particulier de la question 1.

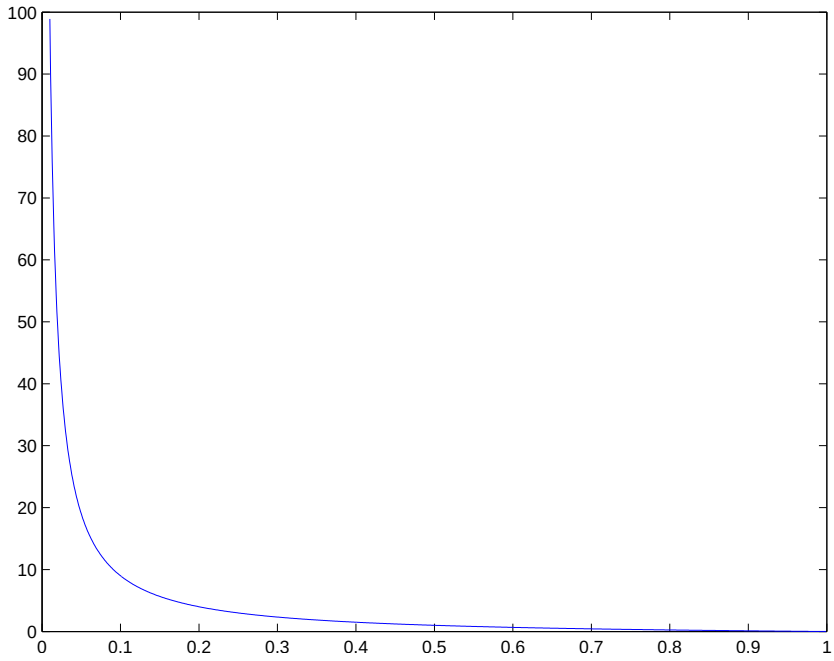


FIGURE Z.3. La fonction $g_x(H(x))/H(x) - 1$ sur $]0,1[$.

(G) Considérons h définie par

$$\forall y \in \mathbb{R}_+, \quad h(y) = \begin{cases} 1, & \text{si } y \in [0,1], \\ +\infty, & \text{si } y \in]1,+\infty[. \end{cases} \tag{Z.47}$$

D'après la définition de h_n, (Z.33), (Z.37), (Z.38) et (Z.40), on a donc démontré que,

$$\forall y \in \mathbb{R}_+, \quad \left(\left(\left((x^x)^x\right)^x\right)^x\right)^{x^{\cdot^{\cdot^{\cdot}}}} = \begin{cases} 1, & \text{si } y \in [0,1], \\ +\infty, & \text{si } y \in]1,+\infty[, \end{cases} \tag{Z.48}$$

l'expression

$$\left(\left(\left((x^x)^x\right)^x\right)^x\right)^{x^{\cdot^{\cdot^{\cdot}}}}$$

étant à prendre comme la limite quand n tend vers l'infini de

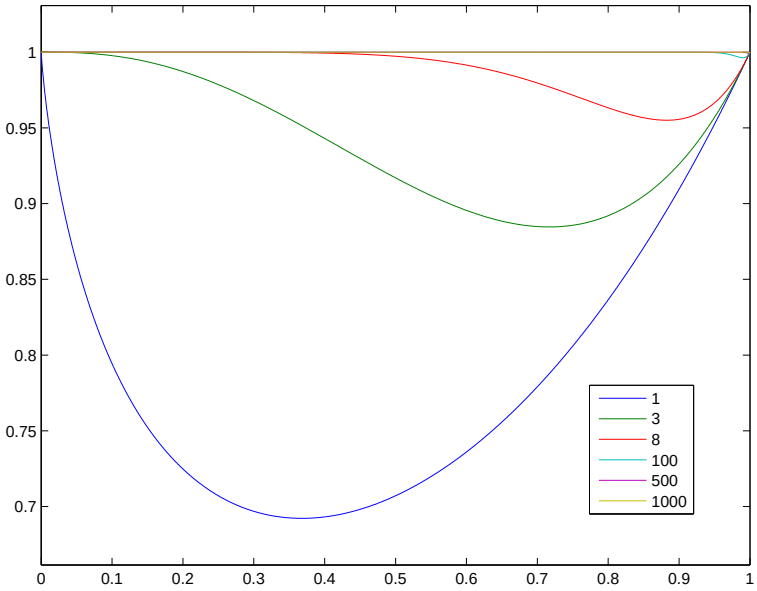
$$\left(\left(\left(\left((x^x)^x\right)^x\right)^x\right)^x\right)^x,$$

où la puissance est "prise n fois" (par "dessus").

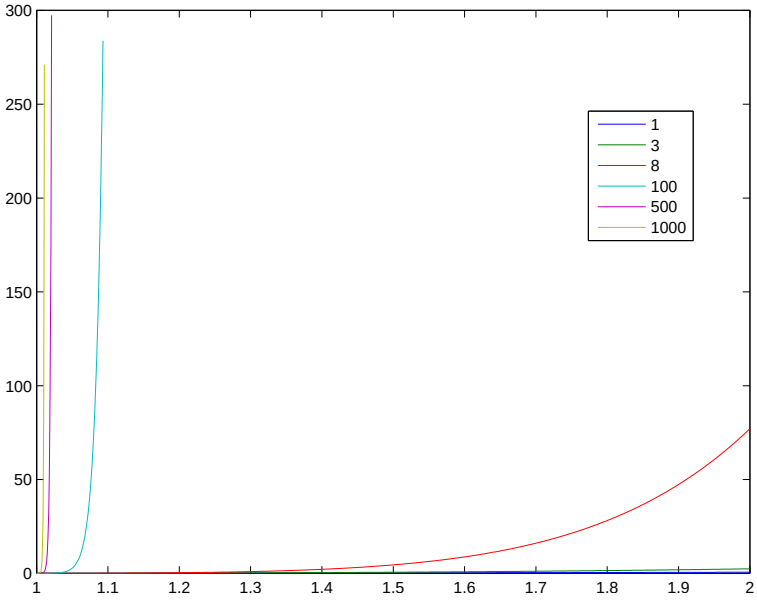
Cette propriété est corroborée par le graphique Z.4.

REMARQUE Z.5. On peut montrer que la convergence de la suite de fonction h_n vers 1 est uniforme sur tout intervalle $[0,\alpha]$ où $\alpha \in [0,1[$. La convergence uniforme sur $[0,1]$ n'a pas été étudiée avec cette méthode. Voir le corrigé de la question 3(a)ii et notamment la remarque Z.6.

(ii) La méthode de la question 3(a)i était surtout intéressante d'un point de vue didactique, puisqu'elle reposait sur l'utilisation du théorème du point fixe. Il est en fait beaucoup plus pertinent d'utiliser la méthode suivante :



(a) : différentes fonctions h_n sur l'intervalle $(0, 1)$.



(b) : différentes fonctions $\ln_{10}(h_n)$ sur l'intervalle $(1, 2)$.

FIGURE Z.4. Le comportement de h_n quand n grandit.

(A) Montrons que l'on a

$$\forall x \in \mathbb{R}_+, \quad \forall n \in \mathbb{N}, \quad h_n(x) = x^{(x^n)}. \tag{Z.49}$$

Il suffit en effet de rappeler que

$$\forall a, b, c \in \mathbb{R}_+^*, \quad \left(a^b\right)^c = a^{bc}. \tag{Z.50}$$

En effet, par définition,

$$\left(a^b\right)^c = e^{c\ln(a^b)} = e^{c\ln\left(e^{b\ln a}\right)} = e^{cb\ln a} = a^{bc}.$$

Ainsi, on peut montrer (Z.49), par récurrence sur n. Avec la convention (Z.39), pour $n = 0$, on a bien $h_0(x) = x$. Supposons (Z.49) vraie pour n et montrons-là pour $n+1$. Par défintion, on a

$$h_{n+1}(x) = (h_n(x))^x,$$

et donc, d'après l'hypothèse de récurrence et en utilisant (Z.50), on a

$$h_{n+1}(x) = \left(x^{(x^n)}\right)^x = x^{(x\times x^n)} = x^{\left(x^{n+1}\right)},$$

ce qui permet de conclure.

On peut aussi écrire (Z.49) sous la forme

$$\forall x \in \mathbb{R}_+, \quad \forall n \in \mathbb{N}, \quad h_n(x) = e^{x^n \ln x}. \tag{Z.51}$$

(B) Concluons alors sur la convergence de la suite $h_n(x)$.

- Si $x > 1$, x^n tend vers $+\infty$ quand n tend vers l'infini et donc, d'après (Z.51), on a
$$\lim_{n+\infty} h_n(x) = +\infty. \tag{Z.52}$$
- Si $x = 0$ ou $x = 1$, avec la convention (Z.39), on a
$$\lim_{n+\infty} h_n(x) = 1. \tag{Z.53}$$
- Enfin, si $x \in]0, 1[$, x^n tend vers 0 quand n tend vers l'infini et donc, d'après (Z.51), on a
$$\lim_{n+\infty} h_n(x) = 1. \tag{Z.54}$$

Remarque Z.6. On pourrait montrer, grâce à l'expression (Z.51), la convergence uniforme de la fonction h_n vers la fonction constante égale à 1 sur $[0, 1]$.

(b) On pourra faire tourner la fonction matlab `etudefonctionf.m` pour obtenir les différentes courbes des corrigés de cette question.

Correction en cours de rédaction

Remarque Z.7. Dans `https://math.stackexchange.com/questions/1317314/derivative-of-xxx-to-infi` est établie la propriété suivante : Si y est dérivable, on a

$$\forall x \in \left[e^{-e}, e^{1/e}, \right[, \quad y'(x) = \frac{y^2(x)}{x(1 - y(x)\ln(x))}, \tag{Z.55a}$$

$$y(1) = 1. \tag{Z.55b}$$

Remarque Z.8. $f_n(x)$ est parfois aussi noté ${}^n x$ ou $x \uparrow\uparrow n$.

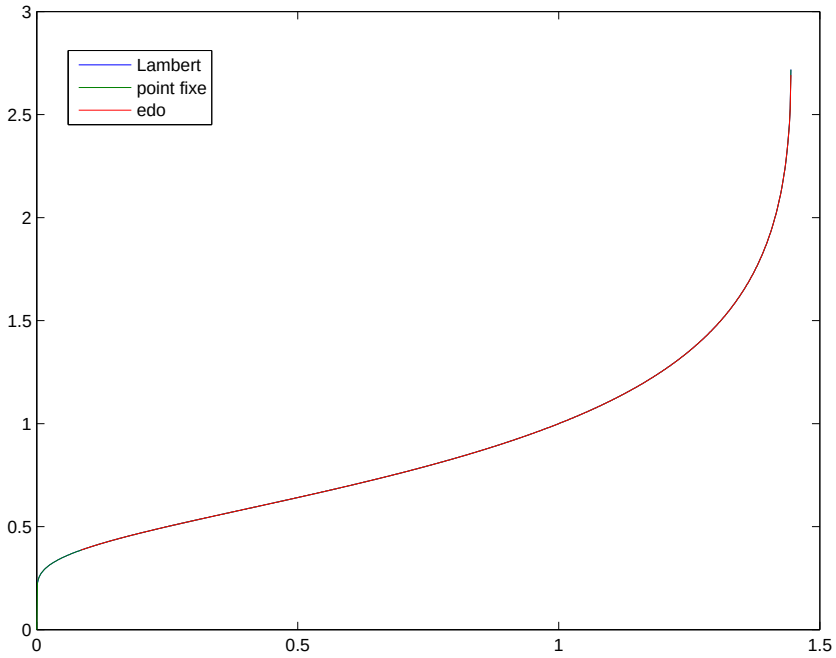


FIGURE Z.5. La fonction obtenue par résolution d'edo, recherche de point fixe ou en utilisant la fonction de Lambert sur $[0, e^{1/e}]$.

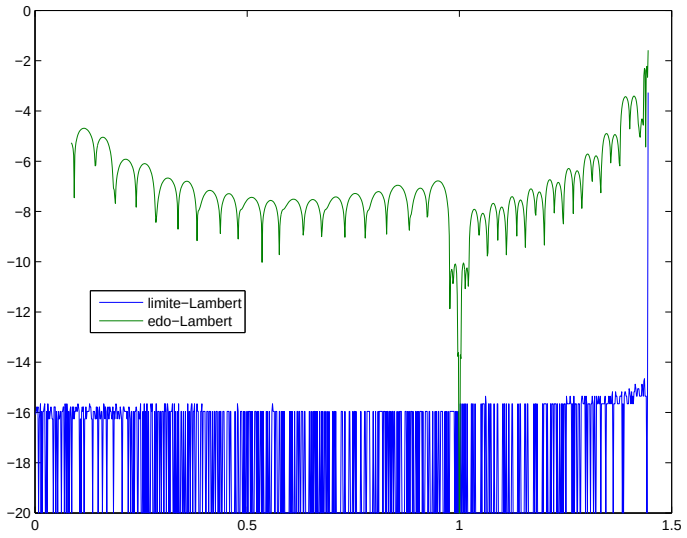


FIGURE Z.6. Logarithme décimal de l'écarts entre les trois modes de calculs de la figure Z.5.

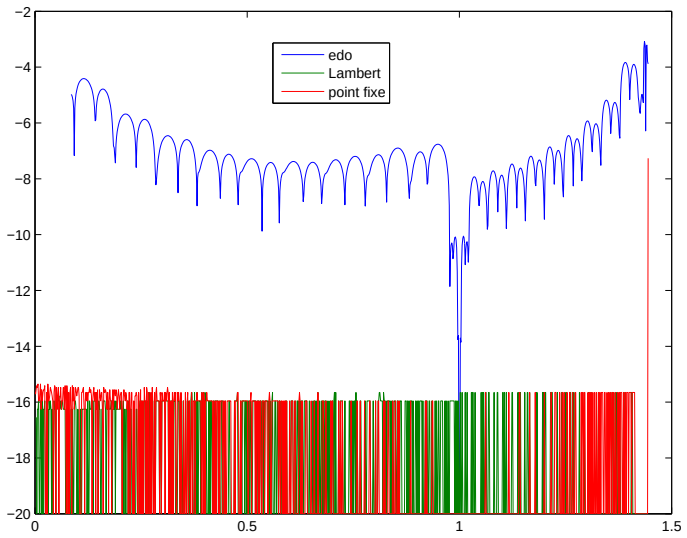


FIGURE Z.7. Logarithme décimal de l'écarts entre $f(x)$ et $x^{f(x)}$ pour les trois modes de calculs de la figure Z.5.

