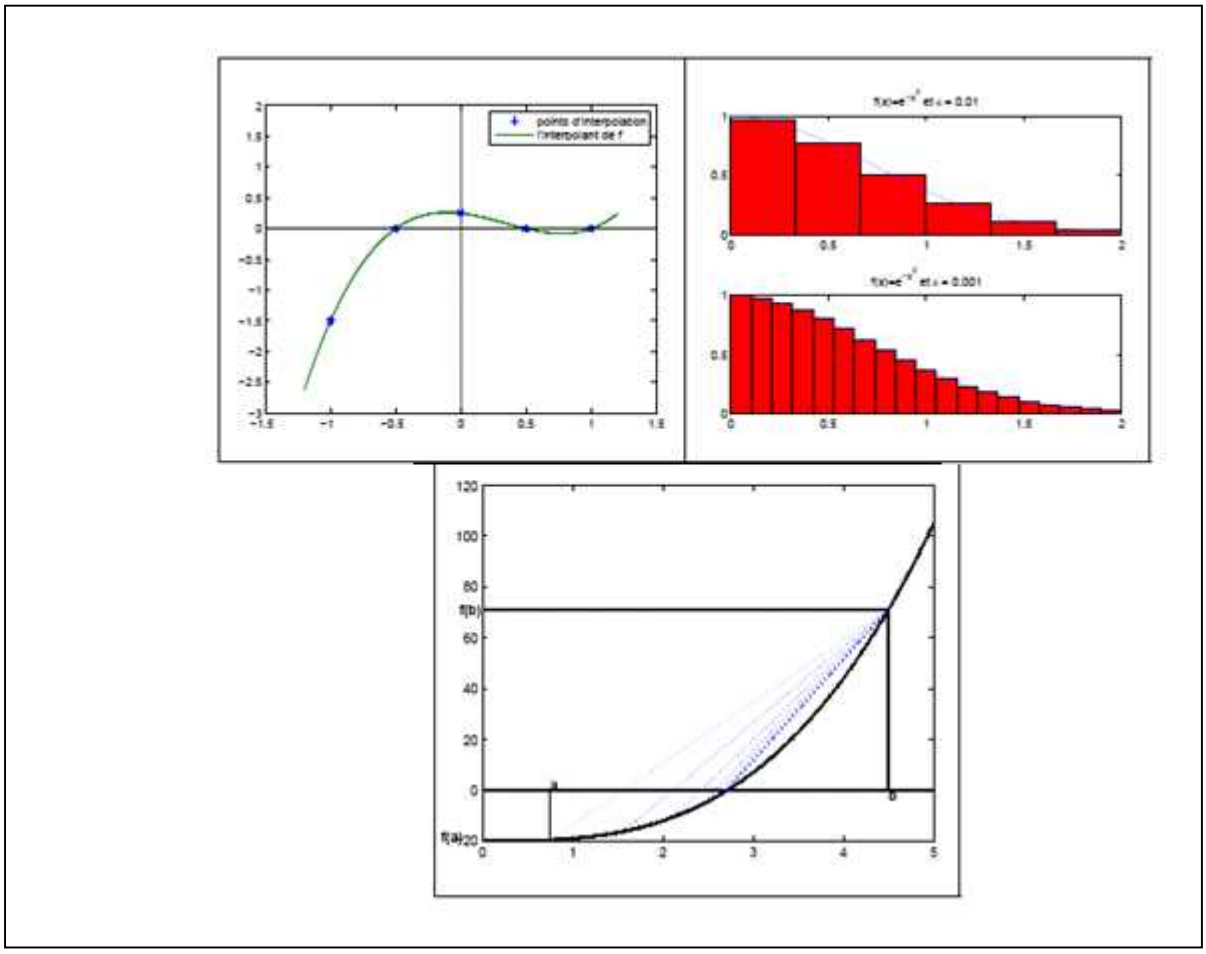
UNIVERSITE MOHAMED PREMIER
FACULTE DES SCIENCES OUJDA
Département de mathématiques et informatique

POLYCOPIE ANALYSE NUMERIQUE

Section SMA-SMI /S2



Abdesslam Boutayeb **Mohamed Derouich**

Année universitaire : 2010/ 2011

UNIVERSITE MOHAMED 1 er
FACULTE DES SCIENCES D'OUJDA-MAROC

Table des matières

Table des figures

Chapitre 1

RAPPELS

1.1 IDENTITES ALGEBRIQUES :

✓ $(a+b)^2 = a^2 + 2ab + b^2$

✓ $(a-b)^2 = a^2 - 2ab + b^2$

✓ $a^2 - b^2 = (a - b(a+b)$

✓ $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

✓ $a^3 - b^3 = (a-b)(a^2 + ab + b^2)$

✓ $a^3 + b^3 = (a+b)(a^2 - ab + b^2)$

✓ $(a+b)^4 = a^4 + 4a^3b^2 + 6a^2b^2 + 4ab^3 + b^4$

✓ $(a+b)^n = a^n + C_n^1 a^{n-1}b + ... + C_n^k a^{n-k}b^k + ... + b^n \quad \left[\text{ avec } C_n^k = \dfrac{n!}{k!(n-k)!} \right]$

✓ $1 + 2 + 3 + ... + (n-1) + n = \dfrac{n(n+1)}{2}$

✓ $a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + ... + ab^{n-2} + b^{n-1}$

✓ $a^{2m+1} + b^{2m+1} = (a+b)(a^{2m} - a^{2m-1}b + ... - ab^{2m-1} + b^{2m})$

✓ Cas particuliers (a=1) $1 - x^{n+1} = (1-x)(1 + x + x^2 + ... + x^n)$

1.2 TRIGONOMETRIE :

1.2.1 valeurs usuelles des fonctions trigonométriques :

1.2.2 Périodes et symétries :

✓ $\sin(x + 2\pi) = \sin x$

✓ $\sin(x + \pi) = -\sin x$

θ	$\sin\theta$	$\cos\theta$	$\text{tg}\,\theta$	$\text{cotg}\,\theta$
$0 = 0^\circ$	0	1	0	-
$\frac{\pi}{6} = 30^\circ$	$1/2$	$\sqrt{3}/2$	$\sqrt{3}/3$	$\sqrt{3}$
$\frac{\pi}{4} = 45^\circ$	$\sqrt{2}/2$	$\sqrt{2}/2$	1	1
$\frac{\pi}{3} = 60^\circ$	$\sqrt{3}/2$	$1/2$	$\sqrt{3}$	$\sqrt{3}/3$
$\frac{\pi}{2} = 90^\circ$	1	0	-	0

✓ $\sin(x + \frac{\pi}{2}) = \cos x$
✓ $\sin(-x) \quad = -\sin x$
✓ $\sin(\frac{\pi}{2} - x) = \cos x$
✓ $\cos(x + 2\pi) = \cos x$
✓ $\cos(x + \pi) \ = -\cos x$
✓ $\cos(x + \frac{\pi}{2}) \ = -\sin x$
✓ $\cos(-x) \quad = \cos x$
✓ $\cos(\frac{\pi}{2} - x) \ = \sin x$
✓ $tg(x + 2\pi) = tgx$
✓ $tg(x + \pi) \ = tgx$
✓ $tg(x + \frac{\pi}{2}) \ = -\cot gx$
✓ $tg(-x) \quad = -tgx$
✓ $tg(\frac{\pi}{2} - x) \ = \cot gx$
✓ $ch(x) = \frac{e^x + e^{-x}}{2}$
✓ $sh(x) = \frac{e^x - e^{-x}}{2}$
✓ $th(x) = \frac{sh(x)}{ch(x)}$
✓ $coth(x) = \frac{ch(x)}{sh(x)}$

1.2.3 Formules trigonométriques

✓ $\cos(n\pi) = (-1)^n$
✓ $\sin^2(\frac{n\pi}{2}) = \frac{1 + (-1)^{n+1}}{2}$

✓ $\cos(x+y) = \cos x \cos y - \sin x \sin y$

✓ $\cos(x-y) = \cos x \cos y + \sin x \sin y$

✓ $\sin(x+y) = \sin x \cos y + \cos x \sin y$

✓ $\sin(x-y) = \sin x \cos y - \cos x \, siny$

✓ $tg(x+y) = \dfrac{tgx + tgy}{1 - tgxtgy}$

✓ $tg(x-y) = \dfrac{tgx - tgy}{1 + tgxtgy}$

✓ $\begin{aligned}\cos(2x) &= \cos^2 x - \sin^2 x \\ &= 2cos^2 x - 1 \\ &= 1 - 2sin^2 x\end{aligned}$

✓ $\sin(2x) = 2 \sin x \cos x$

✓ $tg(2x) = \dfrac{2tgx}{1 - tg^2 x}$

1.3 DEVELOPPEMENTS LIMITES

✓ $\sin(x) = x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} + ... + (-1)^p \dfrac{x^{2p+1}}{(2p+1)!} + o(x^{2p+1})$

✓ $\cos(x) = 1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} + ... + (-1)^p \dfrac{x^{2p}}{(2p)!} + o(x^{2p})$

✓ $\tan(x) = x + x^3 + 2\dfrac{x^5}{15} + o(x^6)$

✓ $sh(x) = x + \dfrac{x^3}{3!} + \dfrac{x^5}{5!} + ... + \dfrac{x^{2p+1}}{(2p+1)!} + o(x^{2p+1})$

✓ $ch(x) = 1 + \dfrac{x^2}{2!} + \dfrac{x^4}{4!} + ... + \dfrac{x^{2p}}{2p!} + o(x^{2p})$

✓ $th(x) = x - \dfrac{x^3}{3!} + \dfrac{2x^5}{5!} + o(x^6)$

✓ $\dfrac{1}{(1-x)} = 1 + x + ... + x^n + o(x^n)$

✓ $\dfrac{1}{(1+x)} = 1 - x + ... + (-1)^n x^n + o(x^n)$

✓ $\ln(1+x) = x - \dfrac{x^2}{2} + ... + (-1)^{n+1} \dfrac{x^n}{n} + o(x^n)$

✓ $\ln(1-x) = -x - \dfrac{x^2}{2} - ... - \dfrac{x^n}{n} + o(x^n)$

✓ $\exp(x) = 1 + x + \dfrac{x^2}{2!} + ... + \dfrac{x^n}{n!} + o(x^n)$

✓ $(1+x)^{\alpha} = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2}x^2 + \cdots + \frac{\alpha(\alpha-1)(\alpha-2.)\cdots(\alpha-n+1)x^n}{n!} + o(x^n)$

✓ $\arctan(x) = x - \frac{x^3}{3} + ... + \frac{(-1)^n x^{2n+1}}{(2n+1)} + o(x^{2n+1})$

✓ $\arcsin(x) = x^3 + \frac{x^3}{2\times 3} + \frac{1\times 3}{2^2\times 2!}\times\frac{x^5}{5} + ... + \frac{1\times 3\times\cdots\times(2n-1)}{2^n n!}\times\frac{x^{2n+1}}{2n+1} + o(x^{2n+1})$

1.4 INTEGRATION

1) $\int udv = uv - \int vdv$

2) $\int a^n du = \frac{a^u}{\ln a} + c, a \neq 1, a > 0$

3) $\int \cos u du = \sin u + c$

4) $\int \sin u du = -\cos u + c$

5) $\int (ax+b)^n dx = \frac{(ax+b)^{n+1}}{a(n+1)} + c, n \neq 1$

6) $\int (ax+b)^{-1} dx = \frac{1}{a}\ln|ax+b| + c$

7) $\int x(ax+b)^n dx = \frac{(ax+b)^{n+1}}{a^2}[\frac{ax+b}{n+2} - \frac{b}{ax+b}] + c$

8) $\int x(ax+b)^{-1} dx = \frac{x}{a} - \frac{b}{a^2}\ln|ax+b| + c$

9) $\int x(ax+b)^{-2} dx = \frac{1}{a^2}[\ln|ax+b| + \frac{b}{ax+b}] + c$

10) $\int \frac{dx}{x(ax+b)} = \frac{1}{b}\ln|\frac{x}{ax+b}| + c$

11) $\int (\sqrt{ax+b})^n dx = \frac{2}{a}\frac{\left(\sqrt{ax+b}\right)^{n+2}}{n+2} + c, n = -2$

12) $\int \frac{\sqrt{ax+b}}{x} dx = 2\sqrt{ax+b} + b\int \frac{x}{x\sqrt{ax+b}}$

13) $(a)\int \frac{dx}{x\sqrt{ax+b}} = \frac{2}{\sqrt{-b}}\tan^{-1}\sqrt{\frac{ax+b}{-b}} + c, b < 0$

(b) $\int \frac{dx}{x\sqrt{ax+b}} = \frac{1}{\sqrt{b}}\ln|\frac{\sqrt{ax+b}-\sqrt{b}}{\sqrt{ax+b}+\sqrt{b}}| + c, b > 0$

14) $\int \frac{\sqrt{ax+b}}{x^2} dx = -\frac{\sqrt{ax+b}}{x} + \frac{a}{2}\int \frac{dx}{x\sqrt{ax+b}} + c$

15) $\int \frac{x}{x^2\sqrt{ax+b}} = -\frac{\sqrt{ax+b}}{dx} - \frac{a}{2b}\int \frac{dx}{x\sqrt{ax+b}} + c$

16) $\int \frac{dx}{a^2+b^2} = \frac{1}{a}\tan^{-1}\frac{x}{a} + c$

17) $\int \frac{dx}{(a^2+x^2)^2} = \frac{x}{2a^2(a^2+x^2)} + \frac{1}{2a^3}\tan^{-1}\frac{x}{a} + c$

18) $\int \frac{dx}{a^2-x^2} = \frac{1}{2a}\ln|\frac{x+a}{x-a}| + c$

19) $\int \frac{dx}{(a^2-x^2)^2} = \frac{x}{2a^2(a^2-x^2)} + \frac{1}{2a^2}\int \frac{dx}{a^2-x^2}$

20) $\int \frac{dx}{\sqrt{a^2+x^2}} = \sinh^{-1}\frac{x}{a} + c = \ln|x+\sqrt{a^2+x^2}| + c$

21) $\int \sqrt{a^2+x^2}dx = \frac{x}{2}\sqrt{a^2+x^2} + \frac{a2}{2}\sinh^{-1}\frac{x}{a} + c$

22) $\int x^2\sqrt{a^2+x^2}dx = \frac{x(a^2+2x^2)\sqrt{a^2+x^2}}{8} - \frac{a^4}{8}\sinh^{-1}\frac{x}{a} + c$

23) $\int \frac{\sqrt{a^2+x^2}}{x}dx = \sqrt{a^2+x^2} - a\sinh^{-1}|\frac{a}{x}| + c$

24) $\int \frac{\sqrt{a^2+x^2}}{x^2}dx = \sinh^{-1}\frac{x}{a} - \frac{\sqrt{a^2+x^2}}{x} + c$

25) $\int \frac{x^2}{\sqrt{a^2+x^2}}dx = -\frac{a2}{2}\sinh^{-1}\frac{x}{a} + \frac{x\sqrt{a^2+x^2}}{2} + c$

26) $\int \frac{dx}{x\sqrt{a^2+x^2}} = -\frac{1}{a}\ln|\frac{a+\sqrt{a^2+x^2}}{x}| + c$

27) $\int \frac{dx}{x^2\sqrt{a^2+x^2}} = -\frac{\sqrt{a^2+x^2}}{a^2x} + c$

28) $\int \frac{dx}{\sqrt{a^2-x^2}} = \sin^{-1}\frac{x}{a} + c$

29) $\int \sqrt{a^2-x^2}dx = \frac{x}{2}\sqrt{a^2-x^2} + \frac{a^2}{2}\sin^{-1}\frac{x}{a} + c$

30) $\int x^2\sqrt{a^2-x^2}dx = \frac{a^4}{8}\sin^{-1}$

31) $\int \frac{\sqrt{a^2-x^2}}{x}dx = \sqrt{a^2-x^2} - a\ln|\frac{a+\sqrt{a^2-x^2}}{x}| + c$

32) $\int \frac{\sqrt{a^2-x^2}}{x^2}dx = -\sin^{-1}\frac{x}{a} - \frac{\sqrt{a^2-x^2}}{x} + c$

33) $\int \frac{x^2}{\sqrt{a^2-x^2}}dx = \frac{a^2}{2}\sin^{-1}\frac{x}{a} - \frac{1}{2}x\sqrt{a^2-x^2} + c$

34) $\int \frac{dx}{x\sqrt{a^2-x^2}}dx = \frac{-1}{a}\ln|\frac{a+\sqrt{a^2-x^2}}{x}| + c$

35) $\int \frac{dx}{x^2\sqrt{a^2-x^2}}dx = \frac{-\sqrt{a^2-x^2}}{a^{2x}} + c$

36) $\int \frac{dx}{\sqrt{x^2-a^2}}dx = \cosh^{-1}\frac{x}{a} + c = \ln|x+\sqrt{x^2-a^2}| + c$

37) $\int \sqrt{x^2-a^2}dx = \frac{x}{2}\sqrt{x^2-a^2} - \frac{a^2}{2}\cosh^{-1}\frac{x}{a} + c$

38) $\int (\sqrt{x^2-a^2})^n dx = \frac{x(\sqrt{x^2-a^2})^{n^n}}{n+1} - \frac{na^2}{n+1}\int (\sqrt{x^2-a^2})^{n-2}dx, n \neq -1$

39) $\int \frac{dx}{(\sqrt{x^2-a^2})^n}dx = \frac{x(\sqrt{x^2-a^2})^{n-2}}{(2-n)a^2} - \frac{n-3}{(n-2)a^2}\int \frac{dx}{(\sqrt{x^2-a^2})^{n-2}}dx$

40) $\int x(\sqrt{x^2-a^2})^n dx = \frac{(\sqrt{x^2-a^2})^{n+2}}{(2-n)a^2} + c, n \neq -2$

41) $\int x^2\sqrt{x^2-a^2}dx = \frac{x}{8}(2x^2-a^2)\sqrt{x^2-a^2} - \frac{a4}{8}\cosh^{-1}\frac{x}{a} + c$

44) $\int \frac{x^2}{\sqrt{x^2-a^2}}dx = \frac{a^2}{2}\cosh^{-1}\frac{x}{a} + \frac{x}{2}\sqrt{x^2-a^2} + c$

45) $\int \frac{dx}{x\sqrt{x^2-a^2}} = \frac{1}{a}\sec^{-1}|\frac{x}{a}| + c = \frac{1}{a}\cos^{-1}|\frac{a}{x}| + c$

46) $\int \frac{dx}{x^2\sqrt{x^2-a^2}} = \frac{\sqrt{x^2-a^2}}{a^2x} + c$

47) $\int \frac{dx}{x\sqrt{2ax-x^2}} = \sin^{-1}(\frac{x-a}{a}) + c$

48) $\int \sqrt{2ax-x^2}dx = \frac{x-a}{2}\sqrt{2ax-x^2} + \frac{a^2}{2}\sin^{-1}(\frac{x-a}{a}) + c$

49) $\int (\sqrt{2ax-x^2})^n dx = \frac{(x-a)(\sqrt{2ax-x^2})^n}{n+1} + \frac{na^2}{n+1}\int (\sqrt{2ax-x^2})^{n-2}dx$

50) $\int \frac{dx}{(\sqrt{2ax-x^2})^n} = \frac{(x-a)(\sqrt{2ax-x^2})^{2-n}}{(n-2)a^2} + \frac{(n-3)}{(n-2)a^2}\int \frac{dx}{(\sqrt{2ax-x^2})^{n-2}}$

51) $\int x\sqrt{2ax-x^2}dx = \frac{(x+a)(2x-3a)\sqrt{2ax-x^2}}{6} + \frac{a^3}{2}\sin^{-1}\frac{x-a}{a} + c$

52) $\int \frac{\sqrt{2ax-x^2}}{x}dx = \sqrt{2ax-x^2} + a\sin^{-1}\frac{x-a}{a} + c$

53) $\int \frac{\sqrt{2ax-x^2}}{x^2}dx = -2\sqrt{\frac{2a-x}{x}} - \sin^{-1}(\frac{x-a}{a}) + c$

54) $\int \frac{xdx}{\sqrt{2ax-x^2}} = a\sin^{-1}\frac{x-a}{a} - \sqrt{2ax-x^2} + c$

55) $\int \frac{dx}{x\sqrt{2ax-x^2}} = -\frac{1}{a}\sqrt{\frac{2a-x}{x}} + c$

56) $\int \sin axdx = -\frac{1}{a}\cos ax + c$

57) $\int \cos axdx = \frac{1}{a}\sin ax + c$

58) $\int \sin^2 axdx = \frac{x}{2} - \frac{\sin 2ax}{4a} + c$

59) $\int \cos^2 axdx = \frac{x}{2} + \frac{\sin 2ax}{4a} + c$

60) $\int \sin^n axdx = \frac{-\sin^{n-1}ax\cos ax}{na} + \frac{n-1}{n}\int \sin^{n-2}axdx$

61) $\int \cos^n axdx = \frac{\cos^{n-1}ax\sin ax}{na} + \frac{n-1}{n}\int \cos^{n-2}axdx$

62)(a) $\int \sin ax\cos bxdx = -\frac{\cos(a+b)x}{2(a-b)} - \frac{\sin(a+b)x}{2(a+b)}, a^2 \neq b^2$

(b) $\int \sin ax \sin bx dx = -\frac{\sin(a-b)x}{2(a-b)} - \frac{\sin(a+b)x}{2(a+b)}, a^2 \neq b^2$

(c) $\int \cos ax \cos bx dx = -\frac{\sin(a-b)x}{2(a-b)} + \frac{\sin(a+b)x}{2(a+b)}, a^2 \neq b^2$

63) $\int \sin ax \cos ax dx = -\frac{\cos 2ax}{4a} + c$

64) $\int \sin^n ax \cos ax dx = \frac{\sin^{n+1} ax}{(n+1)a} + c, n \neq -1$

65. $\int \frac{\cos ax}{\sin ax} dx = \frac{1}{a} \ln |\sin ax| + C$

66. $\int \cos^n ax \sin ax \, dx = -\frac{\cos^{n+1} ax}{(n+1)\, a} + C, \; n \neq -1$

67. $\int \frac{\sin^n ax}{\cos^m ax} dx = -\frac{1}{a} \ln |\cos ax| + C$

68. $\int \sin^n ax \cos^m ax \, dx = \frac{\sin^{n-1} ax \cos^{m+1} ax}{a(m+n)} + \frac{n-1}{m+n} \int \sin^{n-2} ax \cos^m ax \, dx,$
$n \neq -m$ (si n=-m ,utiliser N°.86)

69. $\int \sin^n ax \cos^m ax \; dx \; = \; -\frac{\sin^{n+1} ax \cos^{m-1} ax}{a(m+n)} + \frac{m-1}{m+n} \int \sin^n ax \cos^{m-2} ax \, dx,$
$n \neq -m$ (si m=-n ,utiliser N°.87)

70. $\int \frac{dx}{b + c \sin ax} = \frac{-2}{a\sqrt{b^2 - c^2}} \tan[\sqrt{\frac{b-c}{b+c}} \tan(\frac{\pi}{4} - \frac{ax}{2})] + C \quad b^2 > c^2$

71. $\int \frac{dx}{b + c \sin ax} = \frac{-1}{a\sqrt{c^2 - b^2}} \ln |\frac{c + b \sin ax + \sqrt{c^2 - b^2} \cos ax}{b + c \sin ax}| + C, \quad b^2 < c^2$

72. $\int \frac{dx}{1 + \sin ax} = -\frac{1}{a} \tan(\frac{\pi}{4} - \frac{ax}{2}) + C$

73. $\int \frac{dx}{1 - \sin ax} = \frac{1}{a} \tan(\frac{\pi}{4} + \frac{ax}{2}) + C$

74. $\int \frac{dx}{b + c \cos ax} = \frac{-2}{a\sqrt{b^2 - c^2}} \tan^{-1}[\sqrt{\frac{b-c}{b+c}} \tan \frac{ax}{2}] + C, \qquad b^2 > c^2$

75. $\int \frac{dx}{b + c \cos ax} = \frac{-1}{a\sqrt{c^2 - b^2}} \ln |\frac{c + b \cos ax + \sqrt{c^2 - b^2} \sin ax}{b + c \cos ax}| + C, \quad b^2 < c^2$

76. $\int \frac{dx}{1 + \cos ax} = \frac{1}{a} \tan \frac{ax}{2} + C$

77. $\int \frac{dx}{1 - \cos ax} = -\frac{1}{a} \cot \frac{ax}{2} + C$

78. $\int x \sin ax \, dx = \frac{1}{a^2} \sin ax - \frac{x}{a} \cos ax + C$

79. $\int x \cos ax \, dx = \frac{1}{a^2} \cos ax + \frac{x}{a} \sin ax + C$

80. $\int x^n \sin ax \, dx = -\frac{x^n}{a} \cos ax + \frac{n}{a} \int x^{n-1} \cos ax \, dx$