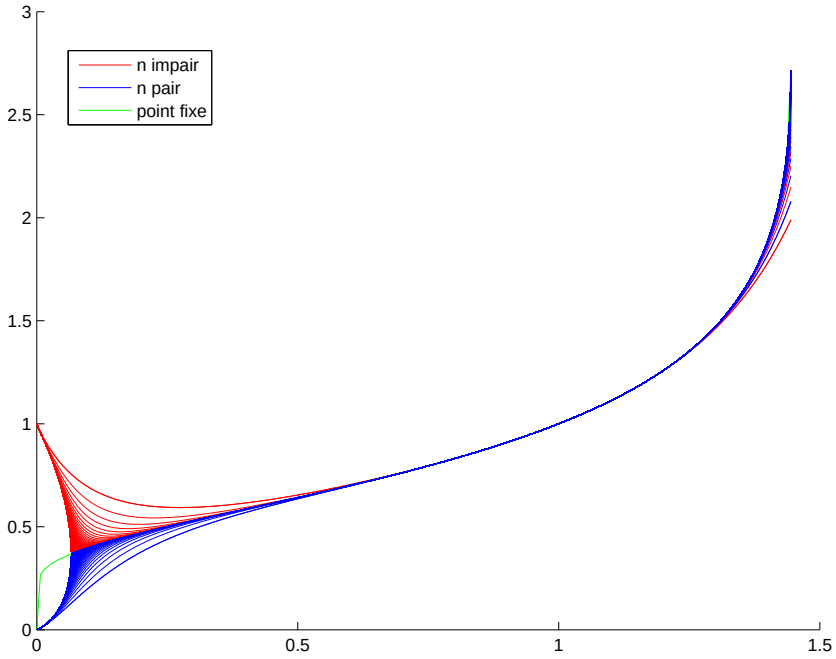


FIGURE Z.8. Plusieurs fonctions f_n sur $[0, e^{1/e}]$.

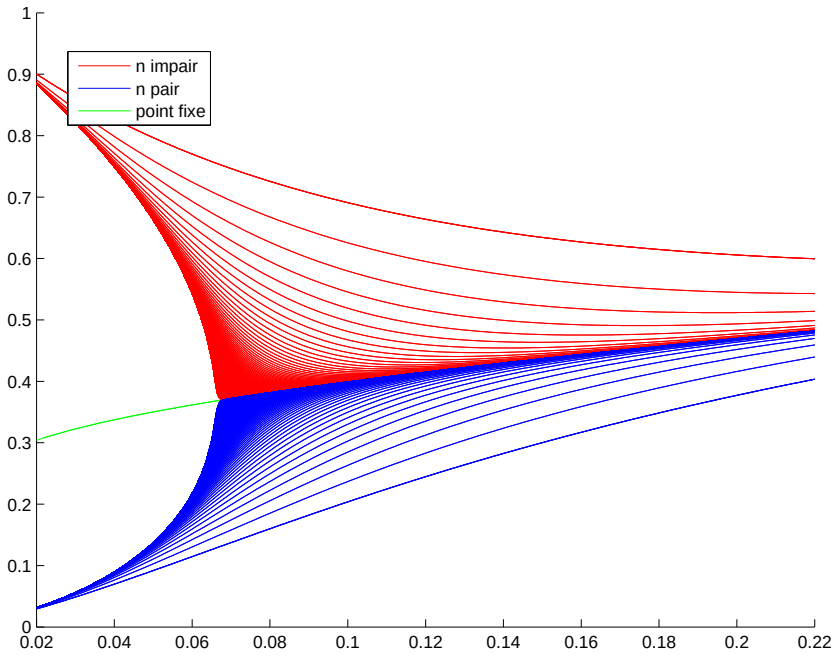


FIGURE Z.9. Plusieurs fonctions f_n sur $[0.02, 0.22]$.

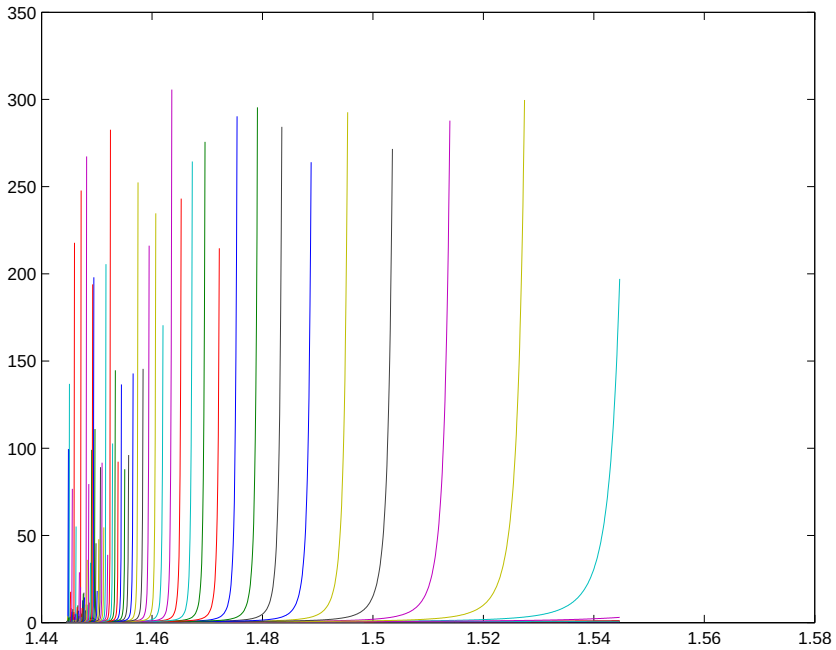


FIGURE Z.10. Plusieurs fonctions $\ln_{10}(f_n)$ sur $[e^{1/e}, e^{1/e} + 0.1]$.

Annexe AA

Approximations de π

Cette annexe est issue de [DB21].

AA.1. Introduction

Il existe de très nombreux moyens d'approcher la valeur de π. On pourra consulter les pages suivantes : `https://fr.wikipedia.org/wiki/Approximation_de_%CF%80`, `http://serge.mehl.free.fr/anx/iso_perim.html`, `https://melusine.eu.org/syracuse/bc/archimede/doc-a4.pdf` et surtout ces dernières [1] qui ont inspiré ce texte : `http://www.pi314.net/fr/archi.php`, `http://www.pi314.net/fr/cues.php` et `http://serge.mehl.free.fr/chrono/Cusa.html`. Les deux méthodes constituent les plus géométriques et les plus intuitives (et naturellement possibles à la règle et du compas) pour approcher π.

AA.2. Méthode d'Archimède

AA.2.1. Principe

L'idée est de construire deux suites de polygone réguliers de même nombre de sommets tels qu'un cercle donnés soit circonscrit à chacun des polygone de la première suite (le cercle passe par tous les sommets) et ce cercle soit inscrit dans chacun des polygone de la seconde suite (le cercle est tangent à tous les cotés de ce polygone). Par la suite, pour simplifier les polygones auxquels est cisconcrit le cercle seront dits inscrits dans le cercle et les polygones dans lequel le cercle est inscrits seront dits circonscrits au cercle. Chacune des deux suite de polygone se rapprochera du cercle donné de telle sorte que les polygones aient des périmètres (faciles à calculer) qui se rapprochent par défaut et par excès du périmètre du cercle. Au début, on considère deux héxagones réguliers inscrit et circonscrit au cercle. Ensuite à chaque étape, on double le nombre de sommet de chacun des deux polygones en construisant deux nouveaux polygones qui soient de nouveau respectivement inscrits et circonscrit au cercle.

AA.2.2. Mise en évidence géométrique

Archimède a tout d'abord construit un hexagone régulier inscrit dans un cercle de rayon $1/2$ et un hexagone régulier circonscrit au cercle, constructions qui peuvent être faites à la règle et au compas. Voir figure AA.1. Le côté de l'hexagone inscrit est égal à $DC = 1/2$ (rayon du cercle) tandis que le côté de celui qui est circonscrit vaut $AB = \sqrt{3}/3$. En effet, d'une part, le triangle OAB est équilatéral et sa hauteur OI (voir figure AA.1) est égal à $1/2$, rayon du cercle. Ainsi, $OC = 1/2$ et puisque le triangle OCD est équilatéral, on a bien $DC = 1/2$ D'autre part, le théorème de Pythagore appliqué au triangle rectangle OIA donne, puisque I est le milieu de $[AB]$

$$OA^2 = OI^2 + IA^2 = \frac{1}{4} + \frac{AB^2}{4} = \frac{1}{4} + \frac{OA^2}{4}$$

et donc

$$OA^2 = \frac{1}{3},$$

1. on pourra aussi consulter `http://www.pi314.net/fr/index.php`.

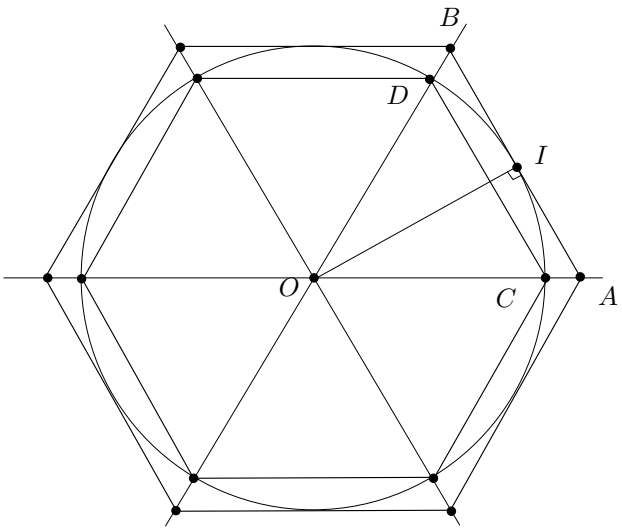


FIGURE AA.1. Hexagones réguliers inscrit et circonscrit au cercle de rayon 1/2.

et

$$OA = \frac{\sqrt{3}}{3},$$

et donc

$$AB = \frac{\sqrt{3}}{3}.$$

Les périmètres de ces deux hexagones, respectivement égaux à $6 \times 1/2 = 3$ et $6 \times \frac{\sqrt{3}}{3} = 2\sqrt{3}$ constituent une approximation par défaut et par excès du périmètre du cercle, égal ici à π :

$$3 \leq \pi \leq 2\sqrt{3} \approx 3,4641. \tag{AA.1}$$

Archimède a ensuite construit deux autres polygones réguliers, inscrit et circonscrit au cercle, en doublant le nombre de cotés, opération qui peut se faire à la règle et au compas. Les périmètres de ces deux polygones constituent un encadrement de π, et quand le nombre de coté grandit, ces deux polygones se confondent avec le cercle, ce qui fournit donc une approximation de π. Nous notons, à la n-ième étape I_n le périmètre du polygone inscrit dans le cercle et C_n le périmètre du polygone circonscrits au cercle.

AA.2.3. Construction à la règle et au compas

Sur les figures AA.2 à AA.4, sont indiquées les constructions de I_0, I_1 et I_2. Le lecteur pourra mesurer, sur chaque figure, le demi-périmètre donné, en utilisant la distance de 1/2 donnée et vérifier qu'il est proche de $\pi/2$ (ou de π en prenant 1 à la place de 1/2 !).

AA.2.4. Étude géométrique

Formons de façon géométrique l'expression des suites I_n et C_n qu'a proposées Archimède pour approcher π. Naturellement, le formalisme utilisé n'est pas l'original. À l'étape numéro $n \in \mathbb{N}$ du procédé, notons

U_n la longueur du côté du polygone régulier inscrit dans le cercle, (AA.2a)

V_n la longueur du côté du polygone régulier circonscrit au cercle. (AA.2b)

Voir les figures AA.5.

Ces deux polygones ont 6.2^n cotés. (AA.3)

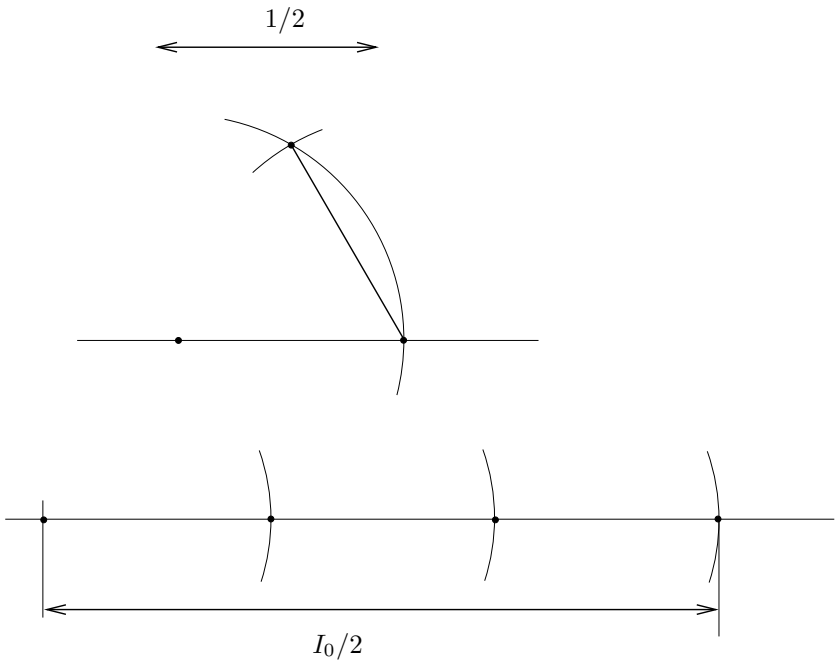


FIGURE AA.2. Construction à la règle et au compas de I_0.

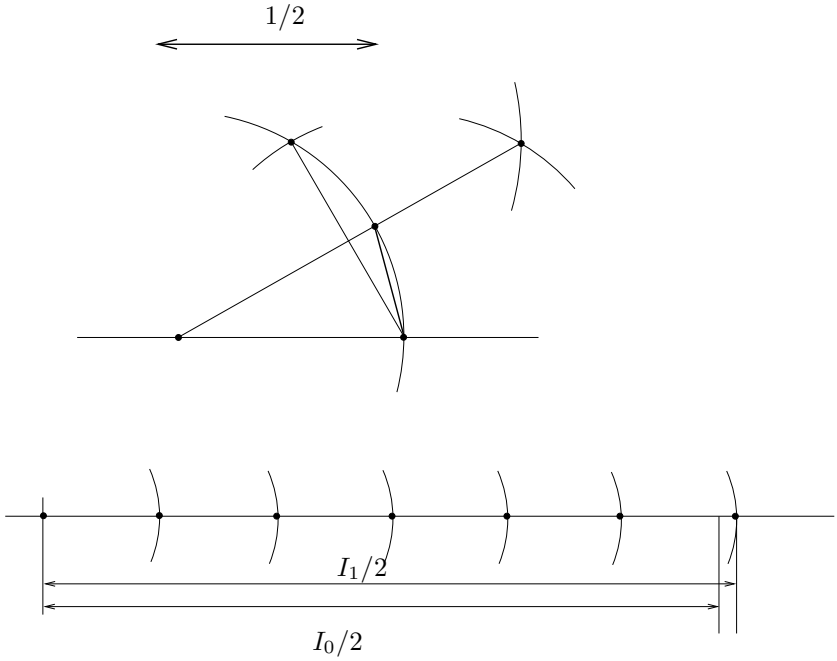


FIGURE AA.3. Construction à la règle et au compas de I_1.

D'après ce qui précède, on a

$$U_0 = \frac{1}{2}, \quad V_0 = \frac{\sqrt{3}}{3}. \tag{AA.4}$$

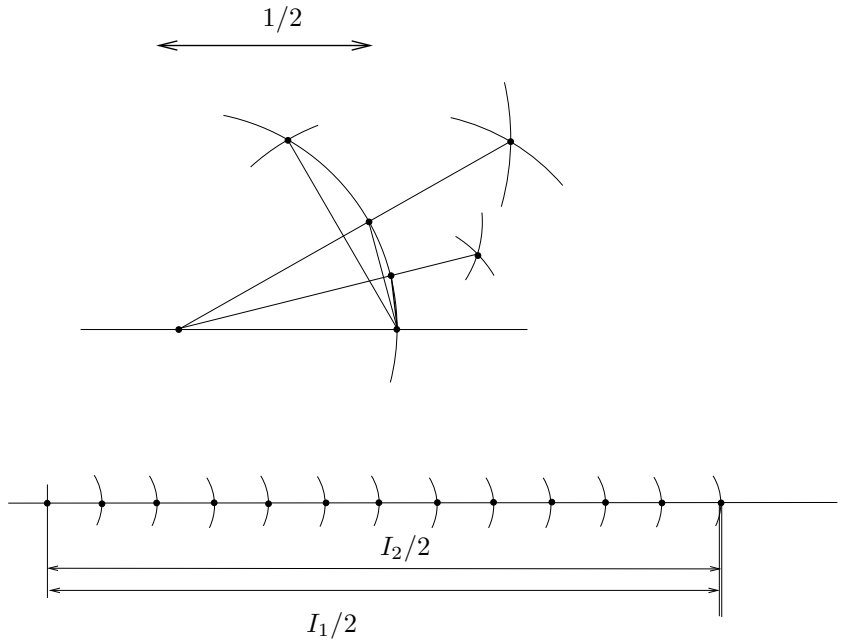
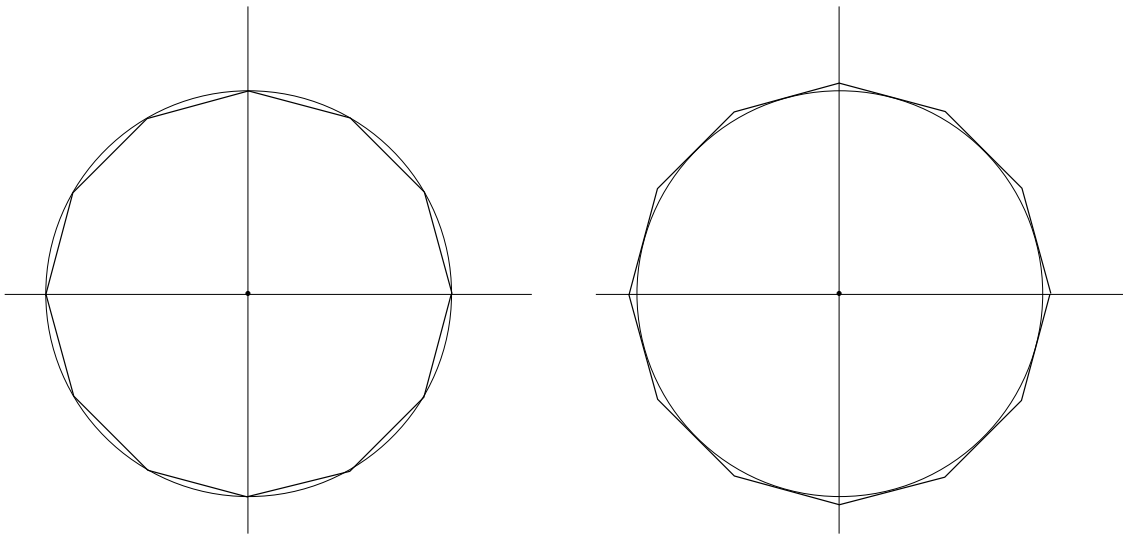


FIGURE AA.4. Construction à la règle et au compas de I_2.



(a) Le polygone inscrit (b) Le polygone circonscrit

FIGURE AA.5. Les polygones inscrit et circonscrit (sur ces figures, ils ont 12 cotés ($n = 1$)).

Contrairement à la page `http://www.pi314.net/fr/archi.php`, on peut trouver une relation de récurrence entre U_n et U_{n+1} et V_n et V_{n+1} directement, de façon géométrique, sans passer par la trigonométrie, inconnue d'Archimède !

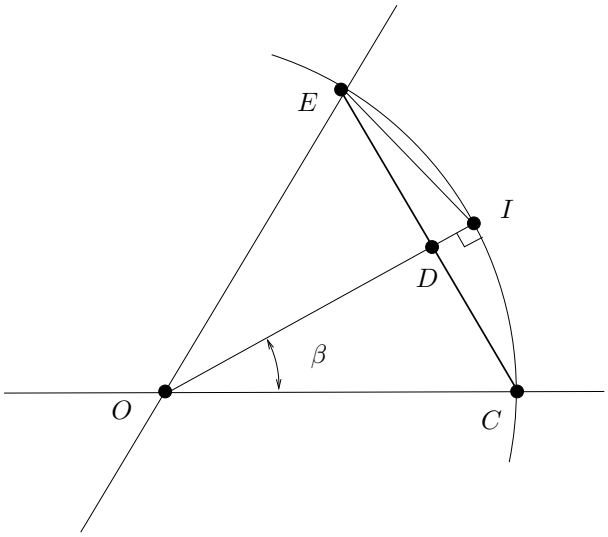


FIGURE AA.6. Relation entre U_n et U_{n+1}.

Sur la figure AA.6, on suppose que $U_n = CE$. Calculons $U_{n+1} = EI$, en appliquant deux fois le théorème de Pythagore aux deux triangles EDI et ODE, ce qui fournit :

$$EI^2 = ED^2 + DI^2,$$
$$OE^2 = OD^2 + DE^2,$$

et donc

$$U_{n+1}^2 = \frac{1}{4}U_n^2 + DI^2,$$
$$\frac{1}{4} = OD^2 + \frac{1}{4}U_n^2$$

et puisque

$$\frac{1}{2} = OI = OD + DI,$$

on a donc, en supposant

$$U_n \leq 1, \tag{AA.5}$$

on a

$$\begin{aligned} U_{n+1}^2 &= \frac{1}{4}U_n^2 + \left(\frac{1}{2} - OD\right)^2, \\ &= \frac{1}{4}U_n^2 + \left(\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{4}U_n^2}\right)^2, \\ &= \frac{1}{4}U_n^2 + \frac{1}{4} + \frac{1}{4} - \frac{1}{4}U_n^2 - \sqrt{\frac{1}{4} - \frac{1}{4}U_n^2}, \\ &= \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{4}U_n^2}, \\ &= \frac{1}{2}\left(1 - \sqrt{1 - U_n^2}\right), \end{aligned}$$

et donc puisque $U_{n+1} \geq 0$ et $1 \geq \sqrt{1-U_n^2}$

$$U_{n+1} = \frac{\sqrt{2}}{2}\sqrt{1-\sqrt{1-U_n^2}}. \tag{AA.6}$$

À ce niveau-là, on peut montrer par récurrence sur n que (AA.5) est valable pour tout n. D'après (AA.4), c'est vrai pour $n = 0$. Si c'est vrai pour un entier n, les calculs donnant U_{n+1} sont valables. D'après (AA.6), $U_{n+1} \leq 1$ est successivement équivalent à

$$\begin{aligned}\frac{\sqrt{2}}{2}\sqrt{1-\sqrt{1-U_n^2}} \leq 1 &\iff \sqrt{1-\sqrt{1-U_n^2}} \leq \sqrt{2},\\ &\iff 1-\sqrt{1-U_n^2} \leq 2,\\ &\iff -1 \leq \sqrt{1-U_n^2},\end{aligned}$$

ce qui est vrai puisque $-1 < 0$.

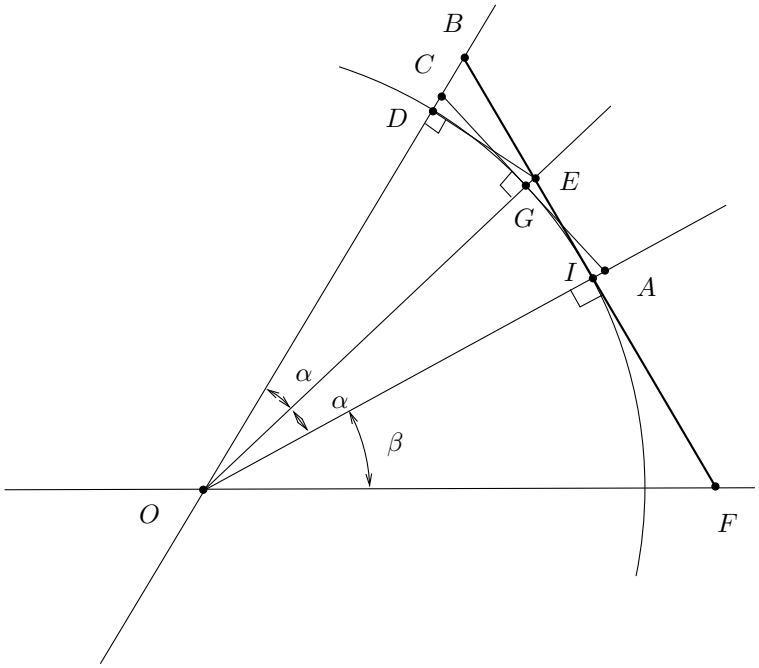


FIGURE AA.7. Relation entre V_n et V_{n+1}.

Faisons de même pour calculer V_{n+1} en fonction de V_n. Sur la figure AA.7, on suppose que $V_n = BF$. Calculons $V_{n+1} = CA$, de nouveau en appliquant deux fois le théorème de Pythagore judicieusement. Calculons tout d'abord CG en fonction de BI. On introduit le point E, intersection des droites (OG) et (IB) (qui ne sont pas parallèles). Une astuce consiste à remarquer que

$$IE = CG = DE = GA, \tag{AA.7}$$

puisque chacun des segments est défini par le même angle α. Voir figure AA.7. Il nous faut donc calculer

$$a = IE \tag{AA.8}$$

en fonction de

$$b = IB. \tag{AA.9}$$

Notons

$$e = DB. \tag{AA.10}$$

Par symétrie, on a aussi l'orthogonalité des droites (ED) et (OD). Le théorème de Pythagore appliqué au triangle DEB fournit donc

$$BD^2 + DE^2 = EB^2,$$

soit encore

$$BD^2 + a^2 = (IB - IE)^2 = (b - a)^2,$$

et donc

$$BD^2 + a^2 = a^2 - 2ab + b^2,$$

d'où

$$2ab = b^2 - BD^2,$$

et puisque b est non nul :

$$a = \frac{1}{2b}\left(b^2 - e^2\right) \tag{AA.11}$$

Appliquons maintenant le théorème de Pythagore au triangle BIO :

$$OI^2 + IB^2 = OB^2,$$

soit puisque le cercle est de rayon $1/2$:

$$\frac{1}{4} + b^2 = (OD + DB)^2 = \left(\frac{1}{2} + BD\right)^2$$

ce qui donne

$$\left(e + \frac{1}{2}\right)^2 = b^2 + \frac{1}{4},$$

et

$$e + \frac{1}{2} = \sqrt{b^2 + \frac{1}{4}}$$

soit encore

$$e = -\frac{1}{2} + \sqrt{b^2 + \frac{1}{4}}. \tag{AA.12}$$

On a plus qu'à réutiliser (AA.11) qui fournit :

$$\begin{aligned}
a &= \frac{1}{2b}\left(b^2 - e^2\right), \\
&= \frac{1}{2b}\left(b^2 - \left(-\frac{1}{2} + \sqrt{b^2 + \frac{1}{4}}\right)^2\right), \\
&= \frac{1}{2b}\left(b^2 - \frac{1}{4} + \sqrt{b^2 + \frac{1}{4}} - b^2 - \frac{1}{4}\right), \\
&= \frac{1}{2b}\left(-\frac{1}{2} + \sqrt{b^2 + \frac{1}{4}}\right), \\
&= \frac{1}{4b}\left(-1 + \sqrt{4b^2 + 1}\right),
\end{aligned}$$

et donc

$$a = \frac{1}{4b}\left(-1 + \sqrt{4b^2 + 1}\right). \tag{AA.13}$$

On a enfin, d'après (AA.7) et (AA.8)

$$V_{n+1} = CA = 2CG = 2IE = 2a,$$

et d'après (AA.9)

$$V_n = BF = 2BI = 2b.$$

Tout cela donne grâce à (AA.13) :

$$\frac{V_{n+1}}{2} = \frac{1}{2(2b)}\left(-1+\sqrt{(2b)^2+1}\right) = \frac{1}{2V_n}\left(-1+\sqrt{V_n^2+1}\right)$$

et finalement

$$V_{n+1} = \frac{1}{V_n}\left(-1+\sqrt{V_n^2+1}\right) \tag{AA.14}$$

On vérifie aisément par récurrence sur n que cette expression assure que pour tout n, on a $0 < V_n$.

Remarque AA.1. Notons qu'en utilisant le langage des lignes trigonométriques, on a (voir l'angle β sur la figure AA.7)

$$V_n = 2BI = 2IF = 2OI\tan\beta,$$

soit

$$V_n = \tan\beta \tag{AA.15}$$

Naturellement, on a aussi

$$V_{n+1} = CA = 2CG = \tan\alpha,$$

soit puisque $\alpha = \beta/2$:

$$V_{n+1} = \tan\left(\frac{\beta}{2}\right). \tag{AA.16}$$

Le calcul qui est fait dans `http://www.pi314.net/fr/archi.php` consiste à calculer classiquement $\tan(2\theta)$ en fonction de $\tan(\theta)$ puis, par le biais de la résolution d'une équation du second degré en $\tan(\theta)$ d'obtenir l'expression de $\tan(\theta)$ en fonction de $\tan(2\theta)$. Ainsi, grâce à (AA.16), on obtient l'expression de V_{n+1} en fonction de celle de V_n, ce que l'on a fait avec deux simples applications du théorème de Pythagore !

Remarque AA.2. Comme dans la remarque AA.1, on a aussi (voir l'angle β sur la figure AA.6) :

$$U_n = CE = 2CD = 2OI\sin\beta,$$

et donc

$$U_n = \sin\beta \tag{AA.17}$$

Les valeurs initiales des suites U_0 et V_0 sont données par (AA.4) et les relations de récurrence vérifiées par les suites U_n et V_n sont données par (AA.6) et (AA.14), ce qui donne finalement

$$U_0 = \frac{1}{2}, \tag{AA.18a}$$

$$V_0 = \frac{\sqrt{3}}{3}, \tag{AA.18b}$$

$$\forall n \in \mathbb{N}, \quad U_{n+1} = \frac{\sqrt{2}}{2}\sqrt{1-\sqrt{1-U_n^2}}, \tag{AA.18c}$$

$$V_{n+1} = \frac{1}{V_n}\left(-1+\sqrt{V_n^2+1}\right). \tag{AA.18d}$$

D'après, (AA.3), puisque chacun des polygones est constitué de 6.2^n cotés chacun de longueur U_n et V_n, on obtient donc l'expression des périmètres I_n et C_n des polygones inscrits et circonscrits :

$$\forall n \in \mathbb{N}, \quad I_n = 6.2^n U_n, \tag{AA.19a}$$

$$C_n = 6.2^n V_n. \tag{AA.19b}$$

On peut donc déterminer numériquement les valeurs de I_n et C_n pour tout n, grâce à (AA.18). Naturellement, chacune des valeurs de I_n et C_n peut s'obtenir à la règle et au compas (voir figure AA.4).