81. $\int x^n \cos ax \, dx = \frac{x^n}{a} \sin ax - \frac{n}{a} \int x^{n-1} \sin ax \, dx$

82. $\int \tan ax\, dx = \frac{1}{a} \ln |\sec ax| + C$

83. $\int \cot ax\, dx = \frac{1}{a} \ln |\sin ax| + C$

84. $\int \tan^2 ax\, dx = \frac{1}{a} \tan ax - x + C$

85. $\int \cot^2 ax\, dx = -\frac{1}{a} \cot ax\, dx - x - C$

86. $\int \tan^n ax\, dx = \frac{\tan^{n-1} ax}{a(n-1)} - \int \tan^{n-2} ax\, dx, \qquad n \neq 1$

87. $\int \cot^n ax\, dx = -\frac{\cot^{n-1} ax}{a(n-1)} - \int \cot^{n-2} ax\, dx, \qquad n \neq 1$

88. $\int \arcsin ax\, dx = x \arcsin ax + \frac{1}{a}\sqrt{1-a^2x^2} + C$

89. $\int \arccos ax\, dx = x \arccos ax - \frac{1}{a}\sqrt{1-a^2x^2} + C$

90. $\int \arctan ax\, dx = x \arctan ax - \frac{1}{2a} \ln(1+a^2x^2) + C$

91. $\int x^n \arcsin ax\, dx = \frac{x^{n+1}}{n+1} \arcsin ax - \frac{a}{n+1} \int \frac{x^{n+1}}{\sqrt{1-a^2x^2}} dx, \quad n \neq -1$

92. $\int x^n \arccos ax\, dx = \frac{x^{n+1}}{n+1} \arccos ax + \frac{a}{n+1} \int \frac{x^{n+1}}{\sqrt{1-a^2x^2}} dx, \quad n \neq -1$

93. $\int x^n \arctan ax\, dx = \frac{x^{n+1}}{n+1} \arctan ax - \frac{a}{n+1} \int \frac{x^{n+1}}{1+a^2x^2} dx, \quad n \neq -1$

94. $\int e^{ax} dx = \frac{1}{a} e^{ax} + C$

95. $\int b^{ax} dx = \frac{1}{a} \frac{b^{ax}}{\ln b} + C, \quad b > 0\,, b \neq 1$

96. $\int x e^{ax} dx = \frac{e^{ax}}{a^2}(ax-1) + C$

97. $\int x^n e^{ax} dx = \frac{1}{a} x^n e^{ax} - \frac{n}{a} \int x^{n-1} e^{ax} dx$

98. $\int x^n b^{ax} dx = \frac{x^n b^{ax}}{a \ln b} - \frac{n}{a \ln b} \int x^{n-1} b^{ax} dx, \qquad b > 0\,, b \neq 1$

99. $\int e^{ax} \sin bx\, dx = \frac{e^{ax}}{a^2+b^2}(a \sin bx - b \cos bx) + c$

100. $\int e^{ax} \cos bx\, dx = \frac{e^{ax}}{a^2+b^2}(a \cos bx + b \sin bx) + c$

101. $\int \ln ax\, dx = x \ln ax - x + C$

102. $\int x^n \ln ax\, dx = \frac{x^{n+1}}{n+1} \ln ax - \frac{x^{n+1}}{(n+1)^2} + C, \qquad n \neq -1$

103. $\int x^{-1} \ln ax\, dx = \frac{1}{2} (\ln ax)^2 + C$

104. $\int \frac{dx}{x \ln ax} = \ln |\ln ax| + C$

105. $\int \sinh ax\, dx = \frac{1}{a} \cosh ax + C$

106. $\int \cosh ax\, dx = \frac{1}{a}\sinh ax + C$

107. $\int \sinh^2 ax\, dx = \frac{\sinh 2ax}{4a} - \frac{x}{2} + c$

108. $\int \cos^2 ax\, dx = \frac{\sinh 2ax}{4a} + \frac{x}{2} + c$

109. $\int \sinh^n ax\, dx = \frac{\sinh^{n-1} ax \cosh ax}{na} - \frac{n-1}{n}\int \sinh^{n-2} ax\, dx\,, \qquad n \neq 0$

110. $\int \cosh^n ax\, dx = \frac{\cosh^{n-1} ax \sinh ax}{na} + \frac{n-1}{n}\int \cos^{n-2} ax\, dx\,, \qquad n \neq 0$

111. $\int x \sinh ax\, dx = \frac{x}{a}\cosh ax - \frac{1}{a^2}\sinh ax + c$

112. $\int x \cosh ax\, dx = \frac{x}{a}\sinh ax - \frac{1}{a^2}\cosh ax + c$

113. $\int x^n \sin ax\, dx = \frac{x^n}{a}\cosh ax - \frac{n}{a}\int x^{n-1}\cos ax\, dx$

114. $\int x^n \cosh ax\, dx = \frac{x^n}{a}\sinh ax - \frac{n}{a}\int x^{n-1}\sinh ax\, dx$

115. $\int \tanh ax\, dx = \frac{1}{a}\ln(\cosh ax) + C$

116. $\int \coth ax\, dx = \frac{1}{a}\ln|\sinh ax| + C$

117. $\int \tanh^2 ax\, dx = x - \frac{1}{a}\tanh ax + C$

118. $\int \coth^2 ax\, dx = x - \frac{1}{a}\coth ax + C$

119. $\int \tanh^n ax\, dx = -\frac{\tanh^{n-1} ax}{(n-1)a} + \int \tanh^{n-2} ax\, dx \qquad n \neq 1$

120. $\int \coth^n ax\, dx = -\frac{\coth^{n-1} ax}{(n-1)a} + \int \coth^{n-2} ax\, dx \qquad n \neq 1$

121. $\int e^{ax}\sinh bx\, dx = \frac{e^{ax}}{2}[\frac{e^{bx}}{a+b} - \frac{e^{-bx}}{a-b}] + c, \quad a^2 \neq b^2$

122. $\int e^{ax}\cosh bx\, dx = \frac{e^{ax}}{2}[\frac{e^{bx}}{a+b} + \frac{e^{-bx}}{a-b}] + c, \quad a^2 \neq b^2$

123. $\int_0^\infty x^{n-1}e^{-x}dx = R(n) = (n-1)!, \quad$ n> 0.

124. $\int_0^\infty e^{-ax^2}dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}, \qquad$ a>0

125. $\int_0^{\frac{\pi}{2}} \sin^n x\, dx = \int_0^{\frac{\pi}{2}} \cos^n x\, dx = \begin{cases} \frac{2\cdot 4\cdot 6\cdots(n-1)}{3\cdot 5\cdot 7\cdots n} & \text{si n enier impaire} \geq 3 \\ \frac{1\cdot 3\cdot 4\cdots(n-1)}{2\cdot 4\cdot 6\cdots n}\cdot\frac{\pi}{2} & \text{si n entier paire } \geq 2 \end{cases}$

Chapitre 2

Approximations des solutions de l'équation $f(x) = 0$

2.1 Rappels et notations

Définition 2.1.1. :
Soit k un réel strictement positif et g une fonction définie sur un intervalle $[a, b]$ de $\mathbb{R}$ à valeurs dans $\mathbb{R}$. La fonction g est dite Lipschitzienne de rapport de k (encore dite $k-$ Lipschitzienne) si pour tous x et y de $[a, b]$ on a : $|g(x) - g(y)| \leq k|x - y|$.

Définition 2.1.2. :
Soit g une fonction $k-$Lipschitzienne sur $[a, b]$. La fonction g est dite contractante de rapport de contraction k si $k \in]0, 1[$.

Exemple 2.1.1. :
La fonction $g(x) = \sin(x)$ est Lipschitzienne de rappport $k = 1$

Exercice 2.1.1. :
Montrer que la La fonction $g(x) = \frac{1}{3}\cos(x)$ est Lipschitzienne et déterminer le rappport k

Définition 2.1.3. :
Soit g une fonction définie sur un intervalle $[a, b]$ de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ la fonction g est dite uniformément continue sur $[a, b]$ si :

$\forall\, \varepsilon \geq 0, \exists\, \eta$ tel que $\forall\, x$ et y de $[a, b]$ vérifiant $|\, y - x| \leq \eta$, on ait $|g(y) - g(x)| \leq \varepsilon$

Remarque 2.1.1. :
Toute fonction Lipschitzienne sur $[a, b]$ est unfiormément continue sur $[a, b]$.

Théorème 2.1.1 (des Valeurs Intermédiaires)**.**
Soit f une fonction définie et continue sur un intervalle fermé borné $[a, b]$ de $\mathbb{R}$.
Alors pour tout réel θ appartenant à $f([a\ ,\ b])$, il existe un réel $c \in [a, b]$ tel que $\theta = f(c)$.
Si de plus f est strictement monotone alors le point c est unique.

Théorème 2.1.2 (des Valeurs Intermédiaires cas particulier $\theta = 0$)**.**
Soit f une fonction définie et continue sur un intervalle $[a, b]$ et vérifiant $f(a) \times f(b) \leq 0$, alors il existe un réel $c \in [a, b]$ tel que $f(c) = 0$.
Si de plus f est strictement monotone alors le point c est unique.

Théorème 2.1.3 (de Rolle)**.**
Soit f une fonction définie sur $[a, b]$ et à valeurs dans $\mathbb{R}$. Si f est continue sur $[a, b]$ et dérivable sur $]a, b[$, alors il existe un réel $c \in]a, b[$ tel que : $f'(c) = 0$.

Théorème 2.1.4 (des Accroissements Finis)**.**
Soit f une fonction définie sur $[a, b]$ et à valeurs dans $\mathbb{R}$ Si f est continue sur $[a, b]$ et dérivable sur $]a, b[$, alors il existe un réel $c \in]a, b[$ tel que :

$$f(b) - f(a) = (b - a) \times f'(c).$$

Théorème 2.1.5 (Formule de Taylor)**.**
Soit f une fonction de classe C^n sur $[a, b]$ dont la dérivée $f^{(n+1)}$ est définie sur $]a, b[$, alors il existe un réel $c \in]a, b[$ tel que :

$$f(b) = f(a) + (b - a)f'(a) + ...\frac{1}{n!}(b - a)^n f^{(n)}(a) + \frac{1}{(n + 1)!}(b - a)^{n+1} f^{(n+1)}(c).$$

Théorème 2.1.6 (Formule de MacLaurin)**.**
Soit f une fonction de classe C^n sur un intervalle I contenant 0 et telle que $f^{(n)}$ soit dérivable à l'intrérieur de I . Alors $\forall x \in I$, il existe un réel c strictement compris entre 0 et x tel que :

$$f(x) = f(0) + xf^{(1)}(0) + \frac{1}{2!}x^2 f''(0) + ...\frac{1}{n!}x^n f^{(n)}(0) + \frac{1}{(n + 1)!}x^{n+1} f^{(n+1)}(c).$$

Définition 2.1.4. :
Soit θ un réel et f une fonction définie sur un intervalle I de $\mathbb{R}$ et à valeurs dans $\mathbb{R}$.
θ est dit zéro de f si $f(\theta\ ) = 0$

Définition 2.1.5. :
Soit θ un réel et g une fonction définie sur un intervalle I de $\mathbb{R}$ et à valeurs dans $\mathbb{R}$.
θ est dit point fixe de g si $g(\theta\ ) = \theta$.

Lemme 2.1.1. *:*
Soit I un intervalle de $\mathbb{R}$ et f une fonction définie sur I et à valeurs dans $\mathbb{R}$.
Alors la recherche des zéros de f est équivalente à la recherche des points fixes de la fonction g définie sur I par : $g(x) = x - f(x)$

Preuve :
En effet, si $f(\theta) = 0$ alors $g(\theta) = \theta - f(\theta) = \theta$ et inversement, si $g(\theta) = \theta$ alors $f(\theta) = \theta - g(\theta) = \theta - \theta = 0$.

Lemme 2.1.2. *:*
Soit g une fonction de classe C^1 sur $[a, b]$. S'il existe un réel $k \geq 0$ tel que : $|g'(x)| \leq k$ $\forall x \in [a, b]$ alors la fonction g est $k-$Lipschitzienne sur $[a, b]$.

Preuve :
Il suffit d'appliquer le théorème des accroissements finis à la fonction g sur $[x, y]$ avec $x \leq y$.
Donc il existe $c \in]x, y[$ tel que : $g(y) - g(x) = (y - x)g'(c)$ et comme on a : $|g'(c)| \leq k$, il s'ensuit que : $|g(y) - g(x)| \leq k|x - y|$

Définition 2.1.6. :
Soit (u_n) une suite admettant pour limite θ.
On appelle erreur de la n^{eme} étape le réel défini par $e_n = u_n - \theta$

Définition 2.1.7. :
On dit que la convergence de (u_n) vers θ est d'order p si :
$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C$ où p et C sont des réels > 0
Si $p = 1$ (avec $C < 1$) la convergence est dite linéaire
Si $p = 2$ on dit que la convergence est quadratique.

Remarque 2.1.2. :
L'ordre de convergence p n'est pas nécessairement un entier.

Définition 2.1.8. :
On dira que le réel δ est une approximation du réel α avec la precision ε si :
$|\alpha - \delta| \leq \varepsilon$.
En particulier, on dira que le terme u_{n_0} d'une suite (u_n) approche la limite θ avec précision ε si $|u_{n_0} - \theta| \leq \varepsilon$.

Exemple 2.1.2. :
la suite $(u_n) = (\frac{1}{n})$ tend vers zéro quand n tend vers l'infini.
Si on veut une précision $\varepsilon = 10^{-1}$, il suffit de prendre n_0 tel que $\frac{1}{n_0} \leq 10^{-1}$ou

encore $n_0 \geq 10$
mais si on exige une precision de 10^{-5} alors on doit prendre n_0 tel que $\frac{1}{n_0} \leq 10^{-5}$
c.a.d $n_0 \geq 10^5$

Remarque 2.1.3. :
Il est important de saisir la notion de vitesse de convergence. Par exemple, les suites $(\frac{1}{n}), (\frac{1}{n^2}), (\frac{1}{n^4})$ convergent vers zéro quand n tend vers l'infini mais la vitesse de convergence diffère d'une suite à l'autre.

Théorème 2.1.7. *:*
Soit g une fonction $k-$contractante sur $[a,b]$ et à valeurs dans $[a,b]$, et (u_n) la suite récurrente définie par :
$u_0 \in [a,b]$, u_0 donné et $u_{n+1} = g(u_n)$ pour tout $n \geq 0$
Alors :
1- la suite (u_n) converge vers un réel θ
2- la fonction g admet un point fixe unique
3- Pour tout $n \in \mathbb{N}^*$ on a : $|u_n - \theta| \leq \frac{k^n}{1-k}|u_1 - u_0|$

Preuve :
Tout d'abord, comme $u_0 \in [a,b]$ et que $g : [a,b] \to [a,b]$, on a $u_n \in [a,b]$ pour tout $n \in \mathbb{N}$.
Ensuite, le fait que g soit une fonction $k-$contractante implique que :

$$|u_{n+1} - u_n| = |g(u_n) - g(u_{n-1})| \leq k|u_n - u_{n-1}| \text{ pour tout } n \geq 1.$$

Par conséquent on obtient :

$$|u_{n+1} - u_n| \leq k^n|u_1 - u_0| \text{ pour tout } n \geq 0 \tag{2.1.1}$$

A l'aide de l'inégalité 2.1.1 on montre que la suite (u_n) vérifie :

$$|u_{n+p} - u_n| \leq \frac{1}{1-k}k^n|u_1 - u_0|$$

En effet : Pour tous $p \in \mathbb{N}^*$ et $n \in \mathbb{N}$ on a :

$$\begin{aligned}
|u_{n+p} - u_n| &\leq |u_{n+p} - u_{n+p-1}| + |u_{n+p-1} - u_{n+p-2}| + ...|u_{n+2} - u_{n+1}| + |u_{n+1} - u_n| \\
&\leq k^{n+p-1}|u_1 - u_0| + k^{n+p-2}|u_1 - u_0|.. + k^{n+1}|u_1 - u_0| + k^n|u_1 - u_0| \\
&\leq \frac{1-k^p}{1-k}k^n|u_1 - u_0| \\
&\leq \frac{1}{1-k}k^n|u_1 - u_0| \quad (\mathbf{2})
\end{aligned}$$

L'inégalité (2) nous permet de prouver que la suite (u_n) est de Cauchy. En effet :
Comme $k^n \xrightarrow[n\longrightarrow+\infty]{} 0$ alors pour tout $\varepsilon > 0$, il existe n_0 tel que pour tout $n \geq n_0$ on ait : $k^n \leq \dfrac{1-k}{|u_1-u_0|}\varepsilon$ et par suite : $\dfrac{1}{1-k}k^n|u_1-u_0| \leq \varepsilon$
Donc pour tout $\varepsilon > 0$, il existe n_0 tel que pour tout $n \geq n_0$ on ait :
$|u_{n+p}-u_n| \leq k^n \dfrac{1-k}{|u_1-u_0|} \leq \varepsilon$
La suite (u_n) est donc de Cauchy et par conséquent elle converge vers une limite θ.
Comme la fonction g est continue sur $[a,b]$, que $u_{n+1} = g(u_n)$ et que
$u_n \in [a,b] \ \ \forall\, n \in \mathbb{N}$
alors on a : $\lim\limits_{n\to\infty} u_n = \theta = g(\theta)$ c-a-d : θ est un point fixe de g
Unicité du point fixe :
Supposons que g admet un autre point fixe α different de θ alors on a :
$|g(\alpha)-g(\theta)| = |\alpha-\theta| \leq k|\alpha-\theta|$ ou encore $(1-k)|\alpha-\theta| \leq 0$ mais comme $k < 1$, alors $\alpha = \theta$
Enfin, en faisant tendre p vers l'infini dans l'inégalité $|u_{n+p}-u_n| \leq \dfrac{k^n}{1-k}|u_1-u_0|$, on obtient :

$$|\theta-u_n| \leq \frac{k^n}{1-k}|u_1-u_0| \quad \forall n \in \mathbb{N}^*$$

Théorème 2.1.8 (condition de convergence locale)**.**
Soit g une fonction de classe C^1 au voisinage θ . Si $g(\theta) = \theta$ et $|g'(\theta)| \prec 1$,
alors il existe ε strictement positif tel que :
$\forall u_0 \in I_\varepsilon = [\theta-\varepsilon, \theta+\varepsilon]$, la suite $(u_n) = (g(u_{n-1}))$ est définie et converge vers θ , l'unique solution de $g(x) = x$ dans I_ε

Preuve :
Puisque g est de classe C^1au voisinage de θ et que $|g'(\theta)| < 1$
on a :$|g'(x)| < 1 \ \ \forall x$ au voisinage de θ.
Par consequent, il existe ε strictement positif tel que :

$$\forall x \in I_\varepsilon\ ,\ \ |g'(x)| < 1$$

et puisque g' est continue sur le fermé borné I_ε ,
on déduit qu'il existe $k \in]0,1[$ tel que :

$$\forall x \in I_\varepsilon\ ,\ \ |g'(x)| \leq k < 1$$

Pour appliquer le théorème , il suffit de vérifier que : $g(I_\varepsilon) \subset I_\varepsilon$.
Or, par application du théorème des accroissements finis on a :
$\forall x \in I_\varepsilon, |g(x)-\theta| \leq |x-\theta|$

Remarque 2.1.4. :

- Si $|g'(\theta)| = 1$, la suite peut converger ou diverger
- Si $|g'(\theta)| \geq 1$ et si la suite possède une infinité de termes différents de θ , alors la suite ne peut converger.

En effet, si on suppose que la suite converge vers θ on obtient :
$u_{n+1} - \theta = (u_n - \theta)g'(c_n)$ avec c_n compris entre u_n et θ et de là on aboutit à une contradiction en supposant que u_n est assez proche de θ de telle sorte que :
$|g'(c_n)| \geq 1 \Longrightarrow |u_{n+1} - \theta| \geq |u_n - \theta|$

Théorème 2.1.9. *:*
Si la suite récurrente définie par : $u_0 \in [a, b]$, u_0 donné et $u_{n+1} = g(u_n)$, $\forall n \geq 0$, converge linéairement vers θ et si g est de classe C^1 sur $[a, b]$, alors $C = \lim\limits_{n\to\infty} \dfrac{|e_{n+1}|}{|e_n|} = |g'(\theta)|$.

Preuve :
Il suffit d'appliquer le théorème des accroissements finis :
$|e_{n+1}| = |u_{n+1} - \theta| = |g(u_n) - g(\theta)| = |(u_n - \theta)g'(c_n)|$ et de là on obtient

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = \lim_{n\to\infty} |g'(c_n)| = |g'(\theta)|$$

Remarque :
On veut résoudre numériquement l'équation $f(x) = 0$. On constate qu'il existe plusieurs façon d'écrire cette équation sous la forme d'un problème de point fixe c'est-à-dire sous la forme $g(x) = x$. Par exemple on à les trois écritures suivantes :

$$\begin{aligned} x^2 - 2x - 3 = 0 &\Longrightarrow x^2 = 2x + 3 \\ &\Longrightarrow x = g_1(x) = \sqrt{2x+3} \end{aligned} \tag{2.1.2}$$

$$\begin{aligned} x^2 - 2x - 3 = 0 &\Longrightarrow 2x = x^2 - 3 \\ &\Longrightarrow x = g_2(x) = \frac{x^2-3}{2} \end{aligned} \tag{2.1.3}$$

$$\begin{aligned} x^2 - 2x - 3 = 0 &\Longrightarrow x^2 = 2x + 3 \\ &\Longrightarrow x = g_3(x) = \frac{2x+3}{x} \end{aligned} \tag{2.1.4}$$

Les trois équations 2.1.2, 2.1.3 et 2.1.4 admettent pour points fixes $-$ et 3. Pour la convergence locale ou globale il faut étudier $|g_i'(x)|$, $|g_i'(-1)|$ et $|g_i'(3)|$ $i = 1, 2, 3$

2.2 Méthode de Newton :

En prenant la fonction g définie par : $g(x) = x - \dfrac{f(x)}{f'(x)}$,
on obtient le procédé de Newton donné par :
x_0 donné , $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$ pour $n \geq 0$ avec $f'(x_n) \neq 0$

Théorème 2.2.1. *:*
Soit une fonction de classe C^2 sur $[a,b]$ satisfaisant les conditions suivantes :

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

ii) $\forall x \in [a,b],\ |f'(x)| \neq 0$

iii) *f'' est de signe constant sur $[a,b]$ (convexité ou concavité)*

iv) $\frac{|f(a)|}{|f'(a)|} < b-a,\ \frac{|f(b)|}{|f'(b)|} < b-a$

Alors la méthode de Newton converge vers l'unique solution θ de $f(x) = 0$ dans $[a,b]$ et ceci pour n'importe quel choix de $x_0 \in [a,b]$.