Remarque AA.3. On peut aussi obtenir des relations de récurrence qui portent uniquement sur I_n et C_n en écrivant

$$\begin{aligned}
I_{n+1} &= 6.2^{n+1}U_{n+1},\\
&= 6.2^{n+1}\frac{\sqrt{2}}{2}\sqrt{1-\sqrt{1-U_n^2}},\\
&= 6.2^{n+1}\frac{\sqrt{2}}{2}\sqrt{1-\sqrt{1-\left(\frac{I_n}{6.2^n}\right)^2}},\\
&= 6.2^{n+1}\frac{\sqrt{2}}{2}\sqrt{1-\frac{1}{6.2^n}\sqrt{(6.2^n)^2-I_n^2}},\\
&= 6.2^{n+1}\frac{\sqrt{2}}{2}\frac{1}{\sqrt{6.2^n}}\sqrt{6.2^n-\sqrt{(6.2^n)^2-I_n^2}},\\
&= 2.6.2^n\frac{\sqrt{2}}{2}\frac{1}{\sqrt{6.2^n}}\sqrt{6.2^n-\sqrt{(6.2^n)^2-I_n^2}},\\
&= \sqrt{6.2^n}\sqrt{2}\sqrt{6.2^n-\sqrt{(6.2^n)^2-I_n^2}},
\end{aligned}$$

et donc

$$I_0 = 3, \tag{AA.20a}$$

$$\forall n \in \mathbb{N}, \quad I_{n+1} = \sqrt{2q_n}\sqrt{q_n-\sqrt{q_n^2-I_n^2}}, \tag{AA.20b}$$

$$q_0 = 6, \tag{AA.20c}$$

$$\forall n \in \mathbb{N}, \quad q_{n+1} = 2q_n. \tag{AA.20d}$$

On montre de même qu'avec la même définition de q_n :

$$C_0 = 2\sqrt{3}, \tag{AA.21a}$$

$$\forall n \in \mathbb{N}, \quad C_{n+1} = \frac{2q_n}{C_n}\left(-q_n+\sqrt{C_n^2+q_n^2}\right). \tag{AA.21b}$$

On consultera les simulations numériques faites plus bas.

AA.2.5. Preuve géométrique de la convergence des deux suites I_n et C_n vers π

Par exemple, pour $n = 0$, on retrouve (AA.1) :

$$I_0 = 3, \quad C_0 = 2\sqrt{3}$$

L'idée d'Archimède est que I_n et C_n constituent une approximation de π par défaut et par excès, puisque qu'elles correspondent aux périmètres des polygones qui se rapprochent du cercle et qui sont respectivement inscrits et circonscrits donc de périmètres respectivement plus petits et plus grands que celui de cercle, dont on sait qu'il vaut π.

Tentons de fournir une preuve qu'aurait pu écrire Archimède en n'utilisant que les calculs précédents, fondés sur la géométrie élémentaire (on n'utilise que les théorèmes de Thalès et de Pythagore, sans utilisation des lignes trigonométriques et de leurs propriétés analytiques).

Donnons tout d'abord le lemme suivant, fondé sur la géométrie élémentaire :

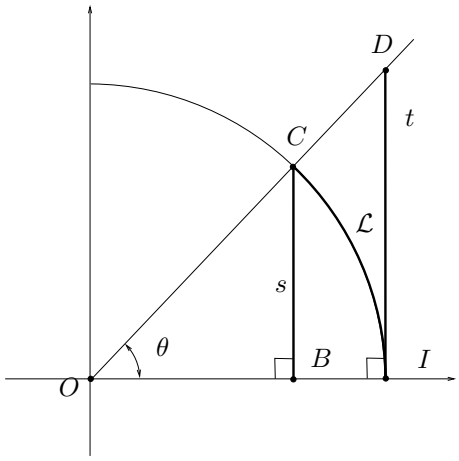


FIGURE AA.8. La longueur de l'arc de cercle $\mathcal{L}$, s et t.

LEMME AA.4.

Soit $\theta \in]0, \pi/2[$ et le cercle de centre O et de rayon r (voir figure AA.8.) On considère C le point du cercle définissant l'angle θ, B son projeté orthogonal sur l'axe de x, I le point de l'axe des x d'abscisse 1 et D le point de la droite (OC) d'abscisse 1. On note $s = BC$, $\mathcal{L}$, la longueur de l'arc engendré par l'arc de cercle de centre O et défini par l'angle θ et $t = IB$. On a alors

$$s < \mathcal{L} < t, \tag{AA.22}$$

et

$$\sin\theta < \theta < \tan\theta. \tag{AA.23}$$

DÉMONSTRATION. Dans le (vrai) triangle rectangle CBI, l'hypoténuse CI est strictement plus grande que le coté BC. La longueur $\mathcal{L}$ est strictement plus grande que la corde CI et on a donc

$$s < \mathcal{L}. \tag{AA.24}$$

L'angle θ étant dans $]0, \pi/2[$, le segment le segment $[ID]$ est à l'extérieur du cercle. On en déduit que l'aire de la portion de disque délimitée par θ est strictement plus petite que l'aire du triangle ODI. Ces aires sont respectivement égales à $\theta r^2/2$ et $rt/2$. Puisque

$$\mathcal{L} = r\theta,$$

ces aires valent donc respectivement $\mathcal{L}r/2$ et $tr/2$ et on a donc

$$\frac{\mathcal{L}r}{2} < \frac{tr}{2},$$

et donc

$$\mathcal{L} < t. \tag{AA.25}$$

L'inégalité (AA.22) provient donc (AA.24) et de (AA.25). En remplaçant dans (AA.22) respectivement s, $\mathcal{L}$ et t par $r\sin\theta$, $r\theta$ et $r\tan\theta$, on déduit (AA.23).

REMARQUE AA.5. Une preuve "moderne" pour démontrer analytiquement directement (AA.23) (et donc en déduire (AA.22) qui est équivalente) consiste à utiliser la convexité et la concavité des fonctions tan et sin. En effet, on a

$$\forall x \in [0, \pi/2[, \quad \tan'(x) = \frac{1}{\cos^2 x} > 0.$$

Donc la dérivée de tan est strictement croissante (car cos est décroissante) et donc tan est strictement convexe sur $[0, \pi/2[$. On a aussi

$$\forall x \in [0, \pi/2[, \quad \sin''(x) = -\cos(x) < 0,$$

Donc sin est strictement convexe sur $[0, \pi/2[$. Ainsi, le graphe de la fonction tan est toujours strictement au-dessus de sa tangente en zéro sur l'intervalle $[0, \pi/2[$, tandis que le graphe de la fonction sin est toujours strictement en-dessous de sa tangente en zéro sur l'intervalle $[0, \pi/2[$. Les deux tangentes des deux fonctions tan et sin en zéro sont la droite d'équation $y = x$. On a donc

$$\forall x \in [0, \pi/2[, \quad \tan(x) > x > \sin(x).$$

Mais ce raisonnement nous pousse à tourner à rond [2] car cette preuve est fondée sur la dérivée du sinus, elle même fondée sur l'inégalité que nous sommes en train de montrer !

□

Si on applique la première inégalité de (AA.22), à la situation de la figure AA.6, en notant $\widetilde{\mathcal{L}}$ la longueur de l'arc de cercle délimité par I et E, on a

$$ED < \widetilde{\mathcal{L}}. \tag{AA.26}$$

En notant $\mathcal{L}$ la longueur de l'arc de cercle délimité par C et E qui vaut $2\widetilde{\mathcal{L}}$, on déduit de (AA.26)

$$U_n < \mathcal{L}. \tag{AA.27}$$

De même, la seconde inégalité de (AA.22), à la situation de la figure AA.7, où cette fois-ci, $\widetilde{\mathcal{L}}$ désigne la longueur de l'arc de cercle délimité par I et D et $\mathcal{L}$ désigne la longueur de l'arc de cercle délimité par F et B, on a

$$IB > \widetilde{\mathcal{L}}, \tag{AA.28}$$

dont on déduit

$$V_n > \mathcal{L}. \tag{AA.29}$$

En multipliant respectivement (AA.27) et (AA.29) par le nombre de cotés des polygones (à l'étape n et qui correspond aussi au nombre de fois où $\mathcal{L}$ est répété pour obtenir le périmètre de cercle), on obtient donc

$$I_n < \pi, \tag{AA.30a}$$

et

$$\pi < C_n. \tag{AA.30b}$$

Si on revient de nouveau sur la figure AA.6, l'hypoténuse du triangle rectangle EDI étant strictement supérieure à son coté ED. On a donc $U_{n+1} = EI > ED = U_n/2$ et donc

$$U_n < 2U_{n+1}$$

Si on multiplie cela par le nombre de cotés, on obtient alors

$$\forall n \in \mathbb{N}, \quad I_n < I_{n+1}. \tag{AA.31}$$

Si on revient de nouveau sur la figure AA.7, l'hypoténuse du triangle rectangle BDE étant strictement supérieure à son coté ED. On a donc $BE > DE$. Ainsi,

$$BI = BE + EI > DE + EI,$$

et d'après (AA.7), il vient donc

$$BI > CG + GA,$$

et donc

$$BI > CA,$$

2. dans le sens trigonométrique ...

soit encore

$$\frac{V_n}{2} > V_{n+1},$$

et finalement

$$V_n > 2V_{n+1},$$

Si on multiplie cela par le nombre de cotés, on obtient alors

$$\forall n \in \mathbb{N}, \quad C_n > C_{n+1}. \tag{AA.32}$$

D'après (AA.30), (AA.31) et (AA.32), les deux suites I_n et C_n sont donc respectivement croissante et majorée et décroissante et minorée et convergent donc respectivement vers deux limites I et C (qui sont strictement positives).

Le raisonnement d'Archimède aurait pu être le suivant : quand θ est "petit", d'après (AA.22), on a

$$s \approx \mathcal{L} \approx t,$$

et, en raisonnant comme dans les inégalités (AA.27) et (AA.29), on a, pour n "grand",

$$U_n \approx \mathcal{L},$$
$$V_n \approx \mathcal{L},$$

dont on déduit, comme précédemment,

$$I_n \approx \pi,$$
$$C_n \approx \pi,$$

et donc que les deux limites I et C sont égales à π.

Plus rigoureusement, on peut utiliser (AA.15) et (AA.17) (qu'Archimède aurait pu écrire sous une autre forme, car n'on utilise que les définitions géométriques des lignes trigonométriques et non leur propriété analytique) qui donnent

$$\frac{U_n}{V_n} = \frac{\sin\beta}{\tan\beta},$$

et donc

$$\forall n \in \mathbb{N}, \quad \frac{U_n}{V_n} = \cos\beta. \tag{AA.33}$$

Quand n tend vers l'infini, β tend vers zéro et d'après l'inégalité (AA.23), $\sin\beta$ tend vers zéro et donc $\cos\beta$ tend vers 1. Ainsi, d'après (AA.33)

$$\lim_{n\to+\infty} \frac{U_n}{V_n} = 1. \tag{AA.34}$$

En multipliant à gauche par le nombre de cotés dans la fraction, on en déduit

$$\lim_{n\to+\infty} \frac{I_n}{C_n} = 1. \tag{AA.35}$$

et donc

$$\frac{I}{C} = 1.$$

et donc

$$I = C = l.$$

Si on passe à la limite n tendant vers l'infini dans (AA.30), on obtient

$$l \leq \pi \leq l$$

et donc

$$l = \pi.$$

On a donc montré que

$$\forall n \in \mathbb{N}, \quad I_n < \pi < C_n, \tag{AA.36a}$$

$$\lim_{n\to+\infty} I_n = \pi, \tag{AA.36b}$$

$$\lim_{n\to+\infty} C_n = \pi. \tag{AA.36c}$$

On peut donc affirmer *a posteriori* que les deux suites I_n et C_n sont adjacentes.

AA.2.6. Définition de la longueur d'un cercle et de π.

À un niveau élémentaire, les résultats précédents peuvent en fait servir à définir la longueur d'un cercle et π. Si on considère la longueur d'un segment comme connue, la longueur d'une courbe quelconque doit se faire par un passage à la limite. Par exemple, pour une courbe paramétrée, on pourra consulter [Bas11, section 3.2.1 Arc rectifiable et longueur] qui nous montre qu'il est nécessaire de passer par une intégrale, ce que l'on veut éviter ici, afin de rester dans le domaine de la géométrie élémentaire.

De façon élémentaire, on peut se passer même de la notion d'angle (comme le faisaient les Mésopotamiennes [3]). Nous n'utilisons maintenant plus la longueur de cercles et d'arcs de cercle, ni même la définition de π (comme périmètre d'un cercle de rayon 1/2 connu). On renvoie à la figure AA.9 où apparaissent U_n et V_n.

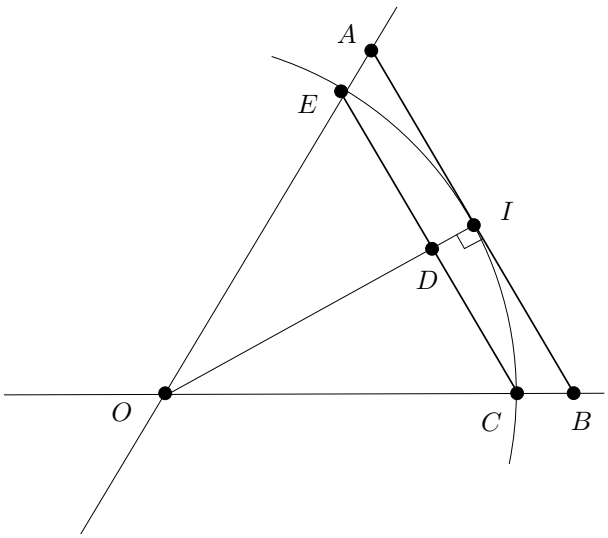


FIGURE AA.9. Relation entre U_n et V_n.

D'après les théorèmes de Thalès et de Pythagore appliqués respectivement aux triangles OAI et ODE, on a successivement

$$\begin{aligned}
\frac{U_n}{V_n} &= \frac{ED}{AI}, \\
&= \frac{OD}{OI}, \\
&= \frac{\sqrt{OE^2 - ED^2}}{OI}, \\
&= \frac{\sqrt{\frac{1}{4} - \frac{U_n^2}{4}}}{\frac{1}{2}},
\end{aligned}$$

3. voir par exemple `https://fr.wikipedia.org/wiki/Mathématiques_mésopotamiennes`

et donc

$$\forall n \in \mathbb{N}, \quad \frac{U_n}{V_n} = \sqrt{1 - U_n^2}. \tag{AA.37}$$

On déduit de (AA.37) que

$$\forall n \in \mathbb{N}, \quad U_n < V_n. \tag{AA.38}$$

Les égalités (AA.31) et (AA.32) impliquent toujours que les deux suites I_n et C_n sont respectivement croissante et minorée. De plus, de (AA.38), on déduit qu'elles sont respectivement majorée et minorée. Elles convergent donc respectivement vers deux limites I et C (qui sont strictement positives). Enfin, si on utilise (AA.19a), on a

$$\forall n \in \mathbb{N}, \quad V_n = \frac{C_n}{6.2^n},$$

et d'après (AA.32), on a donc

$$\forall n \in \mathbb{N}, \quad V_n < \frac{C_0}{6.2^n}. \tag{AA.39}$$

Ainsi, de (AA.38) et (AA.39), on déduit

$$\lim_{n \to +\infty} U_n = 0, \tag{AA.40a}$$

$$\lim_{n \to +\infty} V_n = 0, \tag{AA.40b}$$

et donc, grâce à (AA.37), on obtient de nouveau (AA.34) et (AA.35). Les deux limites des suites I_n et C_n sont donc égales et la limite pourra donc être la définition de π et qui sera aussi la longueur du cercle de rayon $1/2$ dont on est parti.

AA.2.7. Preuve analytique de la convergence des deux suites I_n et C_n vers π et qualité de la convergence

Montrons cette fois-ci analytiquement (AA.36), en précisant la qualité de la convergence et en utilisant les propriétés analytiques des lignes trigonométriques, ce que n'aurait pas probablement fait Archimède !

Montrons maintenant rigoureusement (AA.36) avec un langage "moderne". Pour cela, on fournit maintenant une expression explicite de I_n et de C_n en fonction de n. A la n-ième étape, on note α_n l'angle au sommet correspondant à chacun des secteurs angulaires définissant U_n et V_n. Cet angle est égal à 2π divisé par le nombre de coté soit 6.2^n. On a donc

$$\forall n \in \mathbb{N}, \quad \alpha_n = \frac{2\pi}{6.2^n}. \tag{AA.41}$$

Les angles β donnés sur les figures AA.6 et AA.7 sont égaux à $\alpha_n/2$ (par symétrie) et donc

$$\forall n \in \mathbb{N}, \quad \beta_n = \frac{\pi}{6.2^n}. \tag{AA.42}$$

D'après (AA.17)

$$U_n = \sin\left(\frac{\pi}{6.2^n}\right),$$

et donc, d'après (AA.3) et (AA.19a), il vient :

$$\forall n \in \mathbb{N}, \quad I_n = 6.2^n \sin\left(\frac{\pi}{6.2^n}\right). \tag{AA.43}$$

De même, d'après (AA.15), on a

$$\forall n \in \mathbb{N}, \quad C_n = 6.2^n \tan\left(\frac{\pi}{6.2^n}\right). \tag{AA.44}$$

Les deux équations (AA.43) et (AA.44) n'ont aucune utilité pratique pour déterminer les valeurs numériques de U_n et de V_n puisqu'elles font intervenir π que l'on cherche ! On leur préférera à cet égard, les formules (AA.20) et (AA.21). Mais pour l'étude de la convergence, elles sont fondamentales. On a effet, d'après (AA.23),

$$\forall \theta \in]0, \pi/2[, \quad \sin\theta < \theta < \tan\theta, \tag{AA.45a}$$

et le cours d'analyse (formule de Taylor-Lagrange) nous montre que

$$\sin(\theta) = \theta + O\left(\theta^3\right), \tag{AA.45b}$$
$$\tan(\theta) = \theta + O\left(\theta^3\right). \tag{AA.45c}$$

Les inéquations (AA.45a) donnent, pour $\theta = \frac{\pi}{6.2^n}$

$$\forall n \in \mathbb{N}, \quad 6.2^n \sin\left(\frac{\pi}{6.2^n}\right) < 6.2^n \frac{\pi}{6.2^n} < 6.2^n \tan\left(\frac{\pi}{6.2^n}\right)$$

et donc, d'après (AA.43) et (AA.44), on retrouve (AA.36a). On obtient aussi, grâce à (AA.45b),

$$\begin{aligned} I_n &= 6.2^n \left(\frac{\pi}{6.2^n} + O\left(\frac{\pi}{2^{3n}}\right)\right), \\ &= \pi + O\left(\frac{1}{2^{2n}}\right), \\ &= \pi + O\left(\frac{1}{4^n}\right) \end{aligned}$$

On fait de même pour C_n, grâce à (AA.45c), et on obtient finalement

$$I_n = \pi + O\left(\frac{1}{4^n}\right), \tag{AA.46a}$$
$$C_n = \pi + O\left(\frac{1}{4^n}\right), \tag{AA.46b}$$

ce qui implique bien (AA.36b) et (AA.36c).

Nous constateront aussi cela de façon numérique plus bas.

Remarque AA.6. À partir des équations (AA.20) et (AA.21) il n'était pas possible de montrer la convergence des suites vers π. Notons cependant que les équations (AA.18c) et (AA.18d) nous permettent de retrouver (AA.40). En effet, les suites U_n et V_n sont en effet définies par $U_{n+1} = F(U_n)$ et $V_{n+1} = G(U_n)$ où

$$\forall x \in \mathbb{R}, \quad F(x) = \frac{\sqrt{2}}{2}\sqrt{1 - \sqrt{1 - x^2}}, \tag{AA.47a}$$
$$G(x) = \frac{1}{x}\left(-1 + \sqrt{x^2 + 1}\right). \tag{AA.47b}$$

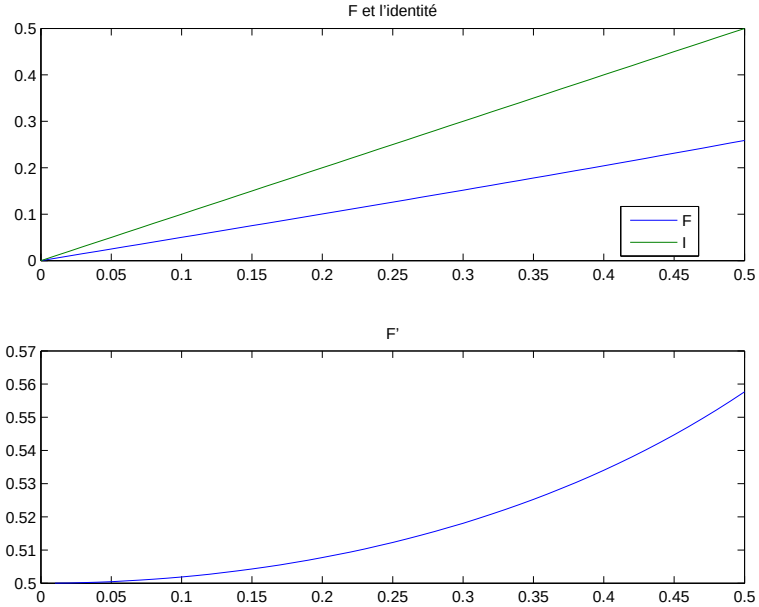
On peut montrer que l'on a les propriétés suivantes (voir figures AA.10) : si on pose $\mathcal{I} = [0, U_0]$ et $\mathcal{J} = [0, V_0]$, F est $\mathcal{C}^1$ sur $\mathcal{I}$, G, prolongée par 0 en 0 est $\mathcal{C}^1$ sur $\mathcal{J}$. De plus,

$$\begin{aligned} &F(\mathcal{I}) \subset \mathcal{I} \text{ et il existe } k \in [0, 1[\text{ tel que } \forall x \in \mathcal{I},\ |F'(x)| \leq k\,; \\ &F(\mathcal{J}) \subset \mathcal{I} \text{ et il existe } l \in [0, 1[\text{ tel que } \forall x \in \mathcal{I},\ |F'(x)| \leq l. \end{aligned}$$

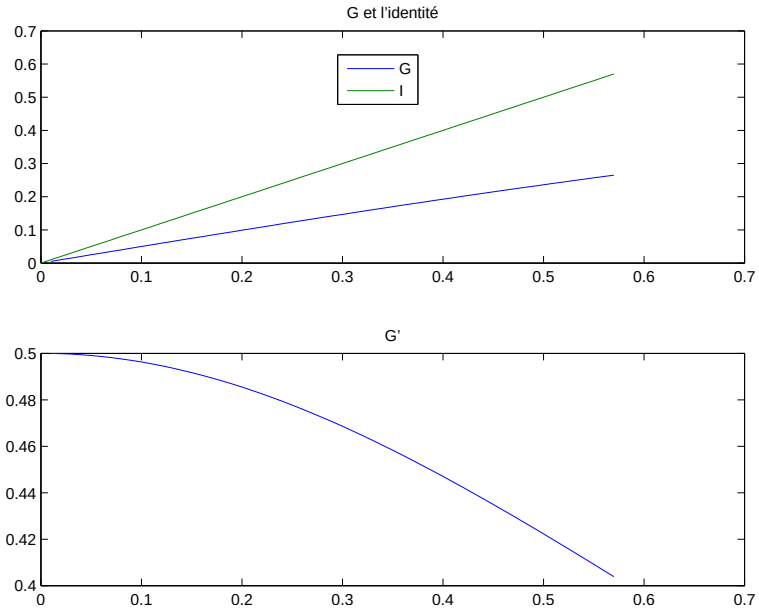
D'après [DB21, Théorème "condition suffisante de convergence de la méthode du point fixe", Chapitre 3], les deux suites convergent respectivement vers l'unique point fixe de F et de G sur $\mathcal{I}$ et $\mathcal{J}$. Puisque $F(0) = G(0) = 0$, la limite commune est donc nulle. Enfin, on peut montrer que

$$\begin{aligned} &\forall x \in]0, U_0], \quad F(x) < x, \\ &\forall x \in]0, V_0], \quad G(x) < x, \end{aligned}$$

ce qui implique que les deux suites U_n et V_n sont strictement décroissantes.



(a) La fonction F et sa dérivée.



(b) La fonction G et sa dérivée.

FIGURE AA.10. Les fonctions F et G.

AA.3. Méthode de Cues

L'idée est cette fois-ci de construire une suite de polygones de périmètres constants, qui vont se rapprocher d'un cercle de périmètre égal à ce périmètre ! Cependant, on fait le contraire de la méthode d'Archimède : on déterminer le rayon d'un cercle dont le périmètre est fixé à l'avance. Là encore, plutôt que de donner la récurrence fondée sur la trigonométrie, comme c'est fait dans `http://www.pi314.net/fr/cues.php`, nous

donnons une preuve géométrique, qui met aussi en avance que la construction est aussi possible à la règle et au compas, ce qui est fait dans `http://serge.mehl.free.fr/anx/iso_perim.html`.

AA.3.1. Principe

Nous partons d'un carré de coté 1/4 et donc de périmètre 1. Nous allons cette fois-ci construire une suite de polygônes réguliers de mêmes périmètres, tous égaux à 1, le coté du nouveau polygône étant égal à la moitié de l'ancien et comme pour la méthode d'Archimède, le nombre de sommet étant le double. Cette suite de polygône se rapprochera d'un cercle de périmètre 1 et donc de rayon $1/(2\pi)$.

AA.3.2. Mise en évidence géométrique

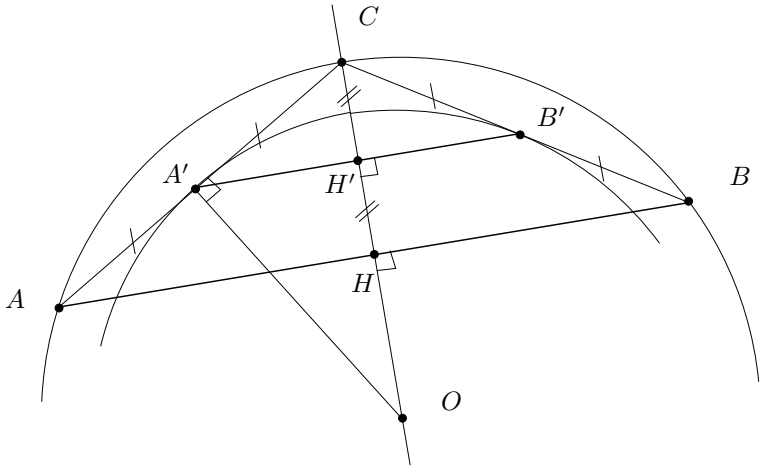


FIGURE AA.11. Passage du polygone à celui d'un polygone à deux fois plus de côtés, de même périmètre.

On se donne tout d'abord un polygone régulier de coté $[AB]$ (ayant en tout n cotés), inscrit dans un cercle de rayon r et de centre O et on construit un polygone ayant deux fois plus de cotés et de périmètre égal au premier (il sera dit isopérimétre au premier). Voir figure AA.11. On considère C le milieu de l'arc AB, A' le milieu de $[AC]$ et B' le milieu de $[BC]$, H le milieu de $[AB]$ et H' le milieu de $[A'B']$. On considère alors le polygone régulier à $2n$ cotés, de centre O, dont l'un des cotés est $[A'B']$ et nous montrons qu'il est bien isopérimètre au premier polygone. Notons r le rayon OA, cercle dans lequel est inscrit le premier polygone et a son apothème, c'est-à-dire, la distance de O au milieu de AB, soit la distance OH. On a donc

$$r = OA, \quad a = OH. \tag{AA.48}$$

On note de même

$$r' = OA', \quad a' = OH'. \tag{AA.49}$$

qui correspondent respectivement au rayon du cercle dans lequel est inscrit le second polygone et à l'apothème de ce polygone. Il évident, grâce au théorème de la droite des milieux que $A'B' = AB/2$ et donc que le périmètre du second polygone est égal à celui du premier (puisqu'il a deux fois plus de côtés). Ce second polygone sera donc inscrit dans le cercle de centre O et de rayon r'. Déterminons maintenant a' et r' en fonction de a et de r. D'après le théorème de la droite des milieux, H' est le milieu de $[CH]$, on a donc

$$OH' = \frac{OH + OC}{2}$$

et donc

$$a' = \frac{a + r}{2}. \tag{AA.50}$$

Par ailleurs, A et C sont équidistants de O (car appartenant au même cercle de centre O) et puisque A' est le milieu de $[AC]$, la droite (OA') est la médiatrice de $[AC]$. Ainsi dans le triangle rectangle $OA'C$, de hauteur $A'H'$ issue de A, on a

$$OA'^2 = OH'.OC. \tag{AA.51}$$

Donnons en effet le petit résultat très classique suivant :

Lemme AA.7. *Soit ABC un triangle rectangle en B et H le pied de la hauteur issue de B. On a*

$$AB^2 = AH.AC. \tag{AA.52}$$

Démonstration.

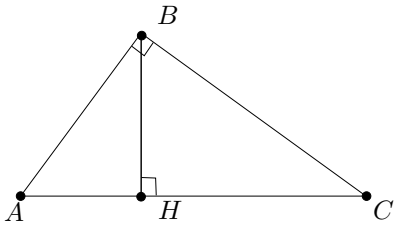


Figure AA.12. Le triangle rectangle ABC et sa hauteur BH.