Preuve :
Considérons le cas $f(a).f(b) < 0$, $f'(x) < 0$ et $f''(x) < 0$
D'après i) et ii) , il existe une solution $\theta \in [a,b]$ qui verifie :
$x_{n+1} - \theta = x_n - \theta + \frac{f(\theta) - f(x_n)}{f'(x_n)} = \frac{1}{2}(x_n - \theta)^2 \frac{f''(c)}{f'(x_n)}$ avec $x_n \leq c \leq \theta$
Par conséquent, le signe de $x_{n+1} - \theta$ est celui de $f''(c)f'(x_n)$.
Si $f'(x) < 0$ et $f''(x) < 0$ alors $x_{n+1} \geq \theta$ pour tout $n \geq 0$, (x_n) est donc minorée par θ à partir de x_1
Pour montrer que la suite (x_n) est décroissante, on distingue deux cas :

1. Si $x_0 > \theta$, $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \leq x_0$ et on montre que $x_{n+1} \leq x_n$ pour tout $n \geq 0$
2. Si $x_0 < \theta$, $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \geq x_0$ et on montre que $x_{n+1} \leq x_n$ pour tout $n \geq 1$

Exemple 2.2.1. Considérons l'équation $f(x) = x^3 + 4x^2 - 10 = 0, \quad x > 0$ On a f une fonction strictement croissante (car $f' = 3x^2 + 8x > 0$ sur $]0, +\infty[$) et comme $f(1) = -5$ et $f(4) = 118$ donc d'après le théorème des valeurs intermédiaires f admet une seule racine dans l'intervalle $[0;4]$. Soit $x_0 = 3.9$. la solution est donnée par la figure (2.1) :

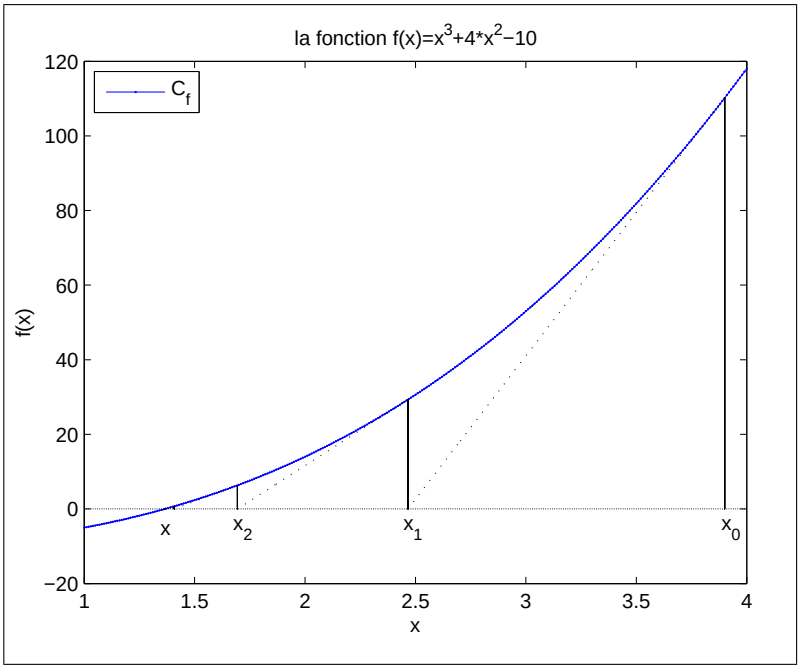


FIG. 2.1 – **la solution est** $x = 1.3652$

2.3 Méthode de Newton modifiée :

En prenant la fonction g définie par : $g(x) = x - \dfrac{f(x)}{f'(x_0)}$, $f'(x_0) \neq 0$
on obtient la méthode de Newton modifiée comme suit :
x_0 donné , $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_0)}$ pour $n \geq 0$
on montre alors que cette méthode est d'ordre 1.
Autres modifications de la méthode de Newton peuvent etre obtenues en prenant $f'(x_{\sigma(n)})$ pour des valeurs intérmédaires.

2.4 Méthode de dichotomie :

Soit une f fonction définie continue sur $[a, b]$ vérifiant $f(a)f(b) \leq 0$.
La fonction f admet donc au moins un zéro $\theta \in [a, b]$.
La méthode de dichotomie consiste à approcher θ par encadrement, en réduisant à chaque étape la longueur de l'intervalle de moitié selon l'algorithme suivant :
Etape 1
On pose $a_0 = a$ et $b_0 = b$ on pose $c_0 = \dfrac{a_0 + b_0}{2}$ puis on teste si $c_0 = \theta$ c'est terminé, sinon :
Si $f(a_0)f(c_0) < 0$ alors $\theta \in [a_0, c_0]$ on pose alors $a_1 = a_0$ et $b_1 = c_0$

puis $c_1 = \dfrac{a_1 + b_1}{2}$
Si $f(b_0)f(c_0) < 0$ alors $\theta \in [c_0, b_0]$ on pose alors $a_1 = c_0$ et $b_1 = b_0$
puis $c_1 = \dfrac{a_1 + b_1}{2}$
Aprés cette étape la longueur de $[a_1, b_1]$ est égale à $\dfrac{b_0 - a_0}{2} = \dfrac{b - a}{2}$
Etape 2

on recommence le procédé de l'étape 1.

Etape k

A chaque étape k du procédé, soit on tombe sur un $c_k = \theta$ soit on diminue la longueur de l'intervalle de moitié.

Théorème 2.4.1. *:*
Les a_k , b_k et c_k satisfont les propriétés suivantes :

1- $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$

2-$b_{k+1} - a_{k+1} = \dfrac{b_k - a_k}{2} = \dfrac{b_0 - a_0}{2^{k+1}}$

3- La suite (c_k) converge vers θ

4- $|c_k - \theta| \leq \dfrac{b - a}{2^{k+1}}$

Preuve :

1- Pour $k \geq 0$ on a $c_k = \dfrac{a_k + b_k}{2}$ et $[a_{k+1}, b_{k+1}] = [a_k, c_k]$ ou $[a_{k+1}, b_{k+1}] = [c_k, b_k]$
Donc $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$

2- On a par construction $b_{k+1} - a_{k+1} = \dfrac{b_k - a_k}{2}$, montrons par récurrence que : $b_k - a_k = \dfrac{b - a}{2^k}$
Pour $k = 0$ la relation est vérifiée
Si on suppose que la relation est vraie à l'ordre k ($b_k - a_k = \dfrac{b - a}{2^k}$) alors on a :
$b_{k+1} - a_{k+1} = \dfrac{1}{2}(b_k - a_k) = \dfrac{1}{2}\dfrac{b - a}{2^k} = \dfrac{b - a}{2^{k+1}}$

3- Par construction $\theta \in [a_k, b_k]$ et $c_k = \dfrac{a_k + b_k}{2}$ est le milieu de $[a_k, b_k]$
Donc $|c_k - \theta| \leq \dfrac{b_k - a_k}{2} \leq \dfrac{b - a}{2^{k+1}}$
En d'autres termes, on a : $\theta - \dfrac{b - a}{2^{k+1}} \leq c_k \leq \theta + \dfrac{b - a}{2^{k+1}}$
et comme $\dfrac{b - a}{2^{k+1}} \xrightarrow[n \longrightarrow +\infty]{} 0$ on déduit que $\lim\limits_{k \to \infty} c_k = \theta$

Remarque 2.4.1. *Le théorème précédent permet de calculer à l'avance le nombre maximal $n \in \mathbb{N}$ d'itérations assurant la précision ε, en effet*
Pour que c_n vérifie $|c_n - \theta| \leq \dfrac{b - a}{2^{n+1}}$ à la n^{ime} itération, il suffit que n vérifie : $\dfrac{b - a}{2^{n+1}} \leq \varepsilon$

On a alors :

$$|c_n - \theta| \leq \frac{b-a}{2^{n+1}} \leq \varepsilon$$

$$\textit{donc } \frac{b-a}{2^{n+1}} \leq \varepsilon \Longleftrightarrow \frac{b-a}{\varepsilon} \leq 2^{n+1} \Longleftrightarrow \log(\frac{b-a}{\varepsilon}) \leq (n+1)\log(2)$$
$$\Longleftrightarrow \frac{\log(b-a) - \log(\varepsilon)}{\log(2)} - 1 \leq n$$

Exemple 2.4.1. Soit $f(x) = x^3 + 4x^2 - 10 = 0$. On vérifie graphiquement que f admet une racine réelle dans l'intervalle $[1;2]$ et que la méthode de dichotomie est applicable (voir figure (2.2)).

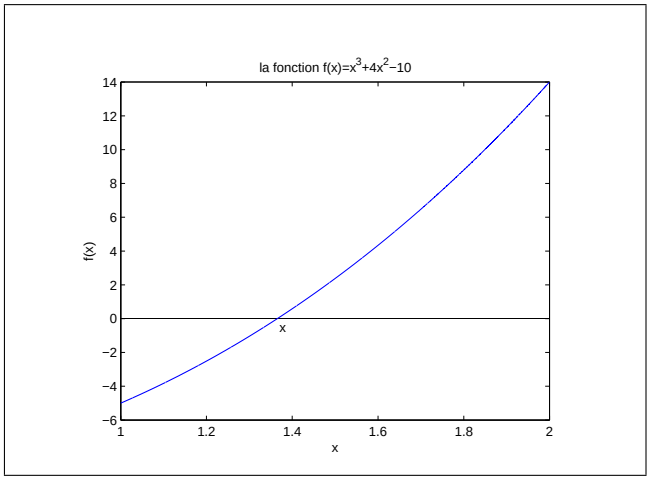


FIG. 2.2 – $f(1).f(2) < 0$

pour trouver une approximation de cette racine on peut utiliser la méthode de dichotomie avec une précision $= 10^{-10}$
on a les résultats suivants : n (numérique) $= 33$ n (théorique) $= 32.219281$
$x = 1.3652300134$ $f(x) = -2.378897e-011$

2.5 Méthode de fausse position (Regula Falsi) :

Au lieu de prendre à chaque étape c_k qui est le milieu de l'intervalle $[a_k, b_k]$, la méthode de fausse position prend le point d'intersection de l'axe des abscisses avec la droite passant par $(a_k, f(a_k))$ et $(b_k, f(b_k))$.
L'équation de cette droite est donnée par :

$$\frac{x-a}{b-a} = \frac{y-fa)}{f(b)-f(a)}.$$

Elle coupe l'axe des abscisses au point : $M_k(c_k, 0)$ où : $c_k = a_k + f(a_k)\dfrac{a_k - b_k}{f(a_k) - f(b_k)}$
En suite on procède comme dans le cas de dichotomie en testant :
Si $f(a_k)f(c_k) < 0$ alors $\theta \in [a_k, c_k]$ on pose alors $a_{k+1} = a_k$ et $b_{k+1} = c_k$
Si $f(b_k)f(c_k) < 0$ alors $\theta \in [c_k, b_k]$ on pose alors $a_{k+1} = c_k$ et $b_{k+1} = b_k$
puis on cherche à nouveau la droite passanrt par $(a_{k+1}, f(a_{k+1}))$ et $(b_{k+1}, f(b_{k+1}))$

Exemple 2.5.1. :
Considérons l'équation $f(x) = x^3 - 20 = 0$ comme $f(0.75) = -19,578125$ et $f(4.5) = 71,125$ alors $f(0.75).f(4.5) < 0$ donc on peut appliquer le méthode de fausse position (Regula Falsi) dans l'intervalle $[0.75; 4.5]$. La solution est donnée par la figure (**??**) :

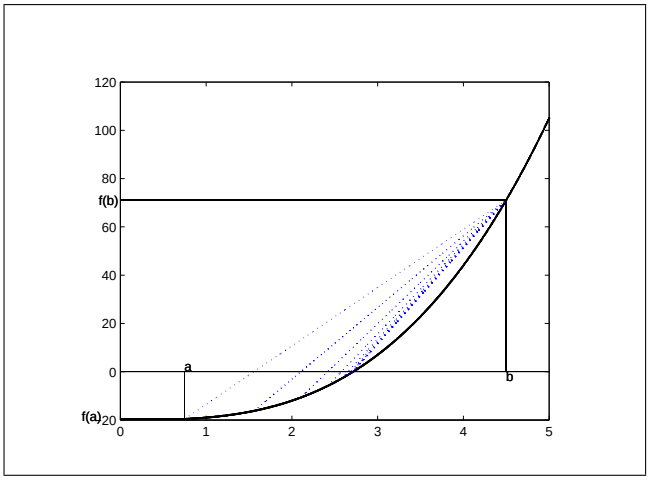


FIG. 2.3 – $x = 2.7133$

2.6 Exercices

Exercice 2.6.1. Considérons la suite récurrente (u_n) définie par : $0 < u_0 < 1$ et $u_{n+1} = h(u_n)\ \forall n \geq 0$
où h est donnée par l'équation logistique $h(x) = rx(1 - x)$

1. Montrer que si $r \in]0, 4]$ alors $0 < u_n \leq 1\ \forall n \geq 1$
2. vérifier que 0 est un point fixe trivial de h et trouver l'autre point fixe θ
3. A quelles conditions sur r la suite (u_n) converge-t-elle vers les points fixes 0 et θ?

Exercice 2.6.2. Soient $g_1,\ g_2, g_3$ et g_4 les fonctions définies par :
$g_1(x) = 2x - \sqrt{A}$, $g_2(x) = \dfrac{x(3A - x^2)}{2A}$, $g_3(x) = \dfrac{1}{2}(x + \dfrac{A}{x})$ et $g_4(x) = \dfrac{3A}{8x} + \dfrac{3x}{4} - \dfrac{x^3}{8A}$

1. vérifier que les quatre fonctions admettent $\sqrt{A}$ comme point fixe
2. Ecrire les formules de Taylor à l'ordre 3 au point $\sqrt{A}$ pour les quatre fonctions
3. On considère les suites (x_n) , (y_n) , (u_n) et (v_n) définies par :
   $x_{n+1} = g_1(x_n)$, $y_{n+1} = g_2(y_n)$, $u_{n+1} = g_3(u_n)$ et $v_{n+1} = h(v_n)$ puis on pose :
   $e_n = \sqrt{A} - x_n$, $\varepsilon_n = \sqrt{A} - y_n$, $d_n = \sqrt{A} - u_n$ et $\delta_n = \sqrt{A} - v_n$
   Trouver l'ordre de convergence et la constante d'erreur asymptotique dans chaque cas, c.a.d :
   $\lim\limits_{n\to\infty} \dfrac{\mid e_{n+1} \mid}{\mid e_n \mid^{p_1}} = C_1$, $\lim\limits_{n\to\infty} \dfrac{\mid \varepsilon_{n+1} \mid}{\mid \varepsilon_n \mid^{p_2}} = C_2$, $\lim\limits_{n\to\infty} \dfrac{\mid \varepsilon_{n+1} \mid}{\mid \varepsilon_n \mid^{p_3}} = C_3$, $\lim\limits_{n\to\infty} \dfrac{\mid e_{n+1} \mid}{\mid e_n \mid^{p_4}} = C_4$,
   où p_i et C_i sont des constantes > 0
4. Conclure.

Exercice 2.6.3. Soit p un entier ≥ 2 et f et g les fonctions définies sur $\mathbb{R}_*^+$ par :
$g(x) = Ax^{1-p}$ et $f(x) = x + \lambda(g(x) - x)$

1. La suite $x_{n+1} = g(x_n)$ converge-t-elle vers $A^{1/p}$?
2. pour quelle valeurs de λ a-t-on convergence de l'itération $x_{n+1} = f(x_n)$ vers $A^{1/p}$
3. Donner la valeur optimale de λ (c.a.d celle qui donne la convergence la plus rapide)
4. Comparer avec la méthode de Newton appliquée à la fonction $h(x) = x^p - A$

Exercice 2.6.4. Soient f et g les fonctions définies sur $\mathbb{R}$ par : $g(x) = \dfrac{1}{4}\cos(x)$ et $f(x) = x - \dfrac{1}{4}\cos(x)$

1. Montrer que la recherche des solutions de $f(x) = 0$ est équivalente à la recherche des points fixes de g, vérifier que de telles solutions existent et étudier l'unicité
2. Montrer que la suite récurrente définie par : $u_0 \in \mathbb{R}$, $u_{n+1} = g(u_n)\ \forall n \geq 0$ est convergente
3. Cette convergence depend-elle du choix de u_0?

Exercice 2.6.5. Soit f la fonction définie sur $\mathbb{R}$ par : $f(x) = x^3 + x - 1$

1. Montrer que la fonction f admet un zéro θ dans $[0,1]$
2. Ecrire la suite (x_n) obtenue à partir de la méthode de Newton
3. Etudier sur $[0,+\infty[$ le signe de la fonction h définie par : $h(x) = 2x^3 - 3\theta x^2 - 1 - \theta$
4. On suppose que $x_0 \in [0,1]$

   i) Montrer que $x_n \in [\theta,1]$ pour tout $n \geq 1$

   ii) Montrer que la suite (x_n) est décroissante et conclure
5. Montrer que pour tout $n \geq 0$ on a : $0 < x_{n+1} - \theta < x_n - x_{n+1}$
6. En prenant $x_0 = 1$, donner une valeur approchée de θ avec une précision $\varepsilon = 10^{-2}$

Exercice 2.6.6. Soit θ un zéro d'ordre 1 de f , x_0 un réel différent de θ et a_n est une approximation de $f'(x_n)$

1. Considerons la méthode de Newton modifiée $(M1)$: $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_0)}$ $\forall n \geq 0$

   Quel est l'ordre de cette méthode
2. Soit la méthode $(M2)$: $x_{n+1} = x_n - \dfrac{f(x_n)}{a_n}$ pour $n \geq 0$

   Montrer que l'ordre de convergence est supérieur à 1 si $a_n \to f'(\theta)$ quand $n \to \infty$,
3. Si θ est un zéro d'ordre 2 de f et $f''(\theta) \neq 0$ quel est l'ordre de convergence de la méthode: $(M3)$ $x_{n+1} = x_n - 2\dfrac{f(x_n)}{f'(x_n)}$ pour $n \geq 0$
4. Quelle méthode $(M4)$ obtient-on en prenant dans $(M2)$ $a_n = \dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$
5. Facultatif : on démontre que l'erreur de la méthode $(M4)$ **vérifie :**

   $\lim\limits_{n\to\infty} \dfrac{e_{n+1}}{e_n e_{n-1}} = \dfrac{f''(\theta)}{2f'(x_n)}$

   et que l'odre de convergence est $p = \dfrac{1+\sqrt{5}}{2} \approx 1.62$

Exercice 2.6.7. Soit f une fonction de classe C^2 sur $\mathbb{R}$. On suppose que f admet une fonction réciproque g.
On cherche un zéro θ de f en passant par g. ($f(\theta) = 0 \Longleftrightarrow g(0) = \theta$)

1. Ecrire les dérivées premières et secondes de g
2. Soit $P(y)$ le polynôme de degré 1 vérifiant :
   $P(y_n) = g(y_n)$ et $P'(y_n) = g'(y_n)$

a) Exprimer $P(y)$ en fonction de y , y_n , $g(y_n)$ et $g'(y_n)$

b) Exprimer $P(y)$ en fonction de y , x_n , $f(x_n)$ et $f'(x_n)$)

c) Quel procédé obtient-on en prenant $x_{n+1} = P(0)$

3. Soit $P(y)$ le polynôme de degré 1 vérifiant :
   $P(y_n) = g(y_n)$ et $P(y_{n-1}) = g(y_{n-1})$

   a) Exprimer $P(y)$ en fonction de y , y_n , $g(y_n)$ et $g(y_{n-1})$

   b) Exprimer $P(y)$ en fonction de y , x_{n-1} , x_n , $f(x_{n-1})$ et $f(x_n)$

   c) Quel procédé obtient-on en prenant $x_{n+1} = P(0)$

4. Soit $P(y)$ le polynôme de degré 2 vérifiant :
   $P(y_n) = g(y_n)$, $P'(y_n) = g'(y_n)$ et $P''(y_n) = g''(y_n)$

   a) Exprimer $P(y)$ en fonction de y , y_n , $g(y_n)$, $g'(y_n)$ et $g''(y_n)$

   b) En exprimant les dérivées de g en fonction de celles de f et en prenant $x_{n+1} = P(0)$,
   Montrer qu'on obtient le procédé (de Tchebychev) suivant :
   $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)} - \dfrac{(f(x_n))^2 f''(x_n)}{2(f'(x_n))^3}$ avec $f'(x_n) \neq 0$

5. Facultatif : On montre que la méthode de Tchebychev est d'ordre 3

Chapitre 3

Inroduction à l'interpolation

3.1 Rappel et définitions

Soit $\mathbb{P}_n(x)$ l'espace des polynômes de degré inférieur ou égal à n .
On rappelle que $\{1, x, x^2, ..., x^n\}$ est une base de $\mathbb{P}_n(x)$ ($\dim \mathbb{P}_n(x) = n+1$)
On note δ_{ij} le symbole de Kronecker : $\delta_{ij} = 1$ si $i = j$; $\delta_{ij} = 0$ si $i \neq j$
Soient f une fonction continue sur $[a, b]$, $x_1, ..., x_n$ n points de $[a, b]$ et $g_1, ..., g_n$ des réels de même signe.
Alors on a : $\sum_{i=1}^{i=n} f(x_i)g_i = f(c) \sum_{i=1}^{i=n} g_i$ où $c \in [a, b]$

Définition 3.1.1. : **Interpolant**
Soit f une fonction réelle définie sur un intervalle $[a, b]$ contenant $n+1$ points distincts $x_0, x_1, ..., x_n$. Soit P_n un polynôme de degré inférieur ou égal à n .
On dit que P_n est un interpolant de f ou interpole f en $x_0, x_1, ..., x_n$ si :

$$P_n(x_i) = f(x_i) \quad \text{pour} \quad 0 \leq i \leq n$$

Définition 3.1.2. : **Polynômes de Lagrange**
Soient $x_0, x_1, ..., x_n$ $(n+1)$ points deux à deux distincts d'un intervalle $[a, b]$ de $\mathbb{R}$
On appelle interpolants de Lagrange les polynômes L_i définis pour $i = 0, ..., n$ par :

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

On a en particulier :

$$L_0(x) = \prod_{j=1}^{j=n} \frac{(x - x_j)}{(x_0 - x_j)} = \frac{(x - x_1)(x - x_2).....(x - x_i).....(x - x_n)}{(x_0 - x_1)(x_0 - x_2)....(x_0 - x_i)....(x_0 - x_n)}$$

$$L_n(x) = \prod_{j=0}^{j=n-1} \frac{(x - x_j)}{(x_n - x_j)} = \frac{(x - x_0)(x - x_1).....(x - x_i).....(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)....(x_n - x_i)....(x_n - x_{n-1})}$$

Si on prend $P_n(x) = L_0(x)f(x_0) + L_1(x)f(x_1) + ... + L_n(x)f(x_n)$
alors : $P_n(x_i) = f(x_i)$ pour $0 \leq i \leq n$

Exemple 3.1.1. :
Si $x_0 = -1$, $x_1 = 0$ et $x_2 = 1$,
$f(x_0) = 2$, $f(x_1) = 1$, $f(x_2) = -1$, on obtient

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{x(x-1)}{(-1)(-1-1)} = \frac{x(x-1)}{2}$$
$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-(-1))(x-1)}{-(-1)(-1)} = \frac{(x+1)(x-1)}{-1}$$
$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x+1)x}{(1-(-1))(1-0)} = \frac{(x+1)x}{2}$$

Donc

$$\begin{aligned} P_2(x) &= L_0(x)f(x_0) + L_1(x)f(x_1) + L_2(x)f(x_2) \\ &= \frac{x(x-1)}{2}f(x_0) + \frac{(x+1)(x-1)}{-1}f(x_1) + \frac{(x+1)x}{2}f(x_2) \\ &= 2\frac{x(x-1)}{2} + \frac{(x+1)(x-1)}{-1} - \frac{(x+1)x}{2} \\ &= -\frac{1}{2}x^2 - \frac{3}{2}x + 1 \end{aligned}$$

On vérifie facilement que :

$$P_2(x_0) = P_2(-1) = \frac{(-)(-1-1)}{2} = 2 = f(x_0)$$
$$P_2(x_1) = P_2(0) = \frac{(1)(-1)}{-1} = 1 = f(x_1)$$
$$P_2(x_2) = P_2(1) = -\frac{(1+1)1}{2} = -1 = f(x_2)$$

Propriétés 3.1.1. :
Les polynômes de Lagrange ont les propriétés suivantes :

$P1)$ $L_j(x)$ est un polynôme de degré n ; $\forall j = 0, ..., n$
$P2)$ $L_j(x_j) = 1 \quad \forall j = 0, ..., n$ et $L_j(x_i) = 0$ pour tout $j \neq i$
$P3)$ la famille $\{L_0(x), L_1(x), ...L_n(x)\}$ est une base de $\mathbb{P}_n(x)$

Preuve :

$P1)$ Par définition $L_j(x)$ est un polynôme de degre n
$P2)$ Cette propriété est aisément vérifiée en remplaçant x par x_i dans $L_j(x)$
$P3)$ On a : $card\ \{L_0(x), L_1(x), ...L_n(x)\} = \dim \mathbb{P}_n(x) = n + 1$
Pour montrer $P3)$, il suffit donc de montrer que $\{L_0(x), L_1(x), ...L_n(x)\}$ est un systéme libre.
Soient $C_0, C_1, ..., C_n$ des constantes telles que :
$C_1L_0(x) + L_1L_1(x) + ... + C_nL_n(x) = 0$

Alors, en prenant succéssivement $x = x_0 ... x_i \; .. \;\; x_n$ et en utilisant $L_j(x_i) = \delta_{ij}$
on déduit que : $C_0 = C_1 = ... = C_n = 0$
La famille $\{L_0(x), L_1(x), ...L_n(x)\}$ est libre et par conséquent c'est une base de $\mathbb{P}_n(x)$.