On peut en donner deux preuves, la première élémentaire n'utilisant que le théorème de Pythagore (donc connue de Cues !) et la seconde utilisant le produit scalaire. On se réfère à la figure AA.12.

(1) Le théorème de Pythagore appliqué aux triangles rectangles ABC, ABH et BHC donne respectivement :

$$\begin{aligned} AC^2 &= AB^2 + BC^2, \\ AB^2 &= AH^2 + HB^2, \\ BC^2 &= BH^2 + HC^2, \end{aligned}$$

et l'on vérifie que le point H est nécessairement sur $[AC]$, ce qui donne :

$$AC = AH + HC.$$

On déduit donc successivement de tout cela que

$$\begin{aligned} AB^2 &= AH^2 + HB^2, \\ &= AH^2 + BC^2 - HC^2, \\ &= AH^2 + AC^2 - AB^2 - HC^2, \end{aligned}$$

et donc

$$\begin{aligned} 2AB^2 &= AH^2 + AC^2 - HC^2, \\ &= AH^2 + AC^2 - (AC - AH)^2, \\ &= AH^2 + AC^2 - AC^2 - AH^2 + 2ACAH, \\ &= 2AC.AH. \end{aligned}$$

(2) Plus rapidement, avec les produits scalaires, on a successivement

$$\begin{aligned} AB^2 &= \overrightarrow{AB}.\overrightarrow{AB}, \\ &= \left(\overrightarrow{AH}+\overrightarrow{HB}\right).\left(\overrightarrow{AC}+\overrightarrow{CB}\right), \\ &= \overrightarrow{AH}.\overrightarrow{AC}+\overrightarrow{AH}.\overrightarrow{CB}+\overrightarrow{HB}.\overrightarrow{AC}+\overrightarrow{HB}.\overrightarrow{CB}, \end{aligned}$$

puisque (BH) est perdendiculaire à (AC), on a $\overrightarrow{HB}.\overrightarrow{AC}=0$ et donc

$$\begin{aligned} &= \overrightarrow{AH}.\overrightarrow{AC}+\overrightarrow{AH}.\overrightarrow{CB}+\overrightarrow{HB}.\overrightarrow{CB}, \\ &= \overrightarrow{AH}.\overrightarrow{AC}+\left(\overrightarrow{AH}+\overrightarrow{HB}\right).\overrightarrow{CB}, \\ &= \overrightarrow{AH}.\overrightarrow{AC}+\overrightarrow{AB}.\overrightarrow{CB}, \end{aligned}$$

puisque (AB) est perdendiculaire à (BC), on a $\overrightarrow{AB}.\overrightarrow{CB}=0$ et donc

$$= \overrightarrow{AH}.\overrightarrow{AC},$$

quantité qui vaut $AH.AC$ puisque le point H est nécessairement sur $[AC]$.

□

De (AA.51), on déduit donc

$$r' = \sqrt{ra'}. \tag{AA.53}$$

Grâce à (AA.50) et (AA.53), on a donc explicité r' et a' en fonction de r et a (et implicitement vérifié que la construction du second polygone en fonction du premier est possible à la règle et au compas.).

Comme dans la méthode d'Archimède, on construit deux suites r_n et a_n correspondant respectivement au rayon du n-ième polygone construit à 2^n côtés et à l'apothème de ce polygone. Le périmètre sera choisi constant, égal à 1. Au début de la construction, on a $n=2$. On remplace donc (AA.3) par

$$\text{Pour } n \geq 2, \text{ le polygone a } 2^n \text{ côtés.} \tag{AA.54}$$

Nous supposerons que

$$\text{le périmètre constant du polygone vaut 1.} \tag{AA.55}$$

On a donc la valeur c_n du côté du polygone à l'étape n :

$$\forall n \geq 2, \quad c_n = \frac{1}{2^n}. \tag{AA.56}$$

Nous avons les relations suivantes de récurrence, obtenues grâce à (AA.50) et (AA.53)

$$\forall n \geq 2, \quad a_{n+1} = \frac{a_n + r_n}{2}, \tag{AA.57a}$$

$$r_{n+1} = \sqrt{r_n a_{n+1}}. \tag{AA.57b}$$

Attention, ce ne sont pas excactement les fameuses suites arithmético-géométriques que l'on verra plus bas.

L'avantage de cette méthode est de proposer une construction de suite a_n et r_n dont la limite est liée à π. Dans la méthode d'Archimède, au contraire, il fallait multiplier par 2^n une longueur qui tendait vers zéro, ce qui pose des problèmes pour la construction géométrique (dupliquer un grand nombre de fois le même segment) ou numériquement, comme on le vera plus bas.

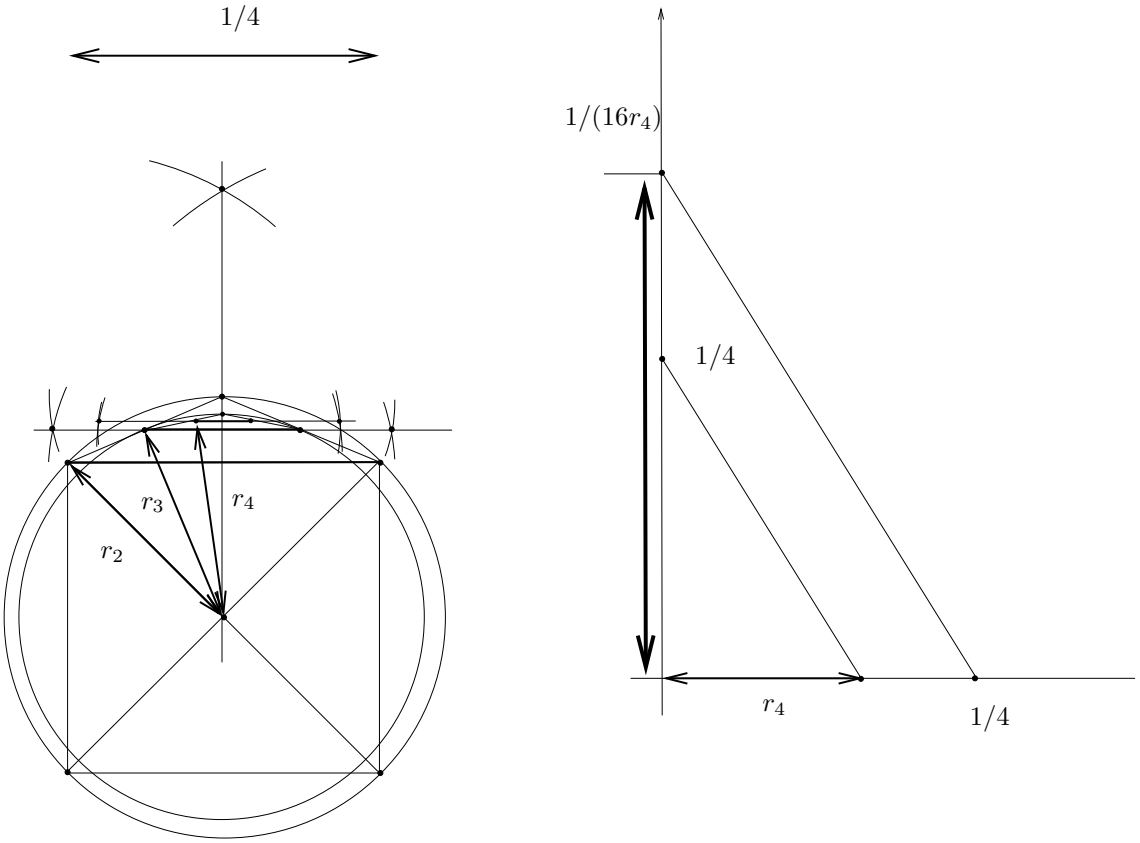


FIGURE AA.13. Construction de a_4 $1/(2a_2)$ à la règle et au compas par la méthode de Cues.

AA.3.3. Construction à la règle et au compas

On donnera en figure AA.13, la construction de a_4 et l'approximation de π en construisant $1/r_4$ (voir [Car84, p. 21]), puis $1/(2r_4)$. La construction du carré initial n'est pas faite et laissée au lecteur !

Le lecteur pourra vérifier que le segment représenté à droite représentant $1/(16r_4)$ vaut bien à peu près $\pi/8$ (puisque $1/(2r_4) \approx \pi$) !

AA.3.4. Étude géométrique

Sur la figure AA.11, on constate que par construction $OA' < OA$ et donc

$$\forall n \geq 2, \quad r_{n+1} < r_n. \tag{AA.58}$$

On a aussi $OH' > 0H$ et donc

$$\forall n \geq 2, \quad a_{n+1} > a_n. \tag{AA.59}$$

Par ailleurs, par construction, à chaque étape, le polygone construit est inscrit dans le cercle de rayon $[OA] = r_n$ et donc son périmètre est strictement inférieur au cercle de rayon r_n soit $1 < 2\pi r_n$, soit encore

$$\forall n \geq 2, \quad r_n > \frac{1}{2\pi}. \tag{AA.60}$$

De même, par construction, à chaque étape, le polygone construit est circonscrit au cercle de rayon $[OH] = a_n$ et donc son périmètre est strictement supérieur au cercle de rayon a_n soit $1 > 2\pi a_n$, soit encore

$$\forall n \geq 2, \quad a_n < \frac{1}{2\pi}. \tag{AA.61}$$

Notons que l'on a aussi (voir figure AA.11)

$$OA < OH + HA,$$

et donc

$$OA - OH < HA,$$

soit

$$\forall n \geq 2, \quad r_n - a_n < \frac{c_n}{2},$$

et donc, d'après (AA.56)

$$\forall n \geq 2, \quad r_n - a_n < \frac{1}{2^{n-1}}. \tag{AA.62}$$

AA.3.5. Preuve géométrique de la convergence des deux suites a_n et r_n vers $1/(2\pi)$

Bref, d'après (AA.59), (AA.58), et (AA.62), les deux suites a_n et r_n sont adjacentes et convergent vers une limite commune l. D'après (AA.60) et (AA.61), par passage à la limite quand n tend vers l'infini, on a

$$l \geq \frac{1}{2\pi}, \quad l \leq \frac{1}{2\pi},$$

et donc $l = 1/(2\pi)$, soit encore

$$\lim_{n\to+\infty} a_n = \frac{1}{2\pi} \tag{AA.63a}$$

$$\lim_{n\to+\infty} r_n = \frac{1}{2\pi}. \tag{AA.63b}$$

Là encore, nous avons fournit une preuve qu'auraient pu écrire Archimède ou Cues, en n'utilisant que les calculs précédents, fondés sur la géométrie élémentaire.

AA.3.6. Preuve analytique de la convergence des deux suites a_n et r_n vers $1/(2\pi)$ et qualité de la convergence

Montrons maintenant le dernier résultat !

Formellement, chaque polygone, de perimètre constant égal à 1, tend vers un cercle, de périmètre $1 = 2\pi l$ où l est la limite commune des suites a_n et r_n. On a donc $l = 1/(2\pi)$ ce qu'on montre maintenant analytiquement.

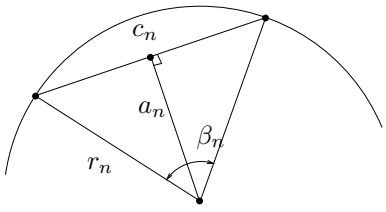


FIGURE AA.14. Les relations entre c_n, a_n et r_n et β_n, l'angle au sommet.

Nous avons des relations entre c_n, a_n et r_n (voir figure AA.14). D'après (AA.54), on a

$$\forall n \geq 2, \quad \beta_n = \frac{\pi}{2^{n-1}}. \tag{AA.64}$$

et, grâce aux relations trigonométriques habituelles :

$$\sin\frac{\beta_n}{2} = \frac{c_n}{2r_n},$$

$$\tan\frac{\beta_n}{2} = \frac{c_n}{2a_n},$$

et donc

$$c_n = 2r_n \sin\frac{\pi}{2^n},$$
$$c_n = 2a_n \tan\frac{\pi}{2^n}$$

et enfin, d'après (AA.56)

$$\frac{1}{2^n} = 2r_n \sin\frac{\pi}{2^n},$$
$$\frac{1}{2^n} = 2a_n \tan\frac{\pi}{2^n}.$$

Cela donne d'une part, pour $n = 2$:

$$\frac{1}{4} = 2r_2\frac{\sqrt{2}}{2},$$
$$\frac{1}{4} = 2a_2,$$

soit

$$a_2 = \frac{1}{8}, \quad r_2 = \frac{\sqrt{2}}{8}. \tag{AA.65}$$

et d'autre part,

$$\forall n \geq 2, \quad a_n = \frac{1}{2}\frac{\frac{1}{2^n}}{\tan\frac{\pi}{2^n}}, \tag{AA.66a}$$
$$r_n = \frac{1}{2}\frac{\frac{1}{2^n}}{\sin\frac{\pi}{2^n}}, \tag{AA.66b}$$

ce qui implique (AA.63). On a aussi, comme pour la méthode d'Archimède :

$$\frac{1}{2a_n} = \pi + O\left(\frac{1}{4^n}\right), \tag{AA.67a}$$
$$\frac{1}{2r_n} = \pi + O\left(\frac{1}{4^n}\right). \tag{AA.67b}$$

D'après (AA.59) et (AA.60), on a aussi

$$\forall n \geq 2, \quad \frac{1}{2r_n} < \pi < \frac{1}{2a_n}. \tag{AA.68}$$

AA.4. Et la méthode originale des isopérimètres de Descartes !

Concluons par une remarque sur la méthode originale des isopérimètres de Descartes présentée dans voir `http://www.pi314.net/fr/descartes.php`. N'en déplaisent à leur auteur ou à M. Descartes, cette méthode n'est rien d'autre que la méthode de Cues, présentée autrement, comme c'est dit dans `https://publimath.univ-irem.fr/glossaire/ME103.htm`. En effet, tentons d'éliminer la variable r_n dans les relations de récurrence (AA.57) afin de ne conserver que la suite des apothèmes a_n. Le raisonnement est même un peu plus simple que celui présenté ci-dessus, puisqu'il n'est même pas nécessaire d'utiliser l'équation (AA.51).