Définition 3.1.3. : **Differences divisées**
Soit f une fonction numérique définie sur un intervalle $[a, b]$ contenant $n+1$ points distincts $x_0, x_1, ..., x_n$.
On définit les differences divisées d'ordre i de f aux points (x_k) comme suit :
$[f(x_0)] = f(x_0)$
$[f(x_0), f(x_1)] = \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$
$[f(x_0), f(x_1), ..., f(x_i)] = \dfrac{[f(x_1), f(x_2), ..., f(x_i)] - [f(x_0), f(x_1), ..., f(x_{i-1})]}{x_i - x_0}$ pour $i \geq 2$

Exemple 3.1.2. :
Si $x_0 = -1, \quad x_1 = 0$ et $\quad x_2 = 1$, $f(x_0) = 2$, $\quad f(x_1) = 1$, $f(x_2) = -1$, on obtient

$$[f(-1)] = 2$$
$$[f(-1), f(0)] = \frac{1-2}{0-(-1)} = -1$$
$$[f(0)] = 1 \qquad\qquad [f(-1), f(0), f(1)] = \frac{-2-(-1)}{1-(-1)} = \frac{-1}{2}$$
$$[f(0), f(1)] = \frac{-1-1}{1-0} = -2$$
$$[f(1)] = -1$$

Propriétés 3.1.2. :
La valeur d'une difference divisée est indépendante de de l'ordre des x_i
On a ainsi :
$$[f(x_0), f(x_1)] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_1)}{x_1 - x_0} + \frac{f(x_0)}{x_0 - x_1} = [f(x_1), f(x_0)]$$

$$\begin{aligned} [f(x_0), f(x_1), f(x_2)] &= \frac{f(x_2)}{(x_2 - x_1)(x_2 - x_0)} + \frac{f(x_1)}{(x_1 - x_2)(x_1 - x_0)} + \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} \\ &= [f(x_2), f(x_1), f(x_0)] = [f(x_1), f(x_0), f(x_2)] \end{aligned}$$

et de façon générale :

$$[f(x_0), f(x_1), ...f(x_k)] = \sum_{i=0}^{i=k} \frac{f(x_i)}{(x_i - x_0)....(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_k)}$$

Définition 3.1.4. : **Interpolant de Newton :**
On appelle interpolant de Newton le polynôme P_n donné par :

$$P_n(x) = f(x_0) + [f(x_0), f(x_1)](x - x_0) + ... + [f(x_0), ..., f(x_n)](x - x_0).....(x - x_{n-1})$$

Exemple 3.1.3. :
Si $x_0 = -1, \quad x_1 = 0 \quad , \quad x_2 = 1, \quad f(x_0) = 2 \, , \quad f(x_1) = 1 \, , f(x_2) = -1 \, ,$ on obtient

$[f(-1)] = 2$

$$[f(-1), f(0)] = \frac{1-2}{0-(-1)} = -1$$

$[f(0)] = 1$

$$[f(-1), f(0), f(1)] = \frac{-2-(-1)}{1-(-1)} = \frac{-1}{2}$$

$$[f(0), f(1)] = \frac{-1-1}{1-0} = -2$$

$[f(1)] = -1$

$$\begin{aligned} P_2(x) &= f(x_0) + [f(x_0), f(x_1)](x - x_0) + [f(x_0), f(x_1), f(x_2)](x - x_0)(x - x_1) \\ &= 2 - 1(x - x_0) - \frac{1}{2}(x - x_0)(x - x_1) = 2 - (x + 1) - \frac{1}{2}(x + 1)(x) \\ &= 1 - \frac{3}{2}x - \frac{1}{2}x^2 \end{aligned}$$

Définition 3.1.5. : Base de Newton
Soient $x_0, x_1, ..., x_n \quad (n+1)$ points deux à deux distincts d'un intervalle $[a, b]$ de $\mathbb{R}$ et les polynômes N_i définis pour $i = 0, ..., n \quad$ par :

$N_0(x) = 1$

$N_j(x) = (x - x_0)(x - x_1).....(x - x_{j-1})$ pour $j = 1, ..., n$

On a en particulier

$N_1(x) = (x - x_0)$

$N_n(x) = (x - x_0)(x - x_1).....(x - x_{n-1})$

Propriétés 3.1.3. :
Les polynômes N_i ont les propriétés suivantes :
$P1)$ $N_i(x)$ est un polynôme de degré i
$P2)$ Pour $i \geq 1$, $N_i(x)$ admet $x_0, x_1, ..., x_{i-1}$ comme racines
$P3)$ la famille $\{N_0(x), N_1(x), ...N_n(x)\}$ est une base de $\mathbb{P}_n(x)$ dite base de Newton

Preuve :
$P1)$ Propriété évidente d'après la définition de $N_i(x)$
$P2)$ Propriété également évidente d'après la définition de $N_i(x)$
$P3)$ On a : $card \, \{N_0(x), N_1(x), ...N_n(x)\} = \dim \mathbb{P}_n(x) = n + 1$
Pour montrer $P3)$, il suffit donc de montrer que $\{N_0(x), N_1(x), ...N_n(x)\}$

est un systéme libre.

Soient $C_0, C_1, ..., C_n$ des constantes telles que :

$C_1N_0(x) + C_1N_1(x) + ... + C_nN_1(x) = 0$

Posons $F(x) = C_1N_0(x) + C_1N_1(x) + ... + C_nN_1(x) = 0$

Comme les x_i sont supposés deux à deux distincts, on obtient successivement :

$F(x_0) = C_0N_0(x_0) = 0 \Longleftrightarrow C_0 = 0$

$F(x_1) = C_1N_1(x_0) = C_1(x_1 - x_0) = 0 \Longleftrightarrow C_1 = 0$

. . . . . .

$F(x_n) = C_nN_n(x_n = C_n(x_n - x_0)(x_n - x_1)...(x_n - x_{n-1}) = 0) \Longleftrightarrow C_n = 0$

La famille $\{N_0(x), N_1(x), ...N_n(x)\}$ est libre et par conséquent c'est une base de $\mathbb{P}_n(x)$.

Théorème 3.1.1. *:*

Soit f une fonction numérique définie sur un intervalle $[a, b]$.

Soit P_n un polynôme interpolant f en $(n + 1)$ points $x_0, x_1, ..., x_n$ de $[a, b]$

Alors :

a) On peut exprimer $P_n(x)$ comme combinaison linéaire des N_i de la base de Newton :

$$P_n(x) = D_0N_0(x) + D_1N_1(x) + ... + D_nN_1(x)$$

b) Les D_i sont des constantes qui peuvent etre détérminées en solvant un système linéaire.

Preuve :

a) Puisque $P_n(x) \in \mathbb{P}_n(x)$ et que $B_N = \{N_0(x), N_1(x), ...N_n(x)\}$ est une base de $\mathbb{P}_n(x)$,

on peut écrire $P_n(x)$ dans la base B_N

b) En écrivant $P_n(x) = D_0N_0(x) + D_1N_1(x) + ... + D_nN_1(x)$ pour $x = x_i$, $0 \leq i \leq n$ on obtient le système triangulaire inférieur suivant :

$$(S1)\begin{cases} P_n(x_0) = D_0 & = & f(x_0) \\ P_n(x_1) = D_0 + N_1(x_1)D_1 & = & f(x_1) \\ P_n(x_2) = D_0 + N_1(x_2)D_1 + N_2(x_2)D_2 & = & f(x_2) \\ \qquad\qquad \vdots & \vdots & \quad \vdots \\ P_n(x_n) = D_0 + N_1(x_n)D_1 + \cdots + N_2(x_n)D_n & = & f(x_n) \end{cases}$$

Les D_i solutions du système $(S1)$ sont données par :

$D_0 = [f(x_0)] = f(x_0)$

$D_1 = \dfrac{f(x_1) - f(x_0)}{N_1(x_1)} = \dfrac{f(x_1) - f(x_0)}{(x_1 - x_0)} = [f(x_0), f(x_1)]$

$D_i = \dfrac{[f(x_1), f(x_2), ..., f(x_i)] - [f(x_0), f(x_1), ..., f(x_{i-1})]}{x_i - x_0} = [f(x_0), f(x_1), ..., f(x_i)]$

pour $i \geq 2$

Exemple 3.1.4. :
Soit la fonction f telle que

X_k	0.15	2.30	3.15	4.85	6.25	7.95
$f(x)$	4.79867	4.49013	4.2243	3.47313	2.66674	1.51909

Donc Les Coefficients du polynôme d'interpolation de f dans la base de newton sont :
$D_0 = 4.798670 \quad D_1 = -0.143507 \quad D_2 = -0.056411 \quad D_3 = 0.001229$
$D_4 = 0.000104 \quad D_5 = -0.000002$
Et son graphe est donné par la figure (3.2)

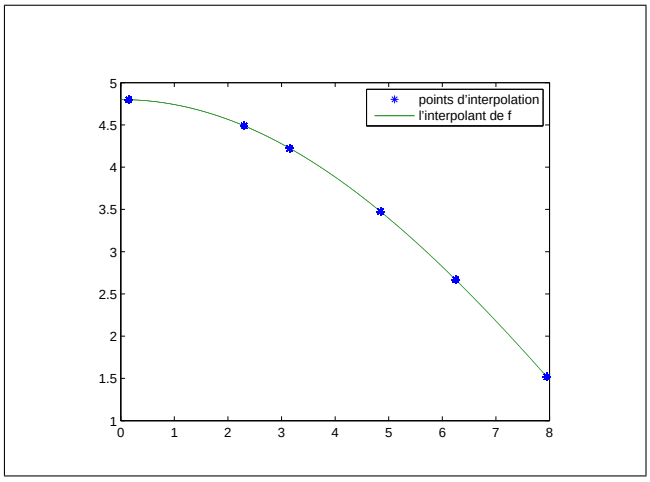


FIG. 3.1 – **Interpolation de Newton**

3.2 Interpolation de Lagrange

3.2.1 Existence et Unicité de l'interpolant

Théorème 3.2.1. *:*
Il existe un polynôme P_n unique de degré $\leq n$, interpolant f en $(n+1)$ points, c.a.d : tel que : $P_n(x_i) = f(x_i),\ x_i = 0, 1, ..., n$

Preuve :
i)Existence :

$$\text{Soit } L_i(x) = \frac{(x-x_0)(x-x_1)....(x-x_{i-1})(x-x_{i+1})....(x-x_n)}{(x_i-x_0)(x_i-x_1)....(x_i-x_{i-1})(x_i-x_{i+1})....(x_i-x_n)}$$

$$\begin{aligned} \text{et } P_n(x) &= L_0(x)f(x_0) + L_1(x)f(x_1) + ... + L_n(x)f(x_n) \\ &= \sum_{i=0}^{i=n} L_i(x)f(x_i) \\ &= \sum_{i=0}^{i=n} \left\{ \prod_{j=0, j\neq i}^{j=n} \frac{(x-x_j)}{(x_i-x_j)} \right\} f(x_i) \end{aligned}$$

Pour chaque $i = 0, ..., n$, L_i est un polynôme de degré n vérifiant : $L_j(x_i) = \delta_{ij}$ et par conséquent on a :

$$P_n(x_i) = f(x_i), \quad i = 0, 1, ..., n$$

ii) Unicité :
Supposons qu'il existe deux polynômes doifférents P_n et Q_n de degré $\leq n$, interpolant f aux points x_i. Alors, en posant $D_n(x) = P_n(x) - Q_n(x)$, on arrive à une contradiction.
En effet, D_n est un polynôme de degré $\leq n$ et par conséquent il peut avoir au plus n zéros mais d'un autre côté $D_n(x_i) = 0$ pour $i = 0, 1, ..., n$, ce qui voudrait dire que D_n aurait $(n+1)$ zéros d'où la contradiction.
Donc $P_n \equiv Q_n$

3.2.2 Interpolation linéaire

Dans ce cas, P_1 est un polynôme de degré 1 interpolant f aux points x_0 et x_1 on a donc $P_1(x_i) = f(x_i)$, $i = 0, 1$ et les polynômes de Lagrange donnés par :

$$L_0(x) = \frac{(x-x_1)}{(x_0-x_1)} \text{ et } L_1(x) = \frac{(x-x_0)}{(x_1-x_0)}$$

D'où : $P_1(x) = L_0(x)f(x_0) + L_1(x)f(x_1) = \frac{f(x_1)-f(x_0)}{(x_1-x_0)}(x-x_0) + f(x_0)$
qui est bien la formule d'interpolation linéaire qu'on obtient en cherchant la droite passant par x_0 et x_1
De façon similaire on peut exprimer P_1 dans la base de Newton pour obtenir :

$$\begin{aligned} P_1(x) &= f(x_0) + [f(x_0), f(x_1)](x-x_0) \\ &= f(x_0) + \frac{f(x_1)-f(x_0)}{(x_1-x_0)}(x-x_0) \end{aligned}$$

3.2.3 Erreur d'interpolation (de Lagrange)

Théorème 3.2.2. *:*
Soit le P_n le polynôme interpolant f aux points $a = x_0 < x_1 < .... < x_n = b$

Si $f \in C^{n+1}[a,b]$ *alors :*

a) $\forall x \in [a,b]$, $\exists \theta = \theta(x) \in [a,b]$ *tel que :*

$$e_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\theta)}{(n+1)!}\Pi_{n+1}(x)$$

avec : $\Pi_{n+1}(x) = \prod_{i=0}^{i=n}(x - x_i)$

b) *En posant* $M_{n+1} = \max_{a\leq x\leq b}|f^{(n+1)}(x)|$

on obtient :

$\max_{a\leq x\leq b}|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!}\max_{a\leq x\leq b}|\Pi_{n+1}(x)|$

et en particulier :

$$\max_{a\leq x\leq b}|e_n(x)| = \max_{a\leq x\leq b}|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!}(b-a)^{n+1}$$

Preuve :

a)

Si $x = x_i$ le résultat est évident.

Si $x \neq x_i$, posons :

$R(t) = f(t) - P_n(t) - \frac{f(x) - P_n(x)}{\Pi_{n+1}(x)}\Pi_{n+1}(t)$

On vérifie alors que $R \in C^{n+1}[a,b]$ et que :

$$R(x_i) = e_n(x_i) - e_n(x)\frac{\Pi_{n+1}(x_i)}{\Pi_{n+1}(x)} = 0, \quad i = 0, 1, ..., n,$$

et $\qquad R(x) = e_n(x) - e_n(x) = 0$

Par conséquent, R admet au moins $n+2$ zéros dans $[a,b]$ et par suite, en appliquant le théorème de Rolle de proche en proche, on montre qu'il existe un point $\theta \in [a,b]$ tel que : $R^{(n+1)}(\theta) = 0$. Le résultat annoncé en découle.

b) $R^{(n+1)}(\theta) = 0 \Rightarrow e_n(x) = \frac{f^{(n+1)}(\theta)}{(n+1)!}\Pi_{n+1}(x)$

$\Rightarrow \max_{a\leq x\leq b}|e_n(x)| \leq \frac{\max_{a\leq x\leq b}|\Pi_{n+1}(x)|}{(n+1)!}M_{n+1}$

Exercice 3.2.1. : Cas particulier : Points équidistants

Si les points sont équidistants et : $x_{i+1} - x_i = h \quad \forall i$, montrer que :

i) pour $n = 1$ on a : $\max_{a\leq x\leq b}|e_1(x)| \leq \frac{h^2}{8}M_2$

ii) pour $n = 3$ on a : $\max_{a\leq x\leq b}|e_3(x)| \leq \frac{h^4}{24}M_4$

Exemple 3.2.1. :

Construisons le graphe du polynôme d'interpolation de la fonction f dont on connaît les valeurs suivantes (voir figure (3.2)) :

X_k	-1	-0.5	0	0.5	1
$f(x)$	-1.5	0	0.25	0	0

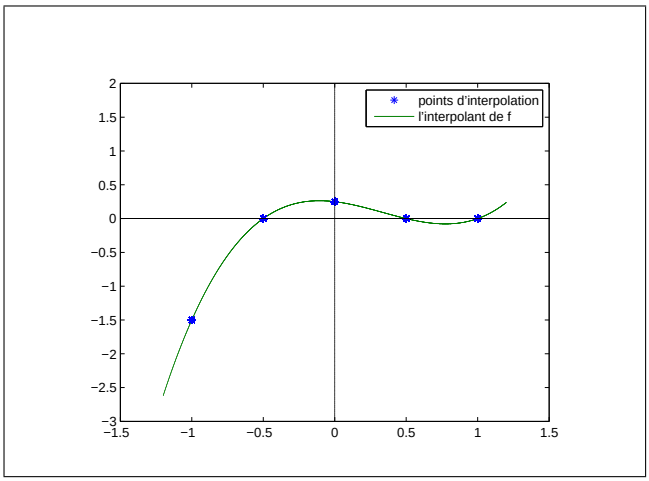


FIG. 3.2 – **Interpolation de Lagrange**

On peut aussi calculer les coefficients de ce polynôme. Le polynôme qui interpole f(x) est donc donné par :

$$P(x) = 0.000000x^4 + 1.000000x^3 + -1.000000x^2 + -0.250000x^1 + 0.250000$$

3.3 Exercices

Exercice 3.3.1. 1. Soit f est une fonction continue sur $[0, 3]$

a) Ecrire le polynôme de Lagrange interpolant f aux points $x_0 = 0$, $x_1 = 1$ et $x_2 = 2$

b) Ecrire le polynôme de Newton interpolant f aux points $x_0 = 0$, $x_1 = 1$ et $x_2 = 2$.

2. On considère un point supplémentaire $x_3 = 3$

a) Ecrire le pôlynome de Lagrange interpolant f aux points $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$

b) Ecrire le polynôme de Newton interpolant f aux points $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$

3. Comparer le temps de calcul entre 2.a) et 2.b)

Exercice 3.3.2. :

1. Calculer $I_1 = \int_0^1 t(t-1)dt$, $I_2 = \int_0^1 t(t-\frac{1}{2})dt$ et $I_3 = \int_0^1 t^2(t-1)^2dt$
2. En déduire que :
   i) $\int_\alpha^\beta (x-\beta)(x-\alpha)dx = (\beta-\alpha)^3 I_1$
   ii) $\int_\alpha^\beta (x-\alpha)(x-\frac{\alpha+\beta}{2})dx = (\beta-\alpha)^3 I_2$,
   iii) $\int_\alpha^\beta (x-\beta)^2(x-\alpha)^2dx = (\beta-\alpha)^5 I_3$
3. Montrer que si f est une fonction continue alors il existe $c \in [\alpha, \beta]$ tel que :
   $\int_\alpha^\beta (x-\beta)^2(x-\alpha)^2 f(x)dx = f(c)\frac{(\beta-\alpha)^5}{30}$
4. Peut-on dire qu'il existe $d \in [\alpha, \beta]$ tel que :
   $\int_\alpha^\beta (x-\alpha)(x-\frac{\alpha+\beta}{2})f(x)dx = f(d)\frac{(\beta-\alpha)^3}{12}$?
5. Soit f est une fonction continue sur $[a, b]$ et soient α et β deux réels dans $[a, b]$
   On suppose $\alpha < \beta$ et on note $P_2(x)$ le polynome interpolant f aux points :
   α , $\frac{\alpha+\beta}{2}$ et β.
   Exprimer $P_2(x)$ dans la base de Newton puis calculer $\int_\alpha^\beta P_2(x)dx$

Chapitre 4

Intégration numérique

4.1 Introduction

Dans le calcul d'intégrales, on n'est pas toujours en mesure d'obtenir des expressions exactes. Il se peut que l'obtention d'une primitive soit impossible ou trop compliquée.
Pour pallier à ce problème, on cherche une approximation de l'integrale $I(f) = \int_a^b f(x)dx$ par une somme de surfaces de rectangles, de trapèzes ou d'autres formes géométriques dont on sait calculer l'aire.
Considérons une subdivision uniforme de l'intervalle $[a, b]$ en n sous intervalles $[x_{i-1}, x_i]$, $i = 1, ..., n$ de même longueur $h = x_i - x_{i-1} = \dfrac{b-a}{n}$
On a donc : $x_0 = a < x_1 < \ ... \ x_i < x_{i+1} < \ ... \ < x_n = b$
où $x_i = a + ih$ pour $i = 0, 1, ..., n$, en particulier $x_0 = a$ et $x_n = b$
Soit f_i la restriction de la fonction f à chaque sous intervalle $[x_i, x_{i+1}]$.
En écrivant $\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + ... + \int_{x_{n-1}}^{x_n} f(x)dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx$,
on obtient alors des approximations de l'intégrale $I(f)$ en remplaçant $f_i(x)$ par une fonction $\Psi_i(x)$ facile à intégrer sur $[x_i, x_{i+1}]$.
Si Ψ_i est la fonction constante λ_i sur chaque sous intervale $[x_i, x_{i+1}]$ ($\Psi_i(x) = \lambda_i$)
on obtient une approximation par les sommes de Riemann :
$I(f) \approx R^n(f) = \sum_{i=0}^{n-1} (x_{i+1} - x_i)\lambda_i$ où $\lambda_i = f(x_i^*)$ avec x_i^* quelconque dans $[x_i, x_{i+1}]$

4.2 Approximation

Théorème 4.2.1. *:*
Soit f une fonction continue sur $[a,b]$ alors :

$$\lim_{n\to\infty} R^n(f) = I(f) = \int_a^b f(x)dx$$

Preuve :
Remarquons d'abord que f est uniformément continue sur $[a,b]$ et par conséquent :
$\forall \varepsilon > 0,\ \exists \eta > 0$ tel que $\forall (x,y) \in [a,b]$ vérifiant: $|x-y| \leq \eta$
on ait : $|f(x) - f(y)| \leq \varepsilon$
et plus particulièrement on a :
$\forall \varepsilon > 0,\ \exists \eta > 0$ tel que $\forall (x,y) \in [x_i, x_{i+1}]$ vérifiant: $|x-y| \leq \eta$
on ait : $|f(x) - f(y)| \leq \dfrac{\varepsilon}{b-a}$
Montrons maintenat que : $\forall \varepsilon > 0,\ \exists n_0 \in \mathbb{N}$ tel que $\forall n > n_0$
on ait : $|I(f) - R^n(f)| \leq \varepsilon$
Soit $n_0 > \dfrac{b-a}{\eta}$ alors pour tout $n \geq n_0$ on a : $h = \dfrac{b-a}{n} \leq \dfrac{b-a}{n_0} < \eta$
Par ailleurs, pour $x_i^* \in [x_i, x_{i+1}]$ et tout $x \in [x_i, x_{i+1}]$ on a : $|x - x_i^*| \leq h < \eta$ et ceci implique que $|f(x) - f(x_i^*)| \leq \dfrac{\varepsilon}{b-a}$
et par suite :

$$|I(f) - R^n(f)| \leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |f(x) - f(x_i^*)|\, dx \leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (\frac{\varepsilon}{b-a})dx$$
$$\leq \frac{\varepsilon}{b-a} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} dx \leq \frac{\varepsilon}{b-a} \sum_{i=0}^{n-1} (x_{i+1} - x_i) = \varepsilon$$

Cas particulier de sommes de Riemann
En particulier, sur chaque sous intervalle $[x_i, x_{i+1}]$,on peut choisir x_i^* et prendre pour constante λ_i les valeurs $f(x_i^*)$ suivantes :
a) $x_i^* = x_i$ et $\lambda_i = f(x_i)$ donne $\int_a^b f(x)dx \approx I_g^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_i)dx$
(I_g^n ,l'indice g pour signifier gauche)
b) $x_i^* = x_{i+1}$ et $\lambda_i = f(x_{i+1})$ donne $\int_a^b f(x)dx \approx I_d^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_{i+1})dx$
(I_d^n , l'indice d pour signifier droit)
c) $x_i^* = \bar{x}_i = \dfrac{1}{2}(x_i + x_{i+1})$ (milieu de $[x_i, x_{i+1}]$) et $\lambda_i = f(\dfrac{x_i + x_{i+1}}{2})$ donne $\int_a^b f(x)dx \approx$
$I_m^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(\dfrac{x_i + x_{i+1}}{2})dx$
(I_m^n , l'indice m pour signifier moyen ou médian)

4.2.1 Approximation par des rectangles à gauche

Soit $x_0 < x_1 < \ ... \ x_i < x_{i+1} < \ ... \ < x_n$ une subdivision uniforme de $[a, b]$
Si on prend $x_i^* = x_i$, on obtient une approxiamtion de $\int_a^b f(x)dx$ comme suit :
$\int_a^b f(x)dx \approx I_g^n$ où :

$$I_g^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_i^*)dx = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i) = h \sum_{i=0}^{n-1} f(x_i) = \frac{b-a}{n} \sum_{i=0}^{n-1} f(x_i)$$

Théorème 4.2.2. *:*
Soit f une fonction continue sur $[a, b]$ alors :

1) La suite (I_g^n) converge vers $I(f)$

2) Si la fonction f est de classe C^1 sur $[a, b]$ alors $\left| I(f) - I_g^n \right| \leq M_1 \frac{(b-a)^2}{2n}$
où $M_1 = \max_{a \leq t \leq b} |f'(t)|$.

Preuve :

1) analogue à celle du théorème 1

2) Si f est de classe C^1 sur $[a, b]$ alors : $\exists M_1 \geq 0$ tel que : $M_1 = \max_{a \leq t \leq b} |f'(t)|$
Le théorème des accroissements finis appliqué à la fonction f sur $[x_i, x]$ où $x \in [x_i, x_{i+1}]$
donne : $\exists c_i \in]x_i, x[$ tel que $f(x) - f(x_i) = (x - x_i) f'(c_i)$
d'où : $\left| I(f) - I_g^n \right| \leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} |(f(x) - f(x_i))| \, dx \leq \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \left| (x - x_i) f'(c_i) \right| dx$
soit encore :

$$\left| I(f) - I_g^n \right| \leq M_1 \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (x - x_i)dx$$
$$\leq M_1 \sum_{i=0}^{n-1} \frac{1}{2} [(x - x_i)^2]_{x_i}^{x_{i+1}} = M_1 \sum_{i=0}^{n-1} \frac{h^2}{2} = M_1 (b-a) \frac{h}{2} = M_1 \frac{(b-a)^2}{2n}$$

Corollaire 4.2.1. *:*
Pour obtenir une approximation avec précision de l'ordre de ε , il suffit de prendre $I_g^{n_0}$
où l'indice n_0 est tel que : $M_1 \frac{(b-a)^2}{2n_0} \leq \varepsilon$ *ou encore* $n_0 \geq M_1 \frac{(b-a)^2}{2\varepsilon}$

Exemple 4.2.1. :
L'approximation de l'intégrale $\int_0^1 e^{-x2} dx$ avec la méthode des rectangles à gauche, avec les précision $\varepsilon = 0.1$ et $\varepsilon = 0.05$

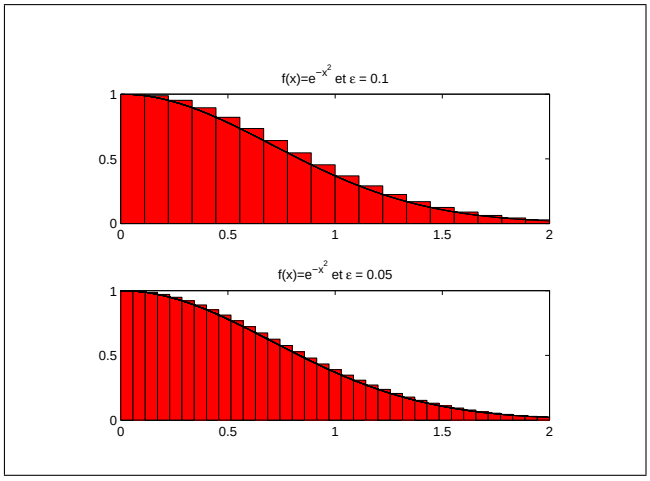


FIG. 4.1 – Approximation par des rectangles à gauche

4.2.2 Approximation par des rectangles à droite

Soit $x_0 < x_1 < \ ... \ x_i < x_{i+1} < \ \ ... \ < x_n$ une subdivision uniforme de $[a, b]$
Si on prend $x_i^* = x_{i+1}$, on obtient une approxiamtion de $\int_a^b f(x)dx$ comme suit :
$\int_a^b f(x)dx \approx I_d^n$ où :

$$I_d^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x_{i+1})dx = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_{i+1}) = \frac{b-a}{n} \sum_{i=1}^{i=n} f(x_i)$$

Théorème 4.2.3. *:*
Soit f une fonction continue sur $[a, b]$ alors :

1) La suite (I_d^n) converge vers $I(f)$

2) Si la fonction f est de classe C^1 sur $[a, b]$ alors $\left|I(f) - I_d^n\right| \leq M_1 \frac{(b-a)^2}{2n}$
où $M_1 = \max_{a \leq t \leq b} |f'(t)|$

Preuve : analogue à celle du théorème 2

Corollaire 4.2.2. *Pour obtenir une approximation avec précision de l'ordre de ε , il suffit de prendre $I_d^{n_0}$ où l'indice n_0 est tel que : $M_1 \frac{(b-a)^2}{2n_0} \leq \varepsilon$ ou encore $n_0 \geq M_1 \frac{(b-a)^2}{2\varepsilon}$*

Exemple 4.2.2. :
L'approximation de l'intégrale $\int_0^1 e^{-x2} dx$ avec la méthode des rectangles à droite, avec les précision $\varepsilon = 0.1$ et $\varepsilon = 0.05$

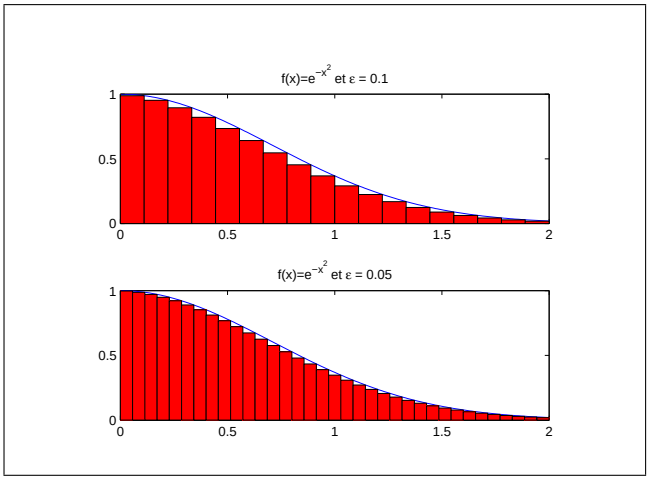


FIG. 4.2 – Approximation par des rectangles à droite

4.2.3 Approximation par des rectangles médians

Soit $x_0 < x_1 < \ldots x_i < x_{i+1} < \ldots < x_n$ une subdivision uniforme de $[a,b]$
Si on prend $x_i^* = \dfrac{x_i + x_{i+1}}{2}$, on obtient une approxiamtion de $\int_a^b f(x)dx$ comme suit :
$\int_a^b f(x)dx \approx I_m^n$ où :

$$I_m^n = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(\frac{x_i + x_{i+1}}{2})dx = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(\frac{x_i + x_{i+1}}{2}) = \frac{b-a}{n} \sum_{i=0}^{n-1} f(\frac{x_i + x_{i+1}}{2})$$

Théorème 4.2.4. *:*
Soit f une fonction continue sur $[a,b]$ alors :

1) La suite (I_m^n) converge vers $I(f)$

2) Si la fonction f est de classe C^2 sur $[a,b]$ alors $|I(f) - I_m^n| \leq M_2 \dfrac{(b-a)^3}{24n^2}$ où $M_2 = \max_{a \leq t \leq b} |f''(t)|$

Preuve :
1)analogue à celle du théorème 2
2) Ici, au lieu du théorème des accroissements finis, on utilise la formule de Taylor à l'ordre 2 sur l'intervalle $[\bar{x}_i, x]$ où $\bar{x}_i = \dfrac{1}{2}(x_i + x_{i+1})$ et $x \in [x_i, x_{i+1}]$.
On obtient l'expression suivante :
$f(x) - f(\bar{x}_i) = (x - \bar{x}_i) f'(\bar{x}_i) + \dfrac{1}{2}(x - \bar{x}_i)^2 f''(c_i)$ avec $c_i \in [\bar{x}_i, x]$
On a alors :

$$|I(f) - I_m^n| = \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (f(x) - f(\bar{x}_i))dx \right|$$

$$\leq \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}(x-\bar{x}_i)f'(\bar{x}_i)dx\right| + \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}\frac{1}{2}(x-\bar{x}_i)^2 f''(c_i)dx\right|$$
$$\leq A + B$$

où $A = \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}(x-\bar{x}_i)f'(\bar{x}_i)dx\right| = |f'(\bar{x}_i)|\left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}(x-\bar{x}_i)dx\right| = 0$

et $B = \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}\frac{1}{2}(x-\bar{x}_i)^2 f''(c_i)dx\right| \leq \frac{1}{2}M_2\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}(x-\bar{x}_i)^2 dx$

$$\leq \frac{1}{2}M_2\sum_{i=0}^{n-1}\frac{1}{3}[(x-\bar{x}_i)^3]_{x_i}^{x_{i+1}} = \frac{1}{6}M_2\sum_{i=0}^{n-1}2(\frac{x_{i+1}-x_i}{2})^3 = M_2\frac{(b-a)^3}{24n^2}$$

Corollaire 4.2.3. *:*

Pour obtenir une approximation avec précision de l'ordre de ε , il suffit de prendre $I_m^{n_0}$ où l'indice n_0 est tel que : $M_2\frac{(b-a)^3}{24n_0^2} \leq \varepsilon$ c-a-d $n_0^2 \geq M_2\frac{(b-a)^2}{24\varepsilon}$

ou encore $n_0 \geq \sqrt{M_2\frac{(b-a)^2}{24\varepsilon}}$

Exemple 4.2.3. :

L'approximation de l'intégrale $\int_0^1 e^{-x2}dx$ avec la méthode des rectangles médians, avec les précision $\varepsilon = 0.01$ et $\varepsilon = 0.001$

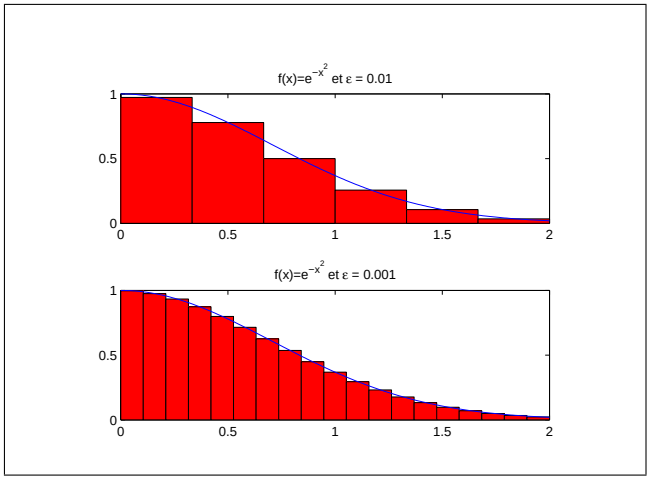


FIG. 4.3 – Approximation par des rectangles médians

4.2.4 Approximations par des trapèzes

Soit $x_0 < x_1 < \ldots x_i < x_{i+1} < \ldots < x_n$ une subdivision uniforme et $P_1(x)$ un polynôme de degré 1 interpolant f aux points x_i et x_{i+1} de chaque intervalle $[x_i, x_{i+1}]$.

$[P_1(x_i) = f(x_i)$ et $P_1(x_{i+1}) = f(x_{i+1})]$
En approchant sur chaque sous intervalle $[x_i, x_{i+1}]$, $f(x)$ par $P_1(x)$ on obtient :

$f(x) \simeq P_1(x) = f(x_i) + [f(x_i), f(x_{i+1})](x - x_i)$

$f(x) \simeq P_1(x) = f(x_i) + \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}(x - x_i)$

et en conséquences :

$I(f) = \int_a^b f(x)dx \simeq \int_a^b P_1(x)dx = \dfrac{h}{2}[f(x_i) + f(x_{i+1})]$

4.2.5 Formule de Simpson

Soit $P_2(x)$ un polynôme de degré 2 vérifiant :

$P_2(x_i) = f(x_i)$, $P_2(x_{i+1}) = f(x_{i+1})$ et $P_2(x_{i+2}) = f(x_{i+2})$

En approchant sur chaque sous intervalle $[x_i, x_{i+2}]$, $f(x)$ par $P_2(x)$ on obtient :

$f(x) \simeq P_2(x)$

$\simeq f(x_0) + [f(x_i), f(x_{i+1})](x - x_i) + [f(x_i), f(x_{i+1}), f(x_{i+2})](x - x_i)(x - x_{i+1})$

et en conséquences :

$I(f) = \int_{x_i}^{x_{i+2}} f(x)dx \simeq \int_{x_i}^{x_{i+2}} P_2(x)dx = \dfrac{h}{3}[f(x_i) + 4f(x_{i+1}) + f(x_{i+2})]$

Preuve :

On a $\int_{x_i}^{x_{i+2}} P_2(x)dx = I(P_2) + J(P_2) + K(P_2)$

où : $I(P_2) = \int_{x_i}^{x_{i+2}} f(x_i)\, dx$

$J(P_2) = \int_{x_i}^{x_{i+2}} \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}(x - x_i)\, dx$

$K(P_2) = \int_{x_i}^{x_{i+2}} [f(x_i), f(x_{i+1}), f(x_{i+2})](x - x_i)(x - x_{i+1})dx$

On fait le changement de variables suivant :

$(x - x_i) = ht \rightarrow dx = hdt$

quand $x = x_i \rightarrow t = 0$ et si $x = x_{i+2} \rightarrow t = 2$

d'où on obtient successivement :

$I(P_2) = \int_0^2 f(x_i)\, hdt = 2hf(x_i)$

$J(P_2) = \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \int_0^2 h^2 t dt$

$= h[f(x_{i+1}) - f(x_i)] \left[\dfrac{t^2}{2}\right]_0^2$

$= 2h[f(x_{i+1}) - f(x_i)]$

$K(P_2) = [f(x_i), f(x_{i+1}), f(x_{i+2})] \int_0^2 (x - x_i)(x - x_{i+1})dx$

$= [f(x_i), f(x_{i+1}), f(x_{i+2})] \int_0^2 h^3 t(t-1)dt$

$= \dfrac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{2h^2} h^3 \left[\dfrac{t^3}{3} - \dfrac{t^2}{2}\right]_0^2$

$= [\, f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)]\dfrac{h}{3}$

Soit enfin :

$$\int_{x_i}^{x_{i+2}} P_2(x)dx = I(P_2) + J(P_2) + K(P_2) = \frac{h}{3}[f(x_i) + 4f(x_{i+1}) + f(x_{i+2})]$$

4.3 Interpolation et Erreur d'intégration numérique

Définition 4.3.1. :
On appelle formule de quadrature de type interpolation la formule :
$\int_a^b f(x)dx \simeq \int_a^b P_n(x)dx$
où P_n est le polynôme d'interpolation associé à f.

4.4 Applications :

4.4.1 Interpolation linèaire et la formule du trapèze :

Soit $a = x_0 < x_1 < \ ... \ < x_n = b$, une subdivision uniforme de $[a, b]$ et $f \in C^2([a, b])$
Considérons d'abord le sous intervalle $[\ x_i, x_{i+1}]$, avec $h = x_{i+1} - x_i$,
Soit P_1 le polynôme de degré 1 interpolant f aux points x_i et x_{i+1}
Alors l'intégrale $I_i(f) = \int_{x_i}^{x_{i+1}} f(x)dx$ peut être approchée par :
$I_i(f) \simeq \int_{x_i}^{x_{i+1}} P_1(x)dx = \frac{h}{2}[f(x_i) + f(x_{i+1})]$
L'erreur commise par cette approximation étant donnée par :
$E_i(f) = \int_{x_i}^{x_{i+1}} e_1(x)dx = \frac{1}{2}\int_{x_i}^{x_{i+1}}(x - x_i)(x - x_{i+1})f''(\theta_i)dx$
$\Pi_2(x) = (x - x_i)(x - x_{i+1})$ garde un signe constant dans $[x_i, x_{i+1}]$
d'où, en appliquant la formule de la moyenne :
$E_i(f) = \frac{f''(\eta_i)}{2}\int_{x_i}^{x_{i+1}}(x - x_i)(x - x_{i+1})dx = -\frac{h^3}{12}f''(\eta_i)\,;\,\eta_i \in [x_i, x_{i+1}]$

4.4.2 Formule du trapèze composée

Pour chercher une approximation de l'intégrale sur tout l'intervalle $[a, b]$, il suffit d'écrire :

$$\begin{aligned}\int_{a=x_0}^{b=x_n} f(x)dx &= \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx \\ &= \sum_{i=0}^{n-1}[\frac{h}{2}[f(x_i) + f(x_{i+1})] + \sum_{i=0}^{n-1} -\frac{h^3}{12}f''(\eta_i) \\ &= \frac{h}{2}[f(x_0) + 2f(x_1) + ...2f(x_{n-1}) + f(x_n)] - \frac{h^3}{12}\sum_{i=0}^{n-1} f''(\eta_i) \\ &= \frac{h}{2}[f(x_0) + 2f(x_1) + ...2f(x_{n-1}) + f(x_n)] - \frac{(b-a)}{12}h^2 f''(\eta)\end{aligned}$$

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

4.4.3 Erreur de la formule de Simpson

En posant : $e_2(x) = f(x) - P_2(x)$ et $E_2(f) = \int_a^b e_2(x)dx = \int_a^b [f(x) - P_2(x)]dx$

On démontre que : $E_2(f) = -\frac{h^5}{90} f^{(4)}(\theta)\,;\theta \in [a,b]$

Exercice 4.4.1. :

Soit $f \in C^2([a,b])$

En posant $I_n = \frac{h}{2}\sum_{i=0}^{n-1}[f(x_{i+1}) + f(x_i)]$, montrer que :

1) $I^n = \frac{I_g^n + I_d^n}{2}$

2) $\lim_{n\to\infty} I^n = I(f)$

3) $|I(f) - I^n| \leq M_2 \frac{(b-a)^3}{12n^2}$

Exemple 4.4.1. :

On veut calculer une valeur approcher de ln 2 pour cela on va calculer l'intégrale de la fonction f(x)=1/(x+1) sur un intervalle [0,1] par la méthode de Simpson (voir **??**) la méthode des trapèzes (voir **??**), et la méthode des rectangles. Et on compare les résultats avec différentes valeurs de n (le nombre de subdivision de l'intervalle [0, 1]) alors on a le tableau suivant :

n	méthode de Simpson	méthode des trapèzes	méthode des rectangles
100	0.69314718057947511	0.6931534304818241	0.69565343048182404
200	0.69314718056116631	0.69314874305506269	0.69439874305506266
400	0.69314718056002178	0.69314757118464021	0.69377257118464031
800	0.69314718055995039	0.69314727821617628	0.69345977821617644
1600	0.69314718055994573	0.69314720497400739	0.69330345497400736
3200	0.6931471805599464	0.69314718666346065	0.69322531166346069
6400	0.69314718055994629	0.69314718208582515	0.69318624458582534
12800	0.69314718055994318	0.69314718094141725	0.69316671219141723
25600	0.69314718055994495	0.69314718065531489	0.69315694628031488
51200	0.69314718055994629	0.6931471805837861	0.69315206339628599

TAB. 4.1 – $Ln(2) \simeq 0,69314718055994530941723212145818$

4.5 Exercices

Exercice 4.5.1. :

Soit $a = x_0, x_1 .... , x_n = b$, une subdivision de $[a,b]$ et $f \in C^2([a,b])$

1. En faisant deux intégrations par parties succéssives, montrer que :
   $\int_{x_i}^{x_{i+1}} f(x)dx = (x_{i+1}-x_i)\frac{f(x_i)+f(x_{i+1})}{2} + \frac{1}{2}\int_{x_i}^{x_{i+1}}(x-x_i)(x-x_{i+1})f^{(2)}(x)dx$
2. Si $x_{i+1} - x_i = h$ pour tout $i = 0, 1, ..., n-1$, montrer que l'erreur d'approximation de $\int_{x_i}^{x_{i+1}} f(x)dx$ par la méthode des trapèzes est de la forme :
   $-\frac{h^3}{12}f^{(2)}(\theta)$, $\theta \in [x_i, x_{i+1}]$

Exercice 4.5.2. Soit P_3 un polynôme de degré 3 et $f \in C^2([a,b])$ avec $f^{(2)} < 0$,
Soit $I(f) = \int_a^b f(x)dx$ et I_n^T l'approximation de $I(f)$ par la méthode des trapèzes,
Soit I_m^n (resp. I_g^n) l'approximation de $I(f)$ par la méthode des triangles droits (resp. gauches)

1. Ecrire les expressions de I_g^n, I_m^n et I_n^T
2. Montrer que $I_n^T = \frac{I_g^n + I_d^n}{2}$
3. En déduire que $\lim_{n\to\infty} I_n^T = I(f)$
4. Prouver que $\left|I(f) - I_n^T\right| \leq M_2\frac{(b-a)^3}{12n^2}$
5. Etablir les inégalités : $(b-a)f(\frac{b+a}{2}) \geq \int_a^b f(x)dx \geq (b-a)(\frac{f(b)+f(a)}{2})$
6. Etablir l'égalité : $\int_a^b g(x)dx = \frac{(b-a)}{6}[P_3(a) + 4P_3(\frac{b+a}{2}) + P_3(b)]$
7. Soit $a = x_0 < x_1 < ... < x_n = b$, une subdivision de $[a,b]$
   Montrer que :
   $\int_{x_i}^{x_{i+1}} f(x)dx = (x_{i+1}-x_i)\frac{f(x_i)+f(x_{i+1})}{2} + \frac{1}{2}\int_{x_i}^{x_{i+1}}(x-x_i)(x-x_{i+1})f^{(2)}(x)dx$
8. Si $x_{i+1} - x_i = h$ pour tout $i = 0, 1, ..., n-1$, montrer que l'erreur d'approximation de $\int_{x_i}^{x_{i+1}} f(x)dx$ par la méthode des trapèzes est de la forme :
   $-\frac{h^3}{12}f^{(2)}(\theta)$, $\theta \in [x_i, x_{i+1}]$

Chapitre 5

Méthodes directes de résolution des systèmes linéaires $Ax = b$

Dans ce chapitre, on s'intéresse à la résolution numérique d'un système linéaire

$$Ax = b \tag{5.0.1}$$

où A est une matrice carrée supposée inversible, b un vecteur second membre et x le vecteur des inconnues, $A = \left(a_{ij}\right)_{i,j=1,\cdots,n}$, $b = (b_1, \cdots, b_n)^\top$, $x = (x_1, \cdots, x_n)^\top$.
Théoriquement, le fait que A soit inversible entraîne que le système (5.0.1) admet une solution unique $x = A^{-1}b$.
Mais cette écriture suppose que l'on dispose de la matrice A^{-1}, or l'obtention de A^{-1} est équivalente à la résolution de n systèmes linéaires, $A.\left(A^{-1}\right)_j = e_j = (0, \cdots, 1, 0, \cdots, 0)^\top$ en plus de la multiplication $x = A^{-1}b$.
Une autre méthode consisterait à obtenir les x_i à l'aide des formules de Cramer $x_i = \dfrac{\det(A_i)}{\det(A)}$ où $\det(A_i)$ désigne le déterminant de la matrice obtenue en remplaçant la i^{eme} colonne de A par le vecteur b.
Le calcul de chaque déterminant nécessite $n.n!$ multiplications et $(n! - 1)$ additions.
Soit au total : $(n+1)!n$ multiplications, $(n+1)(n! - 1)$ additions et n divisions.
A titre d'exemple, on a besoin de 4319 opérations si $n = 5$ et environ 41000 opérations pour $n = 10$. Comme les problèmes d'analyse numérique donnent lieu à des matrices de grandes tailles (n assez grand), la méthode de Cramer et les méthodes similaires s'avèrent inutilisables.

5.1 Résolution d'un système par les méthodes de descente ou de remontée

C'est le cas où on a à résoudre un système de la forme

$$Ux = b \tag{5.1.1}$$

ou

$$Lx = b \tag{5.1.2}$$

avec U triangulaire supérieure et L triangulaire inférieure.
Si on prend l'équation (5.1.1) par exemple on obtient

$$Ux = b \Leftrightarrow \begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n & = b_1 \\ 0 \quad + u_{22}x_2 + \cdots + u_{2n}x_n & = b_2 \\ 0 \qquad\qquad \ddots & \vdots \\ \qquad\qquad\qquad\qquad u_{nn}x_n & = b_n \end{cases}$$

En supposant que les u_{kk} sont non nuls, on obtient les x_i de façon évidente en commençant par le bas et en remontant. On obtient ainsi $x_n = b_n/u_{nn}$ puis
$x_i = \left(b_i - \sum_{j=i+1}^{n} u_{ij}x_i \right) / u_{ii}$, pour $i = n-1$ à 1.
L'algorithme de résolution est le suivant :
$x_n = b_n/u_{nn}$
Pour $i = n-1$ *à* 1
    $x_i = b_i$
    Pour $j = i+1$ *à* n
        $x_i = x_i - u_{ij} * x_j$
    fin j
    $x_i = x_i/u_{ii}$
fin i
Le nombre d'opérations nécessaire est : $\frac{n(n-1)}{2}$ multiplications $\frac{n(n-1)}{2}$ additions et n divisions. Soit au total n^2 opérations.

Remarque 5.1.1. Le cas d'un système avec matrice triangulaire inférieure se traite de façon similaire en obtenant x_1 d'abord puis en descendant.

5.2 Matrices élémentaires

5.2.1 Matrices élémentaires de Gauss

Soient les matrices

$$M_1 = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & 0 & \\ \vdots & 0 & \ddots & \\ -m_{n1} & & & 1 \end{pmatrix}, \qquad M_k = \begin{pmatrix} 1 & & & & \\ 0 & \ddots & & 0 & \\ 0 & & 1 & & \\ \vdots & & -m_{k+1k} & \ddots & \\ \vdots & & \vdots & & \ddots \\ 0 & & -m_{k+1k} & & 1 \end{pmatrix} \begin{matrix} \\ \\ k \\ \\ \\ \\ \end{matrix}$$

k

en posant $e_k = (0, \cdots, 1, 0, \cdots, 0)^\top$ et $m_k = (0, \cdots, 0, -m_{k+1k}, \cdots, -m_{nk})^\top$, on obtient $M_k = I - m_k e_k^\top$ et on vérifie facilement que M_k est inversible et que $M_k^{-1} = I + m_k e_k^\top$.

5.2.2 Matrices élémentaires de Danilevski

Elles sont de la forme :

$$M_k = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{k1} & m_{k2} & \cdots & \ddots & m_{kn} \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

et permettent de transformer une matrice A en une matrice de Frobenius P de la forme

$$P = \begin{pmatrix} p_1 & p_2 & \cdots & \cdots & p_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}.$$

Ces matrices seront utilisées dans le chapitre 6.

5.2.3 Matrices élémentaires de Householder

Ces matrices sont de la forme

$$P_k = I_n - 2\omega_k \omega_k^\top \quad \text{avec } \omega_k = (0, \cdots, 0, \omega_{k+1k}, \cdots, \omega_{nk})^\top \quad \text{et } \omega_k^\top \omega_k = 1.$$

On vérifie aussi qu'on a $P_k = P_k^{-1} = P_k^\top$, P_k est donc une matrice orthogonale symétrique. Elle peut aussi s'écrire sous la forme explicite et en blocs :

$$P_k = \left(\begin{array}{c|c} I_k & \\ \hline & \widehat{P}_k \end{array}\right)$$

avec $\widehat{P}_k = I_{n-k} - 2\widehat{\omega}_k\widehat{\omega}_k^\top$ et $\widehat{\omega}_k = (\omega_{k+1k}, \cdots, \omega_{nk})^\top$.

5.2.4 Matrices élémentaires de permutation

Elles sont de la forme

$$I_{k,l} = \begin{array}{c} \\ \\ \\ \\ k \\ l \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccc} k & & & & l & & & \end{array} \\ \left(\begin{array}{cccccccc} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & & & & & & \\ & & & 0 & 1 & & & \\ & & & 1 & 0 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{array}\right) \end{array}$$

Remarque 5.2.1. $I_{k,l}A$ échange les lignes k et l de A alors que $AI_{k,l}$ échange les colonnes k et l de A . On a encore $I_{k,l} = I_{k,l}^{-1} = I_{k,l}^\top$.

Exemple 5.2.1. Soit A la matrice donnée par : $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ et $I_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

alors $I_{13}A = \begin{pmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 3 \end{pmatrix}$ et $AI_{13} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{pmatrix}$.

Définition 5.2.1. Une matrice de permutation est un produit de matrices élémentaires de permutation.

5.2.5 Matrices élémentaires de Perlis

Soient les matrices suivantes :

- I_{ij}, une matrice I_n dont on a permuté les i^{eme} et j^{eme} lignes ;
- $I_i(d)$, une matrice I_n dont la i^{eme} ligne a été multipliée par un scalaire d ;
- $I_{il}(d)$, une matrice I_n dont la i^{eme} ligne est multipliée par : $e_{ij} + de_{lj}$, où $I_{ij} = \left(e_{ij}\right)_{i,j=1,\cdots,n}$;

Les matrices I_{ij}, $I_i(d)$, $I_{il}(d)$ s'appellent matrices élémentaires de Perlis.

Exemple 5.2.2. $I_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, $I_2(d) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $I_{12}(d) = \begin{pmatrix} 1 & d & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.
Ces matrices sont régulières et leurs inverses s'écrivent :

$$\begin{aligned} I_{ij}^{-1} &= I_{ij} \\ I_i^{-1}(d) &= I_i(1/d) \\ I_{il}^{-1}(d) &= I_{il}(-d) \end{aligned}$$

Les transformations élémentaires sur une matrice A peuvent se ramener à la pre-multiplication de A par l'une des matrices élémentaires précédentes. Ainsi avec $A' = I_{ik}A$, A' est la matrice obtenue après permutation des lignes i et k.
$A' = I_{il}(d)A$, A' est la matrice dont la i^{eme} ligne est remplaçée par la somme de la i^{eme} ligne de A et d fois la l^{eme} ligne de A.

Remarque 5.2.2. On peut aussi définir des matrices élémentaires pour les opérations sur les colonnes, on les note P_{ij}, $P_j(d)$, $P_{jl}(d)$.

Exemple 5.2.3. $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.
Ici la matrice élémentaire M_1 de Gauss est donnée par

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ -7 & 0 & 1 \end{pmatrix}.$$

Par ailleurs

$$A' = I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 7 & 8 & 9 \end{pmatrix}$$

$$\text{et} \quad A'' = I_{31}(-7)I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{pmatrix}$$

$$\text{et on a} \quad A^{(2)} = M_1A = I_{31}(-7)I_{21}(-4)A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -6 & -12 \end{pmatrix}.$$

5.2.6 Matrices élémentaires de Givens (ou de rotation)

Elles sont données par

$$R_{k,l} = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & & & & & & \\ & & c & s & & & & k \\ & & -s & c & & & & l \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \end{pmatrix} \qquad \text{avec } c = \cos\theta,\ \ s = \sin\theta \ \text{ et } \ k \neq l$$

Remarque 5.2.3. On vérifie facilement que $R_{k,l}^{\top} = R_{k,l}^{-1}$.

5.3 Méthodes de Gauss

5.3.1 Méthode de Gauss sans pivot

Elle consiste à transformer le système $Ax = b \quad (1)$ en un système $Ux = c, \quad (2)$ avec U triangulaire supérieure puis à résoudre le nouveau système par la méthode de remontée.

Soit (S_1) le système de départ :

$$(S_1) \begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n & = b_1^{(1)} \\ \vdots & \vdots \\ a_{n1}^{(1)} x_1 + a_{n2}^{(1)} x_2 + \cdots + a_{nn}^{(1)} x_n & = b_n^{(1)} \end{cases}$$

On pose $m_{i1} = \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ en supposant que $a_{11}^{(1)} \neq 0\ , \quad i = 2, \cdots, n.$

Ensuite, on remplace la ligne L_i par $L_{i'} = L_i - m_{i1} * L_1$, ce qui donne :

$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} * a_{1j}^{(1)} \quad i = 2, \cdots, n$ et $j = 2, \cdots, n$

$b_i^{(2)} = b_i^{(1)} - m_{i1} * b_1^{(1)} \quad i = 2, \cdots, n$

On obtient alors le systéme (S_2) suivant :

$$(S_2) \begin{cases} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n & = b_1^{(1)} \\ 0 \qquad a_{22}^{(2)} x_2 + \cdots + a_{2n}^{(2)} x_n & = b_2^{(2)} \\ \vdots \qquad \vdots & \vdots \\ 0 \qquad a_{n2}^{(2)} x_2 + \cdots + a_{nn}^{(2)} x_n & = b_n^{(2)} \end{cases}$$

On suppose que $a_{22}^{(2)} \neq 0$ et on recommence avec $m_{i2} = \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, $i = 3, \cdots, n$.

A l'étape k, le système se présente sous la forme :

$$k \quad (S_k) \begin{cases} a_{11}^{(1)} x_1 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ 0 \quad \ddots \quad \vdots \\ \vdots \quad a_{kk}^{(k)} x_k + \cdots + a_{kn}^{(k)} x_n = b_k^{(k)} \\ \vdots \quad \vdots \quad \vdots \\ 0 \quad a_{nk}^{(k)} x_k + \cdots + a_{nn}^{(k)} x_n = b_n^{(k)} \end{cases} \quad k$$

En supposant $a_{kk}^{(k)} \neq 0$, on pose $m_{ik} = \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$, $\quad i = k+1, \cdots, n$ puis on remplace la ligne L_i par la ligne $L_{i'} = L_i - m_{ik} * L_k$.

On aboutit alors au système final (S_n) de la forme :

$$(S_n) \Longleftrightarrow Ux = c \Longleftrightarrow \begin{cases} a_{11}^{(1)} x_1 + \cdots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ 0 \quad \ddots \quad \vdots \\ \vdots \quad 0 + a_{kk}^{(k)} x_k + \cdots + a_{kn}^{(k)} x_n = b_k^{(k)} \\ \vdots \quad \vdots \quad \vdots \\ 0 \cdots \quad 0 + \cdots \quad + a_{nn}^{(n)} x_n = b_n^{(n)} \end{cases}$$

où $c = \left(b_1^{(1)}, b_2^{(2)}, \cdots, b_n^{(n)}\right)^\top$.

Remarque 5.3.1. Matriciellement, la première étape est équivalente au produit matriciel $A^{(2)} = M_1 A^{(1)}$ où M_1 est la matrice élémentaire de Gauss . L'étape finale est alors donnée par : $A^{(n)} = U = M_{n-1} M_{n-2} \cdots M_2 M_1 A^{(1)}$. Evidemment, l'étape finale n'est accessible par ce procédé que si tous les $a_{kk}^{(k)}$ sont non nuls. Si à une étape donnée $a_{kk}^{(k)}$ est nul, et il y a au moins un $a_{ik}^{(k)}$ non nul, avec $i > k$ on permute les lignes L_i et L_k et on poursuit le procédé. Sinon la matrice A n'est pas inversible et le procédé ne peut continuer.

5.3.2 Méthode de Gauss avec pivot partiel

Exemple 5.3.1. Soit à résoudre le système

$$\begin{cases} 10^{-50} x_1 + x_2 = 0 \\ x_1 - x_2 = 0 \end{cases}$$

La solution exacte est $x_1 = x_2 = 1/(1 + 10^{-50}) \simeq 1$. Cependant, la résolution du système par la méthode de Gauss donne des résultats différents selon qu'on

l'applique avec ou sans pivot.

i) Si on applique la méthode de Gauss sans pivot on obtient

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{1}{10^{-50}} = 10^{50}$$

et

$$(S_2) \begin{cases} 10^{-50}x_1 + x_2 & = 0 \\ (-1 - 10^{50})x_2 & = 10^{-50} \end{cases}$$

qui donne pour solution approchée $x_1 \simeq 1$ et $x_2 \simeq 0$.

ii) Si on adopte la stratégie du pivot partiel qui consiste à mettre en première ligne celle dont le coefficient de x_1 est le plus grand en module alors on permute les lignes pour obtenir le système

$$(S_1) \begin{cases} x_1 - x_2 & = 0 \\ 10^{-50}x_1 + x_2 & = 0 \end{cases}$$

Pour lequel $m_{21} = \dfrac{10^{-50}}{1} = 10^{-50}$ et qui conduit à la solution approchée : $x_2 \simeq 1$ et $x_1 = x_2$.

A travers cet exemple simple, on voit donc le problème que peut poser un pivot trop petit. Pour éviter de diviser par des pivots trop petits pouvant conduire à des solutions absurdes, on peut adopter automatiquement la stratégie du pivot partiel de la manière suivante :

A chaque étape k : choisir $a_{kk}^{(k)}$ tel que : $a_{kk}^{(k)} = \max\limits_{i \geq k} \left| a_{ik}^{(k)} \right|$.

Matriciellement, cette opération revient à multiplier la matrice $A^{(k)}$ par une matrice de permutation I_{kl} avant d'appliquer l'élimination de Gauss. La méthode de Gauss avec pivot partiel s'écrit donc :

$$A^{(2)} = M_1 I_{1i} A^{(1)}, \ \cdots\cdots, \ A^{(n)} = M_{n-1} I_{n-1i} \cdots M_1 I_{1i} A^{(1)} = U$$

où les M_i sont des matrices élémentaires de Gauss et les I_{ki} des matrices de permutation pour $i \geq k$. Si à une étape k on n'a pas besoin de pivoter, l'écriture reste valable avec $I_{ki} = I$ où I désigne la matrice identité.

Théorème 5.3.1. *Soit A une matrice carrée, inversible ou non. Il existe (au moins) une matrice inversible M tellle que la matrice MA soit triangulaire supérieure.*

Preuve :

Si A est inversible, le résultat est déjà prouvé en appliquant la méthode de Gauss sans pivot et en posant $M = M_{n-1} \cdots M_1$. Si A n'est pas inversible cela signifie qu'à une certaine étape k on trouve $a_{ik}^{(k)} = 0$ pour tout $i \geq k$. Mais dans ce cas, il suffit de passer à l'étape suivante.

Matriciellement, cela reviendrait à prendre $I_{ik} = M_k = I$.

5.3.3 Méthode de Gauss avec pivot total

On pourrait aussi adopter la stratégie du pivot total qui consiste, à chaque étape k, à prendre $a_{kk}^{(k)}$ tel que : $a_{kk}^{(k)} = \max_{\substack{i\geq k\\ j\geq k}} \left|a_{ij}^{(k)}\right|$. Ce qui reviendrait à multiplier la matrice $A^{(k)}$ par deux matrices de permutation P et Q, l'une à droite pour permuter les lignes et l'autre à gauche pour permuter les colonnes .

$$A^{(2)} = M_1 I_{1i} A^{(1)} I_{1j}, \cdots, \; A^{(n)} = M_{n-1} I_{n-1i} \cdots M_1 I_{1i} A^{(1)} I_{1j} \cdots I_{n-1j}.$$

5.3.4 Méthode de Gauss-Jordan

C'est une variante qui ressemble à la méthode de Gauss sauf qu'elle aboutit directement à une matrice diagonale. Au lieu des matrices M_k élémentaires on considère les matrices

$$\widetilde{M_k} = \begin{pmatrix} 1 & & & -m_{1k} & & \\ 0 & \ddots & & \vdots & & \\ 0 & & \ddots & -m_{k-1k} & 0 & \\ 0 & & & 1 & & k \\ \vdots & & & -m_{k+1k} & \ddots & \\ \vdots & & & \vdots & & \ddots \\ 0 & & & -m_{k+1k} & & 1 \end{pmatrix}$$

5.4 Factorisation LU

Si on suppose que la méthode de Gauss sans pivot a été appliquée à toutes les étapes du procédé on aboutit à

$$A^{(n)} = M_{n-1} M_{n-2} \cdots M_2 M_1 A^{(1)} = U \qquad \text{avec } \; A^{(1)} = A,$$

ou encore $MA = U$ avec $M = M_{n-1} \cdots M_1$. Comme les matrices élémentaires de Gauss sont triangulaires inférieures et inversibles. En posant $L = M^{-1}$ on obtient $A = LU$.

Remarque 5.4.1. Si on utilise des permutations, alors les matrices $M_k I_{ik}$ ne sont plus triangulaires inférieures. On démontre dans ce cas qu'on aboutit à la forme $PA = LU$.

Théorème 5.4.1 (Condition suffisante de la factorisation LU)**.**
Soit A une matrice carrée d'ordre n telle que toutes les sous-matrices d'ordre k ($k \leq$

n) soient inversibles, alors il existe une matrice triangulaire inférieure L avec $l_{ii} = 1$ et une matrice triangulaire supérieure U telles que $A = LU$. De plus, cette factorisation est unique.

Preuve :

Si $a_{11} \neq 0$, la matrice a_{11} (d'ordre 1) est inversible, donc on peut choisir la matrice de permutation égale à l'identité et appliquer la méthode de Gauss sans pivot à la première étape. Supposons qu'on ait pu choisir toutes les matrices de permutation égales à l'identité jusqu'à l'étape k, il s'ensuit que

$$A^{(k)} = M_{k-1}M_{k-2}\cdots M_1 A = \prod_{i=k-1}^{i=1} M_i A.$$

Avec

$$A_k = \left(\begin{array}{ccccc} a_{11}^{(k)} & & & & \\ 0 & \ddots & & & \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & 0 & \vdots & \ddots & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{array}\right)$$

$$= \left(\begin{array}{ccccc} 1 & & & & \\ ** & \ddots & & & \\ ** & * & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ ** & * & \cdots & \cdots & 1 \end{array}\right) \left(\begin{array}{ccccc} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} & & \\ \vdots & \ddots & & & \\ a_{k1}^{(k)} & \cdots & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1}^{(k)} & \cdots & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{array}\right)$$

en écrivant A sous forme de blocs et en effectuant le produit matriciel par blocs, on obtient $a_{11}^{(1)} \times \cdots \times a_{kk}^{(k)} = \det(B_k)$.

Comme $\det(B_k) \neq 0$ on a $a_{kk}^{(k)} \neq 0$ et par suite on peut choisir $a_{kk}^{(k)}$ comme pivot et poursuivre le procédé.

Unicité

Supposons qu'il existe L_1, L_2, U_1 et U_2 telles que $A = L_1U_1 = L_2U_2$, comme L_2 et U_1 sont inversibles alors $L_2^{-1}L_1 = U_2U_1^{-1}$.

Ce qui impose $L_2^{-1}L_1 = U_2U_1^{-1} = I$ et donc $L_1 = L_2$ et $U_1 = U_2$.

5.5 Factorisation de Choleski (matrice symétrique)

Théorème 5.5.1. *Si A est une matrice symétrique, définie positive, il existe (au moins) une matrice réelle triangulaire inférieure L telle que $A = LL^\top$.*
Si de plus on impose aux éléments diagonaux de L d'être strictement positifs, alors la factorisation est unique.

Preuve :
Remarquons d'abord que si A est définie positive, alors toutes les sous-matrices d'ordre k sont inversibles. Le théorème 5.4.1 permet d'affirmer l'existence de deux matrices L et U telles que $A = LU$. Ce que nous cherchons ici c'est de factoriser en utilisant une seule matrice L. Raisonnons par récurrence sur n.
Si $k = 1$, $A = a_{11} > 0$ donc $a_{11} = \sqrt{a_{11}}.\sqrt{a_{11}}$.
Supposons qu'on ait pu factoriser jusqu'à l'ordre $k-1$ et soit A_k une matrice d'ordre k alors A_k peut s'écrire :

$$A_k = \left(\begin{array}{c|c} A_{k-1} & v \\ \hline v^\top & a_{kk} \end{array}\right) \qquad \text{avec } A_{k-1} = L_{k-1}L_{k-1}^\top.$$

Considérons alors la matrice L_k obtenue à partir de L_{k-1} et telle que :

$$L_k = \left(\begin{array}{c|c} L_{k-1} & l \\ \hline l^\top & l_{kk} \end{array}\right)$$

Le produit matriciel $L_kL_k^\top$ donne :

$$L_kL_k^\top = \left(\begin{array}{c|c} L_{k-1}L_{k-1}^\top & L_{k-1}l \\ \hline l^\top L_{k-1}^\top & l^\top l + l_{kk}^2 \end{array}\right)$$

Par identification on obtient :

$$L_{k-1}l = v \tag{5.5.1}$$
$$L_{k-1}L_{k-1}^\top = A_{k-1} \tag{5.5.2}$$
$$l^\top l + l_{kk}^2 = a_{kk} \tag{5.5.3}$$

i) L'équation (5.5.1) permet alors de résoudre un système et d'obtenir la solution qui est le vecteur l.
ii) L'équation (5.5.3) permet d'obtenir la dernière inconnue du problème, à savoir
$l_{kk} = \sqrt{a_{kk} - l^\top l}$ et on peut choisir $l_{kk} > 0$.

Exemple 5.5.1. Soit A la matrice de Hilbert d'ordre 6, la factorisation de Choleski est donnée par $A = LL^\top$ où

$$L = \begin{pmatrix} 1 & 0.5 & 0.33 & 0.25 & 0.2 & 0.16 \\ 0 & 0.28 & 0.28 & 0.25 & 0.23 & 0.2 \\ 0 & 0 & 0.07 & 0.11 & 0.12 & 0.13 \\ 0 & 0 & 0 & 0.01 & 0.03 & 0.05 \\ 0 & 0 & 0 & 0 & 0.004 & 0.01 \\ 0 & 0 & 0 & 0 & 0 & 0.0012 \end{pmatrix}.$$

5.6 Factorisation de Householder (matrice unitaire)

Soit $P_0 = I - 2\omega_0\omega_0^\top$ une matrice élémentaire de Householder avec

$$\omega_0^\top \omega_0 = 1. \tag{5.6.1}$$

On cherche une matrice unitaire P_0 telle que

$$P_0 a = ke_1, \tag{5.6.2}$$

pour tout vecteur $a = (a_1, \cdots, a_n)^\top$, avec $k \in \mathbb{R}$ et $e_1 = (1, 0, \cdots, 0)^\top$.
P_0 est orthogonale c'est à dire $P_0^\top P_0 = I$ et par suite, on doit avoir

$$\left(a^\top P_0^\top\right)(P_0 a) = k^2 = a^\top a.$$

Soit $k = \pm \left(a^\top a\right)^{1/2}$, les équations (5.6.1) et (5.6.2) donnent :
$P_0 a = a - 2\omega_0\omega_0^\top a = ke_1$ et parsuite $2\omega_0\omega_0^\top a = -ke_1 + a = v$, si on pose $\alpha = 2\omega_0^\top a$.
On obtient $\alpha\omega_0 = v$, et comme on cherche ω_0 tel que $\omega_0^\top \omega_0 = 1$, il vient : $\alpha^2 = v^\top v$.
Par suite $P_0 = I - \dfrac{2}{\alpha^2} vv^\top = I - 2\dfrac{vv^\top}{v^\top v}$.

Remarques 5.6.1.

i) Le choix de k se fait au signe près , on peut choisir le signe +.

ii) Le même procédé peut être appliqué pour obtenir une matrice
$P_k = I_k - 2\omega_k\omega_k^\top$ avec $\omega_k = (0, \cdots, 0, \omega_{k+1k}, \cdots, \omega_{nk})^\top$.

On a constaté que P_k peut être décomposée sous la forme

$$P_k = \left(\begin{array}{c|c} I_k & \\ \hline & \widehat{P}_k \end{array}\right) \quad \text{avec } \widehat{P}_k = I_{n-k} - 2\widehat{\omega}_k\widehat{\omega}_k^\top \text{ (voir paragraphe 5.2.3).}$$

iii) La factorisation de Householder permet d'écrire :

$$P_{n-2}P_{n-3}\cdots P_1 P_0 A = U,$$

ou encore $A = QU$ avec $Q = P_0 P_1 \cdots P_{n-2}$ une matrice orthogonale.

5.7 Conditionnement

Exemple 5.7.1. [dû à R.S. Wilson]
Soit à résoudre le système linéaire $Ax = b$ avec

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

La solution exacte est donnée par $x = (32, 23, 33, 31)^\top$, si on perturbe le second membre d'un δb, quel est l'effet de cette perturbation sur la solution ?
Soit $b + \delta b = (32.1, 22.9, 33.1, 30.9)^\top$, alors il en résulte une solution
$\widetilde{x} = x + \delta x = (9.2, -12.6, 4.5, -1.1)$.
Soit une erreur de l'ordre de $\frac{1}{200}$ sur les données et un rapport d'amplification de l'erreur relative de l'ordre 2000 ce qui montre que le système n'est pas stable, il reste vulnérable à toute perturbation, on caractérise la stabilité intrinsèque d'un tel système indépendamment de la méthode de résolution en disant qu'il est mal conditionné. De même si on perturbe les éléments a_{ij} de la matrice A de δA, on aboutit à une solution approchée $\widetilde{x}$ qui est très différente de la solution exacte.

Définition 5.7.1. Soit $\|.\|$ une norme subordonnée et A une matrice inversible . On appelle conditionnement de A relativement à la norme $\|.\|$, le nombre $\|A\|\|A^{-1}\|$ noté $C(A)$ ou $cond(A)$.

Propriétés du conditionnement

On vérifie facilement les propriétés suivantes
i) $C(A) \geq 1$.
ii) $C(A) = C\left(A^{-1}\right)$.
iii) $C(\alpha A) = |\alpha| C(A), \quad \alpha \neq 0$.
iv) $C_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2 = \dfrac{\mu_n(A)}{\mu_1(A)}$, μ_i valeur singulière de A.
v) $C(QA) = C(A)$ pour toute matrice orthogonale Q.

Théorème 5.7.1. *Soit A une matrice inversible, x et $x + \delta x$ les solutions respectives de $Ax = b$ et $A\widetilde{x} = b + \delta b$ où $\widetilde{x} = x + \delta x$, alors on a*

$$\frac{1}{C(A)} \left(\frac{\|\delta b\|}{\|b\|} \right) \leq \frac{\|\delta x\|}{\|x\|} \leq C(A) \left(\frac{\|\delta b\|}{\|b\|} \right). \tag{5.7.1}$$

De même la solution obtenue après perturbation de A par δA vérifie :

$$\frac{\|\delta x\|}{\|x\|} \leq \left(\frac{C(A)}{1 - C(A)\dfrac{\|\delta A\|}{\|A\|}} \frac{\|\delta A\|}{\|A\|} \right).$$

Si on perturbe en même temps A et b alors on obtient :

$$\frac{\|\delta x\|}{\|x\|} \leq \left(\frac{C(A)}{1 - C(A)\dfrac{\|\delta A\|}{\|A\|}} \right) \left(\frac{\|\delta A\|}{\|A\|} \frac{\|\delta b\|}{\|b\|} \right).$$

Preuve :
On a $Ax = b$ et $Ax + A\delta x = b + \delta b$ d'où $A\delta x = \delta b$ ou encore $\delta x = A^{-1}\delta b$, comme $\|b\| \leq \|A\|\|x\|$, on déduit que $\dfrac{\|\delta x\|}{\|x\|} \leq C(A)\dfrac{\|\delta b\|}{\|b\|}$.
Les autres inégalités s'obtiennent de façons similaires.