On renvoie de nouveau à la figure AA.11. Le théorème de Pythagore appliqué au triangle OAH fournit

$$OA^2 = OH^2 + HA^2,$$

soit encore avec les notations précédemment utilisées : pour tout $n \geq 2$,

$$r_n^2 = a_n^2 + \frac{c_n^2}{4},$$

et donc, d'après (AA.56), on a

$$r_n^2 = a_n^2 + \frac{1}{4.2^{2n}},$$

et donc

$$\forall n \geq 2, \quad r_n^2 = a_n^2 + \frac{1}{2^{2(n+1)}}. \tag{AA.69}$$

Si on utilise (AA.50) ou (AA.57a), on a par ailleurs

$$2a_{n+1} - a_n = r_n$$

et donc

$$4a_{n+1}^2 - 4a_{n+1}a_n + a_n^2 = r_n^2,$$

et en éliminant r_n d'après (AA.69), on a

$$4a_{n+1}^2 - 4a_{n+1}a_n + a_n^2 = a_n^2 + \frac{1}{2^{2(n+1)}}.$$

soit encore

$$\forall n \geq 2, \quad a_{n+1}^2 - a_{n+1}a_n - \frac{1}{2^{2(n+2)}} = 0, \tag{AA.70}$$

ce qui est bien l'équation du second degré en a_{n+1} présentées dans `http://www.pi314.net/fr/descartes.php`. En effet, dans cette dernière référence, on a l'équation (où a_n représente en fait r_n !) soit encore

$$\forall n \geq 2, \quad a_{n+1}^2 - a_{n+1}a_n - \frac{a_0^2}{4^{n+1}} = 0 \tag{AA.71}$$

où $a_0 = 1/8$ et (AA.71) est donc équivalente à

$$\forall n \geq 2, \quad a_{n+1}^2 - a_{n+1}a_n - \frac{1}{2^{2(n+4)}} = 0.$$

et en posant en remplaçant $n+4$ par $n+2$ (car nous partons de 2 et non de 0) :

$$\forall n \geq 2, \quad a_{n+1}^2 - a_{n+1}a_n - \frac{1}{2^{2(n+2)}} = 0,$$

ce qui est bien équivalent à (AA.70). La résolution de (AA.70) donne un discriminant égal à

$$\Delta = a_n^2 + \frac{4}{2^{2(n+2)}} = a_n^2 + \frac{1}{2^{2(n+1)}} > 0$$

et les racines de (AA.70) sont donc données par

$$X = \frac{1}{2}\left(a_n \pm \sqrt{a_n^2 + \frac{1}{2^{2(n+1)}}}\right),$$

et on ne conserve que la racine positive :

$$\forall n \geq 2, \quad a_{n+1} = \frac{1}{2}\left(a_n + \sqrt{a_n^2 + \frac{1}{2^{2(n+1)}}}\right), \tag{AA.72}$$

ce qui est bien la même expression que dans les références données ci-dessus. Puisque a_n tend vers $1/(2\pi)$ on retrouve aussi la convergence annoncée.

AA.5. Approximation quadratique par une méthode arithmético-géométrique

Dans `http://www.pi314.net/fr/salamin.php` est proposée la suite suivante, qui n'a, cette fois-ci, plus de construction à la règle et au compas :

$$\begin{aligned}
a_0 &= 1,\\
b_0 &= \frac{\sqrt{2}}{2},\\
u_0 &= 0,\\
v_0 &= 1,\\
a_{n+1} &= \frac{1}{2}\left(a_n + b_n\right),\\
b_{n+1} &= \sqrt{a_n b_n},\\
u_{n+1} &= \frac{1}{2}\left(u_n + v_n\right),\\
v_{n+1} &= \frac{1}{2b_{n+1}}\left(u_n b_n + v_n a_n\right),\\
w_n &= 2\sqrt{2}\frac{a_n^3}{u_n},
\end{aligned}$$

Les deux suites a_n et b_n sont adjacentes (suite arithmético-géométrique) et on peut montrer assez simplement (voir le lemme AA.8) que la convergence vers la limite commune est quadratique. De plus, on

$$w_n \to \pi,$$

et là encore de façon quadratique.

Lemme AA.8 (Suite arithmético-géométrique). *On définit les deux suites $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ par u_0, $v_0 \in \mathbb{R}_+^*$ et*

$$u_{n+1} = \sqrt{u_n v_n}, \quad v_{n+1} = \frac{u_n + v_n}{2}.$$

Alors les deux suites convergent vers la même limite. De plus, la convergence de chacune des deux suites est quadratique.

Démonstration. On pourra consulter par exemple `https://fr.wikipedia.org/wiki/Moyenne_arithm%C3%A9tico-g%C3%A9om%C3%A9trique`

On raisonne par étape.

(1) (a) Montrons tout d'abord, que, si a et b sont positifs,

$$\sqrt{ab} \leq \frac{a+b}{2}, \tag{AA.73}$$

ce qui s'obtient en élevant au carré :

$$4ab \leq a^2 + b^2 + 2ab,$$

équivalent à

$$0 \leq a^2 + b^2 - 2ab,$$

soit à

$$0 \leq (a-b)^2.$$

(b) On peut vérifier aisément par récurrence sur n que

$$\forall n, \quad u_n > 0, \quad v_n > 0, \tag{AA.74}$$

et, en même temps, que v_n est définie.

(c) Ainsi, (AA.73) appliquée à $a = u_n$ et $b = v_n$ donne

$$\forall n, \quad u_{n+1} \leq v_{n+1}$$

soit

$$\forall n \geq 1, \quad u_n \leq v_n. \tag{AA.75}$$

(d) De (AA.75), on déduit donc

$$\forall n \geq 1, \quad u_{n+1} = \sqrt{u_n v_n} \geq \sqrt{u_n u_n} = |u_n| = u_n,$$

et

$$\forall n \geq 1, \quad v_{n+1} = \frac{u_n + v_n}{2} \leq \frac{v_n + v_n}{2} = v_n,$$

On a donc

$$\forall n \geq 1, \quad u_n \leq u_{n+1} \leq v_{n+1} \leq v_n. \tag{AA.76}$$

(e) La suite u_n est donc croissante et majorée par v_0 et la suite v_n est donc décroissante et minorée par $u_0 > 0$. Ainsi (u_n) converge vers l et (v_n) vers $l' > 0$. Si on passe à la limite quand n tend vers l'infini dans la définition des deux suite, on obtient par continuité : $l = \sqrt{ll'}$ et $l' = (l + l')/2$, ces deux égalités impliquant que $l = l'$. De (AA.76), on déduit donc que les deux suites convergent vers $l > 0$ (difficile à calculer en fonction de u_0 et v_0) tel que

$$\forall n \geq 1, \quad u_n \leq u_{n+1} \leq l \leq v_{n+1} \leq v_n. \tag{AA.77}$$

En fait, les deux suites u_n et v_n sont donc adjacentes.

(2) (a) Par définition,

$$v_{n+1}^2 - u_{n+1}^2 = \frac{1}{4}\left(u_n^2 + v_n^2 + 2u_n v_n - 4u_n v_n\right) = \frac{1}{4}\left(u_n^2 + v_n^2 - 2u_n v_n\right),$$

et donc

$$\forall n, \quad v_{n+1}^2 - u_{n+1}^2 = \frac{1}{4}\left(v_n - u_n\right)^2. \tag{AA.78}$$

(b) On a donc

$$\forall n, \quad (v_{n+1} - u_{n+1})(v_{n+1} + u_{n+1}) = \frac{1}{4}\left(v_n - u_n\right)^2$$

et, grâce à (AA.74) :

$$\forall n, \quad v_{n+1} - u_{n+1} = \frac{1}{4(v_{n+1} + u_{n+1})}\left(v_n - u_n\right)^2. \tag{AA.79}$$

Grâce à (AA.77), on a

$$v_{n+1} + u_{n+1} \geq u_0 + l,$$

et (AA.79) implique

$$\forall n, \quad 0 \leq v_{n+1} - u_{n+1} \leq \frac{1}{4(u_0 + l)}\left(v_n - u_n\right)^2. \tag{AA.80}$$

(c) On a, selon (AA.77),

$$e_n = |u_n - l| + |v_n - l| = l - u_n + v_n - l = v_n - u_n,$$

et donc (AA.80) est équivalent à

$$\forall n, \quad e_{n+1} \leq \frac{1}{4(u_0 + l)} e_n^2. \tag{AA.81}$$

Notons aussi que

$$|u_n - l| \leq e_n,$$
$$|v_n - l| \leq e_n,$$

et donc que les résultats relatifs à e_n sont aussi valables pour les deux erreurs $u_n - l|$ et $|v_n - l|$. La convergence est donc d'ordre 2 et on renvoie à [DB21, chapitre "Équations non-linéaires"].

□

AA.6. Approximations d'ordres plus élevés

Voir `http://www.pi314.net/fr/borwein.php`

AA.7. Simulations numériques

Voir les fonctions matlab aux liens suivants :
`http://utbmjb.chez-alice.fr/Polytech/MNBif/fichiers_matlab/approximation_pi_archimede.m`,
`http://utbmjb.chez-alice.fr/Polytech/MNBif/fichiers_matlab/approximation_pi_cues.m`
`http://utbmjb.chez-alice.fr/Polytech/MNBif/fichiers_matlab/approximation_pi_quadratique.m`.

AA.7.1. Méthode d'Archimède

n	I_n	C_n
0	3.00000000000000000	3.46410161513775440
1	3.10582854123025020	3.21539030917347100
2	3.13262861328123690	3.15965994209749380
3	3.13935020304687210	3.14608621513140200
4	3.14103195089053070	3.14271459964531270
5	3.14145247228534430	3.14187304998012660
6	3.14155760791162210	3.14166274705480710
7	3.14158389214893590	3.14161017659978190
8	3.14159046323676170	3.14159703432307640
9	3.14159210604304920	3.14159374881711620
10	3.14159251658815460	3.14159292787337300
11	3.14159261864078940	3.14159272562285170
12	3.14159264532121570	3.14159267174128450
13	3.14159264532121570	3.14159261890114650
14	3.14159264532121570	3.14159267174128450
15	3.14159264532121570	3.14159193588223350
16	3.14159366984942690	3.14159267174128450
17	3.14159230381173770	3.14158100757962440
18	3.14160869622480380	3.14159267174128450
19	3.14158683965504170	3.14140615473788240
20	3.14167426502175800	3.14054349240083930
21	3.14167426502175800	3.14000686469122850
22	3.14307274017003960	3.13494537565859140
23	3.15980616494113460	3.14000686469122850
24	3.18198051533946420	3.22451524353455100

TABLE AA.1. Valeurs de n, de I_n et de C_n.

n	$\log(\lvert I_n - \pi\rvert)$	$\log(\lvert C_n - \pi\rvert)$
0	−0.84896	−0.49146
1	−1.44655	−1.13196
2	−2.04750	−1.74311
3	−2.64928	−2.34741
4	−3.25127	−2.95003
5	−3.85331	−3.55223
6	−4.45537	−4.15432
7	−5.05742	−4.75639
8	−5.65949	−5.35845
9	−6.26158	−5.96050
10	−6.86327	−6.56180
11	−7.45657	−7.14247
12	−8.08257	−7.74109
13	−8.08257	−7.45981
14	−8.08257	−7.74109
15	−8.08257	−6.14405
16	−5.99300	−7.74109
17	−6.45621	−4.93382
18	−4.79472	−7.74109
19	−5.23553	−3.72932
20	−4.08825	−2.97916
21	−4.08825	−2.79975
22	−2.82971	−2.17736
23	−1.73961	−2.79975
24	−1.39375	−1.08133

TABLE AA.2. Valeurs de n, de $\log(\lvert I_n - \pi\rvert)$ et de $\log(\lvert C_n - \pi\rvert)$.

On utilisant les formules (AA.18) et (AA.19), on a déterminé numériquement les valeurs de I_n et de C_n, ainsi que les logarithmes (décimaux) des erreurs ; voir les tableaux AA.1 et AA.2. On constate sur ces tableaux qu'à partir d'un certain rang ($n = 12$), l'erreur remonte, et ce à cause des arrondis de calcul. On constate aussi que (AA.36a) est vraie mais uniquement jusqu'à cette valeur de n.

Si on utilise cette fois-ci les formules (AA.20) et (AA.21), les calculs ont faits et présentés dans les tableaux AA.3 et AA.4. L'erreur augmente pour une valeur de n un peu plus élevés. Grâce aux logarithmes, on a fait une interpolation pour retrouver les formules (AA.46) ; par interpolation, on obtient les deux valeurs de A tel que $\log(\lvert I_n - \pi\rvert) \approx B/n^A$ et $\log(\lvert C_n - \pi\rvert) \approx B'/n^{A'}$. Numériquement, on obtient en effet

$$A = 0.249641,$$
$$A' = 0.247349.$$

Voir aussi la figure AA.15.

n	I_n	C_n
0	3.00000000000000000	3.46410161513775440
1	3.10582854123024890	3.21539030917347150
2	3.13262861328123470	3.15965994209750670
3	3.13935020304685210	3.14608621513144280
4	3.14103195089047560	3.14271459964548860
5	3.14145247228556150	3.14187304997995120
6	3.14155760791205640	3.14166274705671980
7	3.14158389214893630	3.14161017661176520
8	3.14159046322286570	3.14159703431109350
9	3.14159210604304830	3.14159374871793110
10	3.14159251681049150	3.14159292752788440
11	3.14159262041948390	3.14159272241094990
12	3.14159264532121530	3.14159266783840650
13	3.14159264532121530	3.14159262280402450
14	3.14159264532121530	3.14159266783840650
15	3.14159275915769950	3.14159285047699610
16	3.14159230381173730	3.14159175714657390
17	3.14159412519519070	3.14159649324559310
18	3.14158683965504080	3.14157718609415330
19	3.14155769732548420	3.14153820867771390
20	3.14167426502175710	3.14181032725824180
21	3.14120796828226560	3.14060572479268090
22	3.14493640635228110	3.14927305962702600
23	3.15238005322962240	3.15549011211397310
24	3.12249899919919870	3.08985281321896910

TABLE AA.3. Valeurs de n, de I_n et de C_n (utilisation des formules (AA.20) et (AA.21)).

AA.7.2. Méthode de Cues

On utilise les formules (AA.57) et (AA.65), on a déterminé numériquement les valeurs de $1/(2r_n)$ et de $1/(2a_n)$, ainsi que les logarithmes (décimaux) des erreurs ; voir les tableaux AA.5 et AA.6. Cette fois-ci, contrairement à la méthode d'Archimède, le comportement numérique est meilleur : l'erreur se stabilise au lieu de remonter à partir d'un certain rang, ce qui confirme ce que l'on a observé dans la construction à la règle et au compas, vue plus haut. De plus, la formule (AA.68) reste numériquement vraie. Grâce aux logarithmes, on a fait une interpolation pour retrouver les formules (AA.67) ; par interpolation, on obtient les deux valeurs de A tel que $\log(\pi - 1/(2r_n)) \approx B/n^A$ et $\log(1/(2a_n) - \pi) \approx B'/n^{A'}$. Numériquement, on obtient en effet

$$A = 0.248464,$$
$$A' = 0.252042.$$

On constate à la fin du tableau AA.6 des valeurs complexes du logarithme, due à son argument négatif, à cause des erreurs de calcul, ce qui est normal sous matlab qui calcule le logarithme complexe (Voir [Bas22c, Chapitre "Séries entières et fonctions usuelles sur $\mathbb{C}$"]).

n	$\log(\lvert I_n - \pi \rvert)$	$\log(\lvert C_n - \pi \rvert)$
0	−0.84896	−0.49146
1	−1.44655	−1.13196
2	−2.04750	−1.74311
3	−2.64928	−2.34741
4	−3.25127	−2.95003
5	−3.85331	−3.55223
6	−4.45537	−4.15432
7	−5.05742	−4.75639
8	−5.65948	−5.35845
9	−6.26158	−5.96054
10	−6.86398	−6.56235
11	−7.47925	−7.16228
12	−8.08257	−7.84623
13	−8.08257	−7.51165
14	−8.08257	−7.84623
15	−6.97647	−6.70578
16	−6.45621	−6.04748
17	−5.83221	−5.41571
18	−5.23553	−4.81058
19	−4.45647	−4.26404
20	−4.08825	−3.66219
21	−3.41489	−3.00571
22	−2.47577	−2.11462
23	−1.96708	−1.85706
24	−1.71911	−1.28617

TABLE AA.4. Valeurs de n, de $\log(\lvert I_n - \pi \rvert)$ et de $\log(\lvert C_n - \pi \rvert)$ (utilisation des formules (AA.20) et (AA.21)).