Remarques 5.7.1.
i) L'équation (5.7.1) donne une estimation de l'erreur relative de la solution en fonction de l'erreur relative connue $\dfrac{\|\delta b\|}{\|b\|}$.
ii) Tous les calculs sont effectués sur un ordinateur, des erreurs d'arrondi sont accumulées à chaque étape de calcul. Si ε désigne la précision numérique relative (dépend de la machine), l'erreur relative de la solution explose si $C(A) \times \varepsilon \simeq 1$.
iii) La matrice de Hilbert d'ordre n, $H = \left(\dfrac{1}{i+j-1} \right)_{i,j=1}^{n}$ présente un exemple classique de matrices mal conditionnées.

5.8 Matrices creuses

On appelle matrice creuse une matrice dont la plupart des éléments sont égaux à zéro. Si de plus la structure des éléments non nuls est simple, il n'est pas nécessaire de réserver une quantité de mémoire égale à celle de la matrice complète. Des algorithmes spécifiques permettent de réduire le temps de calcul de manière considérable. Parmi les cas simples des matrices creuses, citons les matrices :

- tridiagonales : les éléments non nuls sont sur la diagonale et de part et d'autre de celle-ci sur les deux lignes adjacentes. On a $a_{ij} = 0$ pour $|i - j| > 1$,
- diagonale par bande de largeur m : les éléments de matrices tels que $|i - j| > m$ sont nuls,

- simplement ou doublement bordées : par rapport à la définition précédente, des éléments non nuls supplémentaires existent le long des lignes ou colonnes du bord de la matrice.

Matrices tridiagonales

Soit le système

$$\begin{pmatrix} d_1 & e_1 & & & & \\ c_2 & d_2 & e_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & d_{n-1} & e_{n-1} \\ & & & & c_n & d_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{pmatrix} \tag{5.8.1}$$

Il y'a un vaste choix de bibliothèques disponibles pour calculer les solutions qui prennent un temps de calcul proportionnel à n.

De manière générale, on peut obtenir un algorithme proportionnel à n pour une matrice à bande. Le préfacteur de l'algorithme est proportionnel à m.

Formule de Sherman-Morison

Supposons que l'on ait une matrice A dont on a facilement calculé l'inverse (cas d'une matrice triangulaire). Si on fait un petit changement dans A en modifiant par exemple un ou quelques éléments de l'ensemble de la matrice, peut on calculer facilement l'inverse de cette nouvelle matrice ?
En effet soit B une matrice telle que

$$B = A + uv^\top \tag{5.8.2}$$

La formule de Sherman-Morison donne l'inverse de B.

$$\begin{aligned} B^{-1} &= \left(I + A^{-1}uv^\top\right)^{-1} A^{-1} \\ &= \left(I - A^{-1}uv^\top + A^{-1}uv^\top A^{-1}uv^\top + \cdots\right) A^{-1} \\ &= A^{-1} - A^{-1}uv^\top A^{-1}\left(1 - \lambda + \lambda^2 + \cdots\right) \\ &= A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+\lambda} \end{aligned}$$

où $\lambda = v^\top A^{-1}u$. Posons $z = A^{-1}u$ et $w = \left(A^{-1}\right)^\top v$ on a $\lambda = v^\top z$ et
$B^{-1} = A^{-1} - \dfrac{zw^\top}{1+\lambda}$.

Remarque 5.8.1. Dans le cas où les vecteurs u et v sont des matrices U et V d'ordre respectif $n \times k$ et $k \times n$ et si de plus la matrice $I + VA^{-1}U$ est inversible, alors

$$(A + UV)^{-1} = A^{-1} - A^{-1}(I + VA^{-1}U)^{-1}VA^{-1}.$$

Supposons que les éléments de la matrice A vérifient

$$\begin{array}{l} i)\ \ |d_1| > |e_1| \\ ii)\ \ |d_i| > |e_i| + |c_i| \quad i = 1, \cdots, n-1 \\ iii)\ \ |d_n| > |e_n| \end{array} \tag{5.8.3}$$

et que

$$c_i e_{i-1} \neq 0, \quad i = 2, \cdots, n \tag{5.8.4}$$

Sous ces conditions la matrice A est irréductible à diagonale dominante donc A est inversible. Wendroff a montré qu'il existe une factorisation de la matrice A sous la forme suivante

$$A = \widehat{L}\widehat{R} \tag{5.8.5}$$

où $\widehat{R}$ est une matrice triangulaire supérieure avec éléments diagonaux égaux à 1 et $\widehat{L}$ est une matrice triangulaire inférieure. Si A est tridiagonale, les matrices $\widehat{R}$ et $\widehat{L}$ sont données par $\widehat{L} = \begin{pmatrix} \alpha_1 & & & \\ c_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & c_n & \alpha_n \end{pmatrix}$ et $\widehat{R} = \begin{pmatrix} 1 & \gamma_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \gamma_n \\ & & & 1 \end{pmatrix}$,

où

$$\begin{array}{l} i)\ \ \alpha_1 = d_1,\ \ \gamma_1 = \dfrac{e_1}{d_1} \\ ii)\ \ \alpha_i = d_i - c_i\gamma_{i-1}, \quad i = 2, \cdots, n \\ iii)\ \ \gamma_i = \dfrac{e_i}{\alpha_i}, \quad i = 2, \cdots, n-1 \end{array} \tag{5.8.6}$$

Puisque A est inversible et $\det(A) = \prod_{i=1}^{n} \alpha_i$, alors tous les α_i sont non nuls et le schéma récursif (5.8.6) est bien défini.

Remarque 5.8.2. On note que la factorisation (5.8.5) est obtenue par application de la méthode d'élimination de Gauss usuelle à $A^\top$.

La résolution du système lnéaire est équivalente à

$$\widehat{L}v = b,\ \ \widehat{R}x = v,$$

qui est obtenue de la manière suivante

$v_1 = b_1/\alpha_1$

Pour $i = 2$ *à* n

$\quad v_i = (b_i - c_i v_{i-1})/\alpha_i$

fin

$x_n = v_n$

Pour $i = 1$ *à n-1*

$\quad x_{n-i} = v_{n-i} - \gamma_{n-i} v_{n-i+1}$

fin

Le théorème suivant donne une majoration des quantités γ_i et α_i indépendamment de l'ordre de la matrice A.

Théorème 5.8.1. *Si les éléments de A vérifient les conditions (5.8.3) alors*

$$\begin{array}{l} i)\ \ |\gamma_i| < 1 \\ ii)\ \ 0 < |d_i| - |c_i| < |d_i| + |c_i|, \quad i = 2, \cdots, n \end{array} \tag{5.8.7}$$

Preuve :
On a $\gamma_i = \dfrac{e_i}{\alpha_i} = \dfrac{e_i}{d_i}$, la condition i) de (5.8.3) donne $|\gamma_1| < \dfrac{|e_1|}{|d_1|} < 1$.
Supposons que $\gamma_j < 1$ pour $j = 1, \cdots, i-1$, en remplaçant (5.8.6 (iii)) dans (5.8.6 (ii)) on obtient

$$\gamma_i = \frac{e_i}{d_i - c_i \gamma_{i-1}}$$

et

$$\begin{aligned} |\gamma_i| &< \frac{|e_i|}{|d_i| - |c_i||\gamma_{i-1}|} \\ &< \frac{|e_i|}{|d_i| - |c_i|} \end{aligned}$$

par hypothèse. Ensuite en considérant (5.8.3 (ii)) il s'ensuit que $|\gamma_i| < 1$.
Pour montrer (5.8.7 (ii)) il suffit de considérer (5.8.6 (ii)) et (5.8.3 (ii)).

L'algorithme présenté précédemment peut être simplifié en éliminant les α_i, c'est le cas où la factorisation $\widehat{L}\widehat{R}$ n'est pas nécessaire. L'algorithme se présente comme suit

$v_1 = b_1/d_1$

$t = d_1$

Pour $i = 2$ *à* n

$\quad \gamma_{i-1} = e_{i-1}/t$

$\quad t = d_i - c_i \gamma_{i-1}$

$\quad v_i = (b_i - c_i v_{i-1})/t$

fin

$x_n = v_n$

Pour $i = 1$ *à n-1*

$x_{n-i} = v_{n-i} - \gamma_{n-i} x_{n-i+1}$

fin

Si A est définie positive, elle possède une factorisation de la forme $L^\top DL$ qui est obtenue facilement du schéma (5.8.6). Si

$$L = \begin{pmatrix} 1 & & & \\ l_1 & \ddots & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{pmatrix} \tag{5.8.8}$$

et $D = diag\,(\delta_1, \cdots, \delta_n)$, par identification des écritures et en posant $c_i = e_{i-1}$ on obtient

$\delta_1 = d_1$

Pour $i = 1$ *à n-1*

$l_i = e_i/\delta_i$

$\delta_{i+1} = d_{i+1} - l_i e_i$

fin

A est définie positive donc $u^\top L^\top DLu > 0$, posons $v = L^\top u$ ($v \neq 0$ car L est inversible), alors $v^\top Dv > 0$, ainsi D est une matrice diagonale définie positive, dont les éléments diagonaux sont tous non nuls, et par suite le schéma récursif est bien défini. Si on pose $Lv = b$, l'algorithme de résolution s'écrit

$v_1 = b_1$

Pour $i = 1$ *à n-1*

$v_{i+1} = b_{i+1} - l_i v_i$

fin

$x_n = v_n/\delta_n$

Pour $i = 1$ *à n-1*

$x_{n-i} = (v_{n-i} - e_{n-i} x_{n-i+1})/\delta_{n-i}$

fin

Si la factorisation n'est pas nécessaire on obtient l'algorithme simplifié suivant

$v_1 = b_1$

$\delta_1 = d_1$

Pour $i = 1$ *à n-1*

$t = e_i/\delta_i$

$\delta_{i+1} = d_{i+1} - te_i$

$v_{i+1} = b_{i+1} - tv_i$

fin
$x_n = v_n/\delta_n$
Pour $i = 1$ *à n-1*
$\quad x_{n-i} = (v_{n-i} - e_{n-i}x_{n-i+1})/\delta_{n-i}$
fin

Exemple 5.8.1. Soit le système linéaire $Ax = b$, où A est une matrice carrée d'ordre n de la forme

$$A = \begin{pmatrix} a_1 & \cdots & \cdots & \cdots & \cdots & a_n \\ 1 & \lambda & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & \lambda & 1 \\ b_1 & \cdots & \cdots & \cdots & \cdots & b_n \end{pmatrix}.$$

Ce type de matrices est issu de la discrértisation de certains problèmes paraboliques voir [**?**, **?**].
Considérons la matrice

$$A_0 = \begin{pmatrix} \alpha & 1 & & & \\ 1 & \alpha & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & \alpha & 1 \\ & & & 1 & \alpha \end{pmatrix}.$$

avec $\alpha = \dfrac{\lambda - \sqrt{\lambda^2 - 4}}{2}$ et $\beta = \dfrac{\lambda + \sqrt{\lambda^2 - 4}}{2}$. Il s'ensuit que

$$A_0 = LU$$

avec

$$L = \begin{pmatrix} \alpha & & & & \\ 1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 1 & \alpha \end{pmatrix}, \quad U = \begin{pmatrix} 1 & \beta & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta \\ & & & & 1 \end{pmatrix}.$$

La solution y de $A_0 y = \omega$, est donnée par

$$\begin{aligned} y_n &= v_n \\ y_m &= v_m - \beta y_{m+1}, \quad m = 0, \cdots, n-1, \end{aligned}$$

où v est donnée par

$$\begin{aligned} v_n &= \beta\omega_0 \\ v_m &= \beta(\omega_m - \omega_{m-1}),\ \ m = 1, \cdots, n. \end{aligned}$$

Soit $z = x - y$, on obtient

$$Az = \left(\omega_0 - \sum_{k=0}^{n} a_k y_k, \cdots, \omega_n - \sum_{k=0}^{n} b_k y_k \right)^\top.$$

Comme

$$z_{m-1} + \lambda z_m + z_{m+1} = 0,\ \ m = 1, \cdots, n-1.$$

Il vient

$$z_m = c_0 \gamma^{n-m} + c_1 \gamma^m,\ \ m = 0, \cdots, n,$$

où $\gamma = \dfrac{-\lambda - \sqrt{\lambda^2 - 4}}{2}$ et c_0, c_1 sont des constantes à déterminer en résolvant le système

$$\begin{aligned} c_0 \sum_{k=0}^{n} a_k \gamma^{n-k} + c_1 \sum_{k=0}^{n} a_k \gamma^k &= \omega_0 - \sum_{k=0}^{n} a_k y_k \\ c_0 \sum_{k=0}^{n} b_k \gamma^{n-k} + c_1 \sum_{k=0}^{n} b_k \gamma^k &= \omega_n - \sum_{k=0}^{n} b_k y_k \end{aligned}$$

Finalement on obtient la solution du système initial en posant $x = y + z$.

5.9 Résultats sur les matrices non carrées

Théorème 5.9.1. *soit A une matrice rectangulaire (m, n) alors il existe deux matrices carrées unitaires U et V d'ordre respectivement m et n tel que : $U^* A V = \Sigma$, où Σ est une matrice rectangulaire (m, n) donnée par :*

$$\Sigma = \begin{array}{c} \\ \\ \\ \\ n \end{array} \overset{\overset{n}{\downarrow}}{\begin{pmatrix} \mu_1 & & & & \\ & \ddots & & & \\ & & \mu_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \\ & & \boldsymbol{0} & & \end{pmatrix}} \qquad m > n$$

où les μ_i sont les valeurs singulières de A.

Corollaire 5.9.1. – *Le rang de A est égal au nombre de valeurs singulières non nulles.*

– *La forme $A = U\Sigma V^*$ est appelée décomposition en valeurs singulières de A et on a $A = \sum_{i=1}^{r} \mu_i u_i v_i^*$ et $A^*A = \sum_{i=1}^{r} \mu_i^2 u_i v_i^*$ où u_i et v_i désignent, respectivement, les i^{eme} colonnes de U et de V.*

Preuve :

$$\text{Soient } I_n = \begin{array}{c} \begin{matrix} e_1 & \cdots & e_n \end{matrix} \\ \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & 1 \end{pmatrix} \end{array} \quad \text{et} \quad I_m = \begin{array}{c} \begin{matrix} e_1 & \cdots & e_m \end{matrix} \\ \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & 1 \end{pmatrix} \end{array}$$

on a par définition

$$V = \sum_{i=1}^{n} v_i e_i^* \,, \quad V^* = \sum_{i=1}^{n} e_i v_i^*, \quad \Sigma e_i = \mu_i \varepsilon_i \text{ pour } i = 1 \cdots n,$$

$$\Sigma V^* = \sum_{i=1}^{n} \mu_i \varepsilon_i v_i^*, \quad U\Sigma V^* = \sum_{i=1}^{r} \mu_i u_i v_i^* \quad \text{et } A^*A = \sum_{i=1}^{r} \mu_i^2 u_i v_i^*. \qquad .$$

5.10 Résolution des systèmes à matrices non carrées

Exemple 5.10.1. $\begin{pmatrix} 1 & 2 \\ -1 & 1 \\ -1 & -3 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 2 \\ 2 & 3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$

Cas où $m \geq n$ et rang $A = n$

Considérons le système $Ax = b$, on définit le i^{eme} résidu

$$r_i = \sum_{j=1}^{n} a_{ij} x_j - b_i \quad i = 1, \cdots, m.$$

(Vectoriellement $r = Ax - b$).
La méthode des moindres carrées consiste à minimiser $\|r\|_2^2$.
Posons

$$I(x) = r^\top r = (Ax - b)^\top (Ax - b).$$

Donc minimiser $I(x)$ revient à chercher des solutions de l'équation

$$\frac{\partial I(x)}{\partial x_i} = 0.$$

Posons $e_n = (0, \cdots, 1, \cdots, 0)^\top$, on a d'une part

$$\begin{aligned} r^\top r &= x^\top A^\top Ax - b^\top Ax - x^\top A^\top b + b^\top b \\ &= x^\top A^\top Ax - 2x^\top A^\top b + b^\top b. \end{aligned}$$

D'autre part

$$\begin{aligned} \frac{\partial I(x)}{\partial x_i} &= e_i^\top A^\top Ax + xA^\top Ae_i - 2e_i^\top A^\top b \\ &= 2e_i^\top (A^\top Ax - A^\top b). \end{aligned}$$

Donc trouver un minimum de $I(x)$ revient à résoudre le système linéaire $A^\top Ax = A^\top b$.

Définition 5.10.1. Soit Σ la matrice rectangulaire (m, n) donnée par

$$\Sigma = \begin{array}{c} \begin{array}{c} \\ n \end{array} \overset{\overset{n}{\downarrow}}{\begin{pmatrix} \mu_1 & & & & & \\ & \ddots & & & & \\ & & \mu_2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ & & & \mathbf{0} & & \end{pmatrix}} \end{array} \qquad m > n$$

on appelle **pseudo-inverse** de Σ la matrice Σ^+ de la forme (n, m) définie par :

$$\Sigma^+ = \begin{array}{c} \begin{array}{c} \\ n \end{array} \overset{\overset{n}{\downarrow}}{\begin{pmatrix} \mu_1^{-1} & & & & & \\ & \ddots & & & & \\ & & \mu_r^{-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ & & & \mathbf{0} & & \end{pmatrix}} \end{array} \qquad m > n$$

Définition 5.10.2. Soit A une matrice de forme (m, n) dont la décomposition en valeurs singulières est $A = U\Sigma V^*$. On appelle **pseudo-inverse**(ou inverse généralisé) de la matrice A, la matrice A^+ de la forme (n, m) donnée par $A^+ = V\Sigma^+ U^*$.

Remarque 5.10.1.

1. $A^+A = V\Sigma^+ U^* U\Sigma V^* = V\Sigma^+ \Sigma V^*$.
2. Sous *Matlab* la commande $svd(A)$ donne la décomposition en valeurs singulières de A.

Exemple 5.10.2. Soit la matrice A donnée

$$A = \begin{pmatrix} 0.1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.5 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.6 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.7 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

La décomposition en valeurs singulières de A est donnée par

$$A = UwV^{\top},$$

où

$$U = \begin{pmatrix} 0.34 & 0.54 & -0.32 & -0.003 & -0.30 & -0.48 & 0.09 & -0.37 \\ 0.34 & 0.39 & 0.55 & -0.23 & -0.3 & 0.12 & 0.02 & 0.5 \\ 0.35 & 0.23 & -0.12 & 0.33 & 0.66 & 0.09 & 0.44 & 0.22 \\ 0.35 & 0.08 & -0.1 & 0.08 & -0.13 & 0.78 & -0.16 & -0.43 \\ 0.35 & -0.07 & -0.23 & 0.19 & 0.14 & -0.16 & -0.78 & 0.32 \\ 0.35 & -0.22 & 0.02 & -0.78 & 0.38 & -0.09 & -0.01 & -0.2 \\ 0.35 & -0.38 & 0.59 & 0.41 & -0.01 & -0.27 & 0.03 & -0.35 \\ 0.36 & -0.53 & -0.39 & -0.001 & -0.43 & 0.02 & 0.37 & 0.31 \end{pmatrix}$$

$$w = \begin{pmatrix} 8.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.64 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$V = \begin{pmatrix} 0.15 & -0.98 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.34 & 0.05 & 0.93 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.34 & 0.05 & -0.13 & 0.92 & 0 & 0 & 0 & 0 & 0 \\ 0.34 & 0.05 & -0.13 & -0.15 & 0.91 & 0 & 0 & 0 & 0 \\ 0.34 & 0.05 & -0.13 & -0.15 & -0.18 & -0.89 & 0 & 0 & 0 \\ 0.34 & 0.05 & -0.13 & -0.15 & -0.18 & 0.22 & -0.5 & -0.5 & -0.5 \\ 0.34 & 0.05 & -0.13 & -0.15 & -0.18 & 0.22 & 0.83 & -0.16 & -0.16 \\ 0.34 & 0.05 & -0.13 & -0.15 & -0.18 & 0.22 & -0.16 & 0.83 & -0.16 \\ 0.34 & 0.05 & -0.13 & -0.15 & -0.18 & 0.22 & -0.16 & -0.16 & 0.83 \end{pmatrix}$$

Théorème 5.10.1. *Si A est une matrice rectangulaire (m, n) alors on a*

1. $A^+ = \sum_{i=1}^{n} \mu_i^{-1} v_i u_i^*$.
2. $AA^+ = \sum_{i=1}^{n} u_i u_i^*$ *matrice de projection orthogonale sur ImA.*
3. $A^+A = \sum_{i=1}^{n} v_i v_i^*$ *matrice de projection orthogonale sur ImA^*.*

Remarques 5.10.2.

1. Le nombre $\mathcal{K}(A) = \|A\|_2 \|A^+\|_2$ est appelé conditionnement généralisé de A.
2. Si $rang(A) = n$ alors $\mathcal{K}^2(A) = \mathcal{K}(AA^\top) = Cond(AA^\top)$.
3. La commande MATLAB \ (backslash) est la commande générique pour résoudre un système linéaire. L'algorithme mis en œuvre dépend de la structure de la matrice A du système. MATLAB utilise dans l'ordre les méthodes suivantes :
   i) Si A est une matrice triangulaire , le système est résolu par simple substitution.
   ii) Si A est une matrice symétrique ou hermitienne, définie positive , la résolution est effectuée par la méthode de Choleski.
   iii) Si A est une matrice carrée, mais n'entrant pas dans les deux cas précédents, une factorisation LU est réalisée en utilisant la méthode d'élimination de Gauss avec stratégie de pivot partiel.
   iv) Si A n'est pas une matrice carrée, la méthode QR est utilisée.
4. La matrice A peut être creuse, elle comporte une forte proportion de coefficients nuls (de nombreux problèmes issus de la physique conduisent à l'analyse de systèmes linéaires à matrices creuses), l'intérêt de telles matrices résulte non seulement de la réduction de la place mémoire (on ne stocke pas les zéros) mais aussi de la réduction des nombres d'opérations. Dans le cas de ces matrices des algorithmes particuliers sont mis en œuvre.

5. Chacune des méthodes précédentes peut être utilisée de manière spécifique grâce aux commandes *chol, lu, qr.*

5.11 Conclusion

Avant d'entamer la résolution des systèmes linéaires de grandes tailles, il est impératif de commencer par une analyse des propriétés de la matrice afin de détermi-ner la méthode la plus adaptée afin d'obtenir une solution avec une précision correcte et pour un temps de calcul qui sera minimal. Les différentes méthodes dans ce chapitre ne sont qu'une introduction à ce très vaste sujet.

5.12 Exercices

Exercice 5.12.1. Une matrice $A = (a_{ij})$ carrée d'ordre n à diagonale strictement dominante en colonnes.

1. Donner l'expression du terme général de la matrice $A^{(2)}$ obtenue après la première étape du procédé d'élimination de Gauss .
2. Montrer que la matrice B d'ordre $(n-1)$ obtenue à partir de $A^{(2)}$ en enlevant la première ligne et la première colonne est à diagonale strictement dominante en colonnes.
3. Quel est l'intérêt de cette propriété ?

Exercice 5.12.2. Soit a et b deux vecteurs ayant le même nombre de composantes et α un scalaire.

1. Montrer que la matrice $I - \alpha ab^\top$ admet une matrice inverse de la même forme en imposant une condition sur α. En déduire l'expression de $(A + ab^\top)^{-1}$ en fonction de A^{-1} et des vecteurs a et b.
2. Appliquer ce résultat aux matrices de Gauss : $I + m_k e_k^\top$
3. Montrer que :
$$(I + m_1 e_1^\top).(I + m_2 e_2^\top) \cdots (I + m_{n-1} e_{n-1}^\top) = I + m_1 e_1^\top + \cdots + m_{n-1} e_{n-1}^\top$$

Exercice 5.12.3. Soit A une matrice rectangulaire (m, n) avec $m > n$.

1. Montrer que il existe deux matrices carrées unitaires U et V d'ordre respectivement m et n telles que : $U^* A V = \widetilde{D}$ où $\widetilde{D}$ est une matrice rectangulaire (m, n) donné par :
$$\widetilde{D} = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_r \\ 0 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \quad \text{avec } \mu_i \text{ valeurs singulières de } A$$
2. Déduire le rang de A

Exercice 5.12.4. Soit $A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 5 & 3 & 4 \\ 0 & 3 & -20 & 15 \\ 0 & 4 & 15 & -5 \end{pmatrix}$

1. Utiliser les transformations de Householder pour obtenir une matrice tridiagonale .
2. Utiliser les transformations de Givens pour obtenir une matrice tridiagonale.

Exercice 5.12.5. Soit $B = (b_{ij})$ une matrice tridiagonale donnée par : $b_{ij} = 1$ et $b_{i-1i} = b_{i+1i} = \frac{1}{4}$ et E une matrice carrée vérifiant :$\|E\|_1 = \varepsilon < \frac{1}{2}$.
Montrer que les systèmes $Bx = b$ et $(B + E)y = b$ ont des solutions uniques x et y et prouver que : $\|y - x\|_1 \leq \frac{4\varepsilon}{1-2\varepsilon}\|b\|_1$.

Chapitre 6

Méthodes indirectes de résolution des systèmes linéaires $Ax = b$

6.1 Introduction

Pour résoudre le système

$$Ax = b, \tag{6.1.1}$$

on utilise des méthodes, dites indirectes, du type

$$x^{(k+1)} = Tx^{(k)} + C \tag{6.1.2}$$

où T est une matrice obtenue à partir de A et C un vecteur dépendant de A et de b. Pour passer de l'équation (6.1.1) à l'équation (6.1.2) on écrit A sous la forme : $A = M - N$ avec M inversible. En remplaçant dans (6.1.2) on obtient $Mx = Nx + b$ et par la suite

$$x = M^{-1}Nx + M^{-1}b, \tag{6.1.3}$$

qui suggère le procédé itératif suivant

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b \tag{6.1.4}$$

qui est de la forme (6.1.2) avec $T = M^{-1}N$ et $C = M^{-1}b$.

Remarque 6.1.1. En général il y a une multitude de façons d'écrire A sous la forme $A = M - N$. Dans le cadre de ce chapitre, nous nous limiterons à deux familles de décomposition. La première comprend les méthodes classiques $A = M - N = D - L - U$, en précisant :

- Méthode de Jacobi : $M = D, N = L + U$.
- Méthode de Gauss-Seidel : $M = D - L, N = U$.

– Méthode de relaxation : $A = A(\omega) = M(\omega) - N(\omega)$, avec $M(\omega) = \frac{1}{\omega}D - L$, $N(\omega) = \frac{1-\omega}{\omega}D - U$ où ω est un scalaire.

La deuxième sera consacrée aux matrices positives.

6.2 Généralités et définitions

Définition 6.2.1. Une méthode de type (6.1.2) est dite convergente si pour toute valeur initiale $x^{(0)}$ on a $\lim_{n\longrightarrow+\infty} x^{(n)} = x$. Si une telle limite x existe alors elle vérifie $Tx + C = x$.

Définition 6.2.2. On appelle erreur de la méthode (à la k^{eme} itération) la quantité $e^{(k)} = x^{(k)} - x$. avec $e^{(0)} = x^{(0)} - x$ on obtient $e^{(k)} = T^k e^{(0)}$ pour tout k.
La méthode est convergente si $\lim_{k\longrightarrow+\infty} T^k = 0$.

Définition 6.2.3. Une matrice carrée B est dite convergente si $\lim_{k\longrightarrow+\infty} B^k = 0$.

Théorème 6.2.1. *Soit A une matrice carrée. Les assertions suivantes sont équivalentes :*

i) $\lim_{k\longrightarrow+\infty} A^k = 0$.

ii) $\lim_{k\longrightarrow+\infty} A^k v = 0$ *pour tout v.*

iii) $\rho(A) < 1$.

iv) $\|A\| < 1$, *pour au moins une norme matricielle induite.*

Preuve : On utilisera les résultats de l'exercice 1.5.3

i)$\Longrightarrow$ii)

On a $\|A^k v\| \leq \|A^k\|\|v\|$; l'assertion i) et la continuité de la norme implique que $\lim_{k\longrightarrow+\infty} \|A^k\| = 0$ et par suite $\lim_{k\longrightarrow+\infty} \|A^k v\| = 0$ et $\lim_{k\longrightarrow+\infty} A^k v = 0$ pour tout v.

ii)$\Longrightarrow$ iii)

Ceci revient à montrer que : Non iii)$\Longrightarrow$ Non ii)

Supposons que $\rho(A) \geq 1$ et soit v le vecteur propre associé à la valeur propre λ qui vérifie $|\lambda| = \rho(A)$, donc $Av = \lambda v$ et on en tire $A^k v = \lambda^k v$ d'où, si $|\lambda| \geq 1$ finalement $A^k v$ ne converge pas vers 0 pour ce v et donc Non ii) est vraie.

iii)$\Longrightarrow$ iv)

D 'après l'exercice 1.5.3, pour tout $\varepsilon > 0$ il existe une norme induite telle que $\|A\| \leq \rho(A) + \varepsilon$ il suffit de considérer un ε tel que $\rho(A) + \varepsilon < 1$, pour ce ε, l'exercice 1.5.3 assure que $\|A\| \leq \rho(A) + \varepsilon$ et par suite $\|A\| < 1$.

iv)$\Longrightarrow$ i)

Elle est évidente car si $\|A\| < 1$ alors $\lim_{k\longrightarrow+\infty} \|A\|^k = 0$ et par ailleurs $\|A^k\| \leq \|A\|^k$ d'où $\lim_{k\longrightarrow+\infty} \|A^k\| = 0$ et $\lim_{k\longrightarrow+\infty} A^k = 0$

Théorème 6.2.2. *Soit A une matrice carrée et $\|.\|$ une norme quelconque, alors* $\lim_{k\longrightarrow+\infty} \left\|A^k\right\|^{1/k} = \rho(A)$.

Preuve :

i) On a $\rho(A) \leq \|A\|$ et comme $\rho\left(A^k\right) = (\rho(A))^k$

il s'ensuit que $\rho\left(A^k\right) = (\rho(A))^k \leq \left\|A^k\right\|$ et par suite $\rho(A) \leq \left\|A^k\right\|^{1/k}$ donc $\rho(A) \leq \lim_{k\longrightarrow+\infty} \left\|A^k\right\|^{1/k}$.

ii) Pour montrer que $\lim_{k\longrightarrow+\infty} \left\|A^k\right\|^{1/k} \leq \rho(A)$, introduisons pour tout $\varepsilon > 0$ la matrice $A_\varepsilon = \dfrac{1}{\rho(A)+\varepsilon}A$, pour cette matrice on a $\rho(A_\varepsilon) = \dfrac{\rho(A)}{\rho(A)+\varepsilon} < 1$ et d'après le théorème précédent on obtient : $\|A_\varepsilon\| < 1$ pour au moins une norme matricielle induite, donc $\lim_{k\longrightarrow+\infty} A_\varepsilon^k = 0$ et $\lim_{k\longrightarrow+\infty} \|A_\varepsilon\|^k = 0$.

Donc $\forall\ \varepsilon > 0$, $\exists\ N(\varepsilon)$, tel que pour tout $k \geq N(\varepsilon)$ on ait : $\left\|A_\varepsilon^k\right\| < 1$

ou encore $\left\|A^k\right\| \leq (\rho(A)+\varepsilon)^k$ soit encore $\left\|A^k\right\|^{1/k} \leq (\rho(A)+\varepsilon)$ pour tout $\varepsilon > 0$

finalement $\lim_{k\longrightarrow+\infty} \left\|A^k\right\|^{1/k} \leq \rho(A)$.

Remarque 6.2.1. $\|A\|_2 = \rho(A^*A)^{1/2}$ et si A est hermitienne alors $\|A\|_2 = \rho(A)$.

Théorème 6.2.3. *Considérons deux méthodes itératives $\widetilde{x}^{(k+1)} = \widetilde{T}x^{(k)} + \widetilde{C}$ et $x^{(k+1)} = Tx^{(k)} + C$, avec $\rho(T) < \rho\left(\widetilde{T}\right)$ et $x^{(0)} = \widetilde{x}^{(0)}$, alors $\forall\ \varepsilon > 0 \quad \exists k_0 > 0$ tel que $\forall\ k \geq k_0$ on a*

$$\sup \frac{\widetilde{e}^{(k)}}{e^{(k)}} \geq \left(\frac{\rho\left(\widetilde{T}\right)}{\rho(T)+\varepsilon}\right)^k.$$

Donc la méthode itérative de matrice T converge plus rapidement que celle de matrice $\widetilde{T}$, en résumé, l'étude des méthodes itératives consiste à étudier les deux problèmes suivants :

1. Etant donné une méthode itérative de matrice T, déterminer si la méthode converge, ie si $\rho(T) < 1$ ou s'il existe une norme telle que $\|T\| < 1$.
2. Etant donné deux méthodes itératives convergentes de matrices T et $\widetilde{T}$, les comparer, la méthode la plus rapide est celle ayant le plus petit rayon spectral.

Définition 6.2.4. On appelle taux moyen de convergence sur k itérations le nombre $\widetilde{R}(k,T) = -\log\left\|T^k\right\|^{1/k}$ et taux asymptotique de convergence le nombre

$$R(T) = \lim_{k\longrightarrow+\infty} \widetilde{R}(k,T) = -\log(\rho(T)).$$

$R(T)$ joue le rôle de vitesse de convergence, plus $R(T)$ est grand plus rapide est la convergence.

6.3 Description des méthodes classiques

6.3.1 Méthode de Jacobi

Elle consiste à choisir $M = D = diag(a_{ii})$ inversible et $N = (-a_{ij})_{i \neq j}$ le schéma itératif est comme suit

$$x^{(k+1)} = D^{-1}(L+U)x^{(k)} + D^{-1}b.$$

La matrice $T_J = D^{-1}(L+U)$ est dite matrice de Jacobi associée à la matrice A. Si $x^{(0)}$ est le vecteur initial (donné), l'algorithme de Jacobi est de la forme

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{i \neq j} a_{ij} x_j^{(k)} + \frac{b_i}{a_{ii}} \quad i = 1, 2, \cdots, n.$$

Cet algorithme nécessite $a_{ii} \neq 0$ pour $i = 1, \cdots, n$ c.à.d D inversible .
Explicitement, on obtient :

$$\begin{array}{ccc} a_{11}x_1^{(k+1)} & = & -a_{12}x_2^{(k)} - \cdots - a_{1n}x_n^{(k)} + b_1 \\ \vdots & \vdots & \vdots \\ a_{nn}x_n^{(k+1)} & = & -a_{n1}x_1^{(k)} - \cdots - a_{nn-1}x_{n-1}^{(k)} + b_n \end{array}$$

On a besoin de stocker les n composantes de $x^{(k)}$ et les n composantes de $x^{(k+1)}$.
Matriciellement, le schéma itératif est du même type que le schéma (6.1.2) avec $T_J = D^{-1}(L+U)$ et $C = D^{-1}b$.
D'après les théorèmes précédents, une condition suffisante pour que la méthode de Jacobi converge est $\rho(T_J) < 1$ ou $\|T_J\|_\infty < 1$.

Théorème 6.3.1. *Si A est une matrice carrée à diagonale strictement dominante en lignes alors la méthode de Jacobi converge.*

Preuve :
On a

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}| \quad i = 1, \cdots, n \tag{6.3.1}$$

d'autre part on a : $t_{ij} = -\dfrac{a_{ij}}{a_{ii}}$ pour $i \neq j$ et $t_{ii} = 0$ d'où
$\|T_J\|_\infty = \max\limits_i \sum\limits_j |t_{ij}| = \max\limits_i \left\{ \dfrac{1}{|a_{ii}|} \sum\limits_{j \neq i} |a_{ij}| \right\}$ et d'après (6.3.1) on a $\|T_J\|_\infty < 1$.

Corollaire 6.3.1. *Si A est une matrice à diagonale strictemant dominante en colonnes, alors la méthode de Jacobi converge.*

Preuve : Identique à celle du théorème 6.3.1.
Pour le cas des matrices irréductibles la stricte dominance en lignes ou en colonnes peut être affaiblie, pour de telles matrices le théorème qui va suivre assure que

Théorème 6.3.2. *Si A est une matrice irréductible et vérifie $(|a_{ii}| \geq \sum_{k \neq i} |a_{ik}|, \quad i = 1, 2, \cdots, n)$, avec inégalité stricte pour au moins un indice i_0, alors la méthode de Jacobi converge.*

Preuve :
On procède d'une manière analogue à celle de la preuve du théorème 6.3.1 pour montrer que $\|T_J\|_\infty \leq 1$.
Donc

$$|T_J|e \leq e, \quad |T_J|e \neq e, \quad e = (1, 1, \cdots, 1)^\top. \tag{6.3.2}$$

Puisque A est irréductible, T_J est aussi irréductible.
Pour montrer le théorème, il suffit de montrer que

$$|T_J|^n e \leq e,$$

car si c'était le cas alors

$$(\rho(T_J))^n = \rho(T_J^n) \leq \|(|T_J|^n)\| < 1.$$

D'après (6.3.2) et le fait que $|T_J| \geq 0$, on a

$$|T_J|^2 e \leq |T_J|e < e,$$

et par suite

$$|T_J|^{i+1} e \leq |T_J|^i e \leq \cdots < e,$$

donc le vecteur $t^{(i)} = e - |T_J|^i e$ satisfait

$$0 < t^{(1)} \leq t^{(2)} \leq \cdots \tag{6.3.3}$$

Montrons que le nombre de composantes non nulles τ_i de $t^{(i)}$ croît avec i.
Si ce n'était pas le cas, (6.3.3) impliquerait qu'il existe $i \geq 1$ tel que $\tau_i = \tau_{i+1}$.
$t^{(i)}$ peut s'écrire sous la forme

$$\begin{pmatrix} a \\ 0 \end{pmatrix}, \quad a > 0, \ a \in \mathbb{R}^p.$$

(6.3.3) et $\tau_i = \tau_{i+1}$ impliquent que $t^{(i+1)}$ est aussi de la forme

$$\begin{pmatrix} b \\ 0 \end{pmatrix}, \quad b > 0, b \in \mathbb{R}^p.$$

Si $|T_J|$ est écrite sous la forme

$$|T_J| = \begin{pmatrix} |T_{11}| & |T_{12}| \\ |T_{21}| & |T_{22}| \end{pmatrix}, \quad |T_{11}| \text{ matrice } p \times p.$$

Il s'ensuit que

$$\begin{pmatrix} b \\ 0 \end{pmatrix} = t^{(i+1)} = e - |T_J|^{i+1} e \geq |T_J| e - |T_J|^{i+1} e$$
$$= |T_J| t^{(i)} = \begin{pmatrix} |T_{11}| & |T_{12}| \\ |T_{21}| & |T_{22}| \end{pmatrix} \begin{pmatrix} a \\ 0 \end{pmatrix}$$

Puisque $a > 0$, ceci n'est possible que si $T_{21} = 0$ ie T_J est réductible. Ceci contredit les hypothèses du théorème.
Donc $0 < \tau_1 < \tau_2 < \cdots$, et $t^{(n)} > 0$. La preuve est ainsi achevée.
Pour remédier au problème du stockage et dans l'espoir d'améliorer les résultats en accélérant la convergence, on cherche une méthode qui utilise les composantes de $x^{(k+1)}$ au fur et à mesure qu'elles sont calculées. C'est ce que réalise la méthode de Gauss-Seidel.

6.3.2 Méthode de Gauss-Seidel

Pour cette méthode, les matrices M et N sont données par : $M = D - L$ inversible et $N = U$ où D, L et U proviennent de l'écriture $A = D - L - U$, le schéma itératif est comme suit :

$$(D - L)x^{(k+1)} = Ux^{(k)} + b \tag{6.3.4}$$

ou encore

$$x^{(k+1)} = (D - L)^{-1} U x^{(k)} + (D - L)^{-1} b \tag{6.3.5}$$

en supposant que $D - L$ est inversible .
Les équations (6.3.4) et (6.3.5) peuvent aussi être présentées sous les formes :

$$Dx^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + b \tag{6.3.6}$$

et (si D est inversible)

$$x^{(k+1)} = D^{-1} L x^{(k+1)} + D^{-1} U x^{(k)} + D^{-1} b \tag{6.3.7}$$

en explicitant (6.3.6) on obtient :

$$\begin{aligned}
a_{11}x_1^{(k+1)} &= -a_{12}x_2^{(k)} - \cdots - a_{1n}x_n^{(k)} + b_1 \\
a_{22}x_2^{(k+1)} &= -a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)} + b_2 \\
\vdots \quad & \vdots \qquad\qquad \vdots \\
a_{ii}x_i^{(k+1)} &= -a_{i1}x_1^{(k+1)} \cdots - a_{ii-1}x_{i-1}^{(k+1)} - a_{ii+1}x_{i+1}^{(k)} - \cdots - a_{in}x_n^{(k)} + b_2 \\
\vdots \quad & \vdots \qquad\qquad \vdots \\
a_{nn}x_n^{(k+1)} &= -a_{n1}x_1^{(k+1)} - \cdots - a_{nn-1}x_{n-1}^{(k+1)} + b_n
\end{aligned}$$

La matrice $T_{GS} = (D-L)^{-1}U$ est dite matrice de Gauss-Seidel associée à la matrice A.

Remarque 6.3.1. Si D est inversible, la matrice de Gauss-Seidel s'écrit $T_{GS} = (I - D^{-1}L)^{-1}D^{-1}U$.

Théorème 6.3.3. *Si A est une matrice carrée à diagonale strictement dominante en lignes alors la méthode de Gauss-Seidel converge .*

Preuve :
Posons $T = (D-L)^{-1}U$ et montrons que $\|T\|_\infty < 1$ où $\|T\|_\infty = \max\limits_{x\neq 0} \dfrac{\|Tx\|_\infty}{\|x\|_\infty}$.
Soit $y = Tx = (D-L)^{-1}Ux$ on a alors $(D-L)y = Ux$ ou encore $Dy = Ly + Ux$ et $y = D^{-1}Ly + D^{-1}Ux$. Considérons l'indice i_0 tel que

$$|y_{i_0}| = \max_i |y_i| = \|y\|_\infty = \|Tx\|_\infty.$$

Il vient :

$$y_{i_0} = \sum_{j=1}^{i_0-1} (D^{-1}L)_{i_0 j} y_j + \sum_{j=i_0+1}^{n} (D^{-1}U)_{i_0 j} x_j.$$

Par suite

$$|y_{i_0}| = \|y\|_\infty \leq \sum_{j=1}^{i_0-1} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right| \|y\|_\infty + \sum_{j=i_0+1}^{n} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right| \|x\|_\infty.$$

En regroupant les termes

$$\left(1 - \sum_{j=1}^{i_0-1} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|\right) \frac{\|y\|_\infty}{\|x\|_\infty} \leq \sum_{j=i_0+1}^{n} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|$$

Par hypothèse, le terme $1 - \sum\limits_{j=1}^{i_0-1} \left|\dfrac{a_{i_0 j}}{a_{i_0 i_0}}\right|$ est strictement positif d'où on en tire :

$$\frac{\|Tx\|_\infty}{\|x\|_\infty} = \frac{\|y\|_\infty}{\|x\|_\infty} \leq \left(\sum_{j=i_0+1}^{n} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|\right) \left(1 - \sum_{j=1}^{i_0-1} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|\right)^{-1},$$

finalement

$$\max_{x \neq 0} \frac{\|Tx\|_\infty}{\|x\|_\infty} < 1.$$

Remarques 6.3.2.
1. Un résultat de convergence similaire a lieu si A est à diagonale dominante en colonnes.
2. Si on se place dans les conditions du théorème 6.3.2, la méthode de Gauss-Seidel est convergente.

6.3.3 Méthode de relaxation

Si on considère des matrices M et N dépendantes d'un paramètre ω
on obtient : $A = M(\omega) - N(\omega)$.
Prenons $M(\omega) = \frac{1}{\omega}D - L$ et $N(\omega) = \frac{1-\omega}{\omega}D + U$, en supposant $M(\omega)$ inversible.
Le schéma itératif qui en résulte est le suivant :

$$x^{(k+1)} = \left(\frac{1}{\omega}D - L\right)^{-1} \left(\frac{1-\omega}{\omega}D + U\right) x^{(k)} + \left(\frac{1}{\omega}D - L\right)^{-1} b \tag{6.3.8}$$

l'équation (6.3.8) peut être remplacée par :

$$x^{(k+1)} = \left(\frac{1}{\omega}D\right)^{-1} Lx^{(k+1)} + \left((1-\omega)I + \frac{1}{\omega}D^{-1}U\right) x^{(k)} + \left(\frac{1}{\omega}D^{-1}\right) b \tag{6.3.9}$$

La matrice de relaxation est donnée par $T_\omega = \left(\frac{1}{\omega}D - L\right)^{-1} \left(\frac{1-\omega}{\omega}D + U\right)$.

Remarques 6.3.3.
– Si D est inversible, $T_\omega = \left(I - \omega D^{-1}L\right)^{-1} \left((1-\omega)I + \omega D^{-1}U\right)$.
– Si $\omega = 1$, on retrouve la méthode de Gauss-Seidel.
– Si $\omega > 1$, on parle de sur-relaxation.
– Si $\omega < 1$, on parle de sous-relaxation.

Ici la condition de convergence $\|T_\omega\| < 1$ dépendra du paramètre ω et par conséquent, on est amené à chercher tous les ω pour lesquels il y a convergence et ensuite choisir la valeur optimale ω_0 de telle sorte que la vitesse de convergence soit la meilleure possible.

Théorème 6.3.4. *Soit A une matrice hermitienne et inversible définie positive ($A = M - N$) telle que M soit inversible, et la matrice $M^* + N$ soit définie positive, alors le schéma (6.1.4) converge si et seulement si la matrice A est définie positive.*

Preuve :

i) Supposons que A est définie positive

$$A = A^* \Longrightarrow (M-N)^* = M^* - N^* = M - N.$$

D'où $M^* + N = M^* + M - A = M + N^* = (M^* + N)^*$. Par conséquent si A est hermitienne alors $M^* + N$ est aussi hermitienne. Comme A est définie positive, l'application $v \longrightarrow (v^*Av)^{1/2}$ définit une norme : $v \longrightarrow \|v\|_A = (v^*Av)^{1/2}$. Considérons alors la norme matricielle induite par $\|.\|_A$ on a :

$$\left\|M^{-1}N\right\|_A = \left\|I - M^{-1}A\right\|_A = \sup_{v/\|v\|_A=1} \left\|(I - M^{-1}A)v\right\|_A$$

En posant $\omega = M^{-1}Av$ il vient que $M\omega = Av$ et on est amené à travailler sur $\|v - \omega\|_A$ avec $\|v\|_A = 1$ on a :

$$\begin{aligned}
\|v-\omega\|_A^2 &= (v-\omega)^*A(v-\omega) \\
&= \|v\|_A^2 - v^*A\omega - \omega^*Av + \|\omega\|_A^2 \\
&= 1 - \omega^*M\omega - \omega^*M^*\omega + \omega^*A\omega \\
&= 1 - \omega^*M^*\omega - \omega^*N\omega \\
&= 1 - \omega^*(M^*+N)\omega
\end{aligned}$$

On a $v \neq 0$ donc $\omega \neq 0$ et $M^* + N$ est définie positive donc $\omega^*(M^* + N)\omega > 0$ par conséquent $\|v - \omega\|_A^2 < 1$ et $\left\|M^{-1}N\right\|_A < 1$.

ii) Réciproquement, posons $T = I - M^{-1}A$ et $R = AM^{*-1}(M^* + N)M^{-1}A$.

$$\langle Rx, x\rangle = \langle (M + M^* - A)y, y\rangle \text{ avec } y = M^{-1}Ax.$$

Or la matrice $M + M^* - A = M + N$ est définie positive et A est inversible, on a donc $\langle Rx, x\rangle > 0 \quad \forall x \neq 0$.
Par suite R est hermitienne définie positive.
En remarquant que $A = R + T^*AT$, on obtient

$$\begin{aligned}
A &= R + T^*(R + T^*AT)T = R + T^*RT + T^{*2}(R + T^*AT)T^2 \\
&= \sum_{i=0}^{k-1} T^{*i}RT^i + T^{*k}RT^k.
\end{aligned}$$

Par hypothèse le schéma (6.1.4) est convergent, donc

$$\rho\left(M^{-1}N\right) < 1,$$

or $M^{-1}N = I - M^{-1}A = T$ donc $\rho(T) < 1$ et on a

$$\lim_{k\longrightarrow\infty} T^k = \lim_{k\longrightarrow\infty} T^{*k} = 0.$$

Donc

$$A = \sum_{i=0}^{\infty} T^{*i} R T^i = R + \sum_{i=0}^{\infty} T^{*i} R T^i.$$

Puisque R est définie positive, et que $\langle T^{*k} R T^k x, x \rangle