AA.7.3. Approximation quadratique

Voir les tableaux AA.7 et AA.8 qui montrent la convergence très rapide de w_n vers π. On peut aussi mettre en évidence l'aspect quadratique de la convergence en passant par le logarithme et en utilisant la technique de [Bas21a, un des exercices du TD "Équations non-linéaires"] et son corrigé [Bas21b]. On obtient la valeur numérique suivante de l'ordre : 2.0691, ce qui confirme l'aspect quadratique de la méthode.

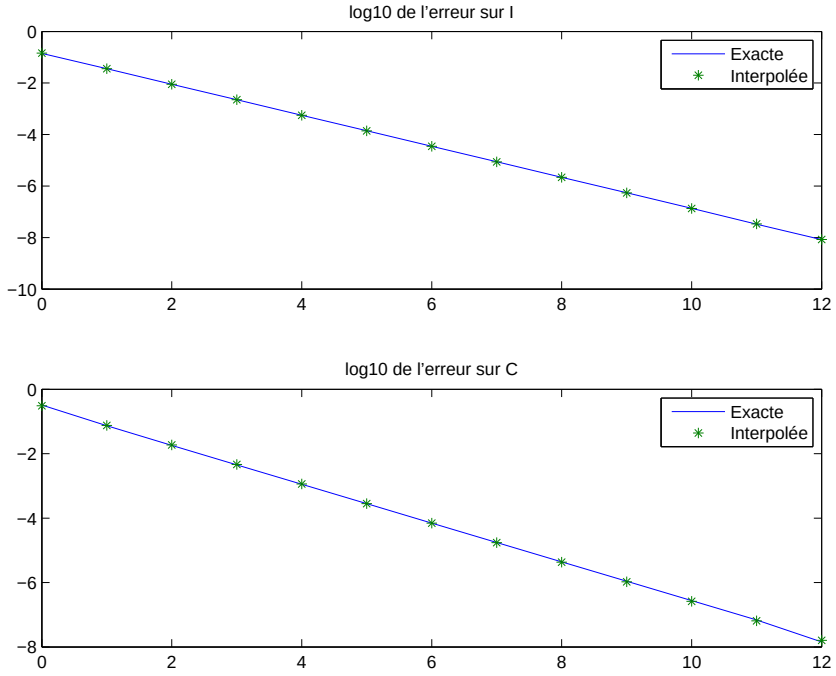


FIGURE AA.15. Estimation de l'erreur pour le calcul de I_n et de C_n.

n	$1/(2r_n)$	$1/(2a_n)$
0	2.82842712474618980	4.00000000000000000
1	3.06146745892071830	3.31370849898476070
2	3.12144515225805200	3.18259787807452810
3	3.13654849054593890	3.15172490742925640
4	3.14033115695475250	3.14411838524590430
5	3.14127725093277290	3.14222362994245640
6	3.14151380114430090	3.14175036916896570
7	3.14157294036709130	3.14163208070318190
8	3.14158772527716000	3.14160251025680900
9	3.14159142151120020	3.14159511774958930
10	3.14159234557011800	3.14159326962930720
11	3.14159257658487290	3.14159280759964420
12	3.14159263433856270	3.14159269209225480
13	3.14159264877698610	3.14159266321540850
14	3.14159265238659160	3.14159265599619710
15	3.14159265328899330	3.14159265419139460
16	3.14159265351459330	3.14159265374019370
17	3.14159265357099260	3.14159265362739280
18	3.14159265358509290	3.14159265359919270
19	3.14159265358861810	3.14159265359214280
20	3.14159265358949870	3.14159265359038020
21	3.14159265358971900	3.14159265358993920
22	3.14159265358977450	3.14159265358982910
23	3.14159265358978820	3.14159265358980200
24	3.14159265358979130	3.14159265358979490
25	3.14159265358979270	3.14159265358979310
26	3.14159265358979270	3.14159265358979270
27	3.14159265358979270	3.14159265358979270
28	3.14159265358979270	3.14159265358979270
29	3.14159265358979270	3.14159265358979270

TABLE AA.5. Valeurs de n, de $1/(2r_n)$ et de $1/(2a_n)$.

n	$\log(\pi - 1/(2r_n))$	$\log(1/(2a_n) - \pi)$
0	-0.50423	-0.06631
1	-1.09623	-0.76418
2	-1.69578	-1.38716
3	-2.29721	-1.99429
4	-2.89911	-2.59761
5	-3.50113	-3.19999
6	-4.10318	-3.80213
7	-4.70524	-4.40421
8	-5.30730	-5.00627
9	-5.90936	-5.60833
10	-6.51142	-6.21039
11	-7.11348	-6.81245
12	-7.71554	-7.41451
13	-8.31760	-8.01657
14	-8.91966	-8.61863
15	-9.52172	-9.22069
16	-10.12378	-9.82275
17	-10.72583	-10.42482
18	-11.32788	-11.02689
19	-11.92994	-11.62899
20	-12.53102	-12.23130
21	-13.12981	-12.83533
22	-13.72928	-13.44404
23	-14.31114	-14.05150
24	-14.75047	-14.75047
25	-15.35253	$-Inf$
26	-15.35253	$-15.35253 + 1.36438i$
27	-15.35253	$-15.35253 + 1.36438i$
28	-15.35253	$-15.35253 + 1.36438i$
29	-15.35253	$-15.35253 + 1.36438i$

TABLE AA.6. Valeurs de n, de $\log(\pi - 1/(2r_n))$ et de $\log(1/(2a_n) - \pi)$.

n	w_n
1	3.51776695296636890
2	3.14278210836401910
3	3.14159266157735220
4	3.14159265358979490
5	3.14159265358979490
6	3.14159265358979490
7	3.14159265358979490
8	3.14159265358979490
9	3.14159265358979490

TABLE AA.7. Valeurs de n et de w_n

n	$\log(\lvert w_n - \pi \rvert)$
1	-0.42461
2	-2.92465
3	-8.09759
4	-14.75047
5	-14.75047
6	-14.75047
7	-14.75047
8	-14.75047
9	-14.75047

TABLE AA.8. Valeurs de n, de $\log(\lvert w_n - \pi \rvert)$.

Bibliographie

[Bac20] N. Bacaër. *Un modèle mathématique des débuts de l'épidémie de coronavirus en France.* hal-02509142v5 dans HAL. disponible sur `https://hal.archives-ouvertes.fr/hal-02509142v5`. 2020.

[Bar77] J. Baranger. *Introduction à l'analyse numérique.* Ouvrage disponible à la bibliothèque de Mathématiques de Lyon 1 (cote : 518.07 BAR, niveau Capes/Agreg). Hermann, Paris, 1977, pages vii+133.

[Bas11] J. Bastien. *Applications de l'algèbre et de l'analyse à la géométrie.* Notes de cours de l'UV MT25 de l'UTBM, disponible sur le web : `http://utbmjb.chez-alice.fr/UTBM/index.html`, rubrique MT25. 2011. 180 pages.

[Bas12] J. Bastien. *Introduction à la statistisque descriptive.* Notes de cours de statistiques du M1 APA de l'UFRSTAPS de Lyon 1, disponible sur le web : `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html`, rubrique M1 APA. 2012. 127 pages.

[Bas14a] J. Bastien. *Savoir se méfier des ordinateurs et de la science.* Transparents de l'UE Zététique de l'INSA de Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/INSA/index.html`. 2014. 38 pages.

[Bas14b] J. Bastien. *Vérité mathématique : paradoxe, preuve et conventions. Se méfier de ses réflexes et de ses habitudes.* Transparents de l'UE Zététique de l'INSA de Lyon. 2014. 80 pages.

[Bas18] J. Bastien. *Biomécanique du mouvement.* Tutorat de l'UE Biomécanique (L2) de l'UFRSTAPS de Lyon 1, disponibles sur le web : `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html`, rubrique L2 Bioméca. 2018. 93 pages.

[Bas21a] J. Bastien. *Méthodes numériques de base.* Travaux Dirigés de l'UV MNB (Département Informatique) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Informatique 3A : Méthodes Numériques de Base". 2021. 23 pages.

[Bas21b] J. Bastien. *Méthodes numériques de base.* Corrigés des Travaux Dirigés de l'UV MNB (Département Informatique) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Informatique 3A : Méthodes Numériques de Base". 2021. 91 pages.

[Bas22a] J. Bastien. *Mathématiques Fondamentales pour l'Informatique.* Notes de cours de l'UV MFI (Département Informatique) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Informatique 3A : Mathématiques Fondamentales pour l'Informatique". 2022. 270 pages.

[Bas22b] J. Bastien. *Mathématiques Fondamentales pour l'Ingénieur.* Notes de cours de l'UV MFImater (Département Matériaux) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Matériaux 3A : Mathématiques Fondamentales pour l'Ingénieur". 2022. 127 pages.

[Bas22c] J. Bastien. *Outils Mathématiques pour l'Ingénieur 3.* Notes de cours de l'UV OMI3 (Département Mécanique) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Mécanique 4A : Outils Mathématiques pour l'Ingénieur 3". 2022. 269 pages.

[Bas78] J. Bass. *Cours de mathématiques, tome 2.* Masson, 1978.

[BD65] W. E. Boyce et R. C. DiPrima. *Elementary differential equations and boundary value problems.* John Wiley & Sons, Inc., New York-London-Sydney, 1965, pages xi+485.

[BM03] J. Bastien et J.-N. Martin. *Introduction à l'analyse numérique. Applications sous Matlab.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 519.4 BAS, 4e étage). Voir `https://www.dunod.com/sciences-techniques/introduction-analyse-numerique-applications-sous-matlab`. Paris : Dunod, 2003. 392 pages.

[Bré19a] C.-E. Bréhier. *Contrôle 2 du TD d'Introduction à l'Analyse Numérique (L2).* Disponible sur le web : `http://math.univ-lyon1.fr/~brehier/teaching/2018-2019/introduction-analyse-numerique-l2`. 2019.

[Bré19b] C.-E. Bréhier. *TD d'Introduction à l'Analyse Numérique (L2).* Disponible sur le web : `http://math.univ-lyon1.fr/~brehier/teaching/2018-2019/introduction-analyse-numerique-l2`. 2019.

[Cam10] J.-B. Campesato. "Sur le problème de la tétration infinie ou infinite power tower". disponible sur `http://citron.9grid.fr/docs/tetration.pdf`. 2010.

[Car84] J.-C. Carrega. *Théorie des corps. La règle et le compas.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 512.3 CAR, 4e étage). Paris : Hermann, 1984.

[CB18] S. D. Conte et C. de Boor. *Elementary numerical analysis.* Tome 78. Classics in Applied Mathematics. An algorithmic approach, Updated with MATLAB, Reprint of the third (1980) edition, For the 1965 edition see [MR0202267]. Society for Industrial et Applied Mathematics (SIAM), Philadelphia, PA, 2018, pages xxiv+456.

[CB81] D. Conte et C. de Boor. *Elementary numerical analysis. An algorithmic approach.* Mc Graw-Hill, 1981.

[Cia82] P. G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Ouvrage disponible à la bibliothèque de Mathématiques de Lyon 1 (cote : 65 CIARLET, niveau -1). Paris : Masson, 1982, pages xii+279.

[CM84] M. Crouzeix et A. L. Mignot. *Analyse numérique des équations différentielles.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1984, pages viii+171.

[CM86] M. Crouzeix et A. L. Mignot. *Exercices d'analyse numérique des équations différentielles.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1986, page 183.

[DB21] N. Débit et J. Bastien. *Méthodes numériques de base.* Notes de cours de l'UV MNB (Département Informatique) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Informatique 3A : Méthodes Numériques de Base". 2021. 288 pages.

[DB22] N. Débit et J. Bastien. *Méthodes numériques de base.* Notes de cours de l'UV MNB (Département Matériaux) de Polytech Lyon, disponible sur le web : `http://utbmjb.chez-alice.fr/Polytech/index.html`, rubrique "Matériaux 3A : Méthodes Numériques de Base". 2022. 326 pages.

[DHB13] O. Diekmann, H. Heesterbeek et T. Britton. *Mathematical tools for understanding infectious disease dynamics.* Princeton Series in Theoretical and Computational Biology. Princeton University Press, Princeton, NJ, 2013, pages xiv+502.

[HM13] F. Holweck et J.-N. Martin. *Géométries pour l'ingénieur.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 516.07 HOL, 4e étage). Paris : Ellipses, 2013.

[Kno81] R. A. Knoebel. "Exponentials reiterated". In : *The American Mathematical Monthly* 04 (1981), pages 235–252.

[LT93] P. Lascaux et R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur. Tome 1.* Second. Ouvrage disponible à la bibliothèque de Mathématiques de Lyon 1 (cote : 518.2 LAS, niveau Capes/Agreg). Masson, Paris, 1993, pages xxiv+327.

[Mer10] D.-J. Mercier. *Cahiers de mathématiques du supérieur. Statistiques, probabilités, homothéties.* Tome 1. Publibook, 2010.

[Nou93] J. Nougier. *Méthodes de calculs numériques.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : SCI 1351). Paris : Masson, 1993.

[QSS00] A. Quarteroni, R. Sacco et F. Saleri. *Méthodes numériques pour le calcul scientifique, Programmes en matlab.* Springer, 2000.

[RDO87] E. Ramis, C. Deschamps et J. Odoux. *Cours de mathématiques spéciales. 4. Séries et équations différentielles.* 2e édition. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 515 RAM, 4e étage). Paris : Masson, 1987, pages VIII+314.

[RDO88] E. Ramis, C. Deschamps et J. Odoux. *Cours de mathématiques spéciales. 3. Topologie et éléments d'analyse.* 2e édition. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 510.7 RAM, 4e étage). Masson, Paris, 1988, pages VIII+362.

[RDO93] E. Ramis, C. Deschamps et J. Odoux. *Cours de mathématiques spéciales. Vol. 1.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 510.7 RAM, 1er étage). Masson, Paris, 1993, pages viii+440.

[Rey98] A. Rey. *Dictionnaire historique de la langue française.* Paris : Dictionnaires Le Robert, 1998.

[Sch01] M. Schatzmann. *Analyse numérique, une approche mathématique, Cours et exercices.* Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 518.1 SCH, 4e étage). Dunod, 2001.

[Sch87] F. Scheid. *Analyse numérique, Cours et problèmes.* Série Schaum. Ouvrage disponible à la bibliothèque Sciences de Lyon 1 (cote : 519.407 SCH, 4e étage). groupe Mc Graw-Hill, 1987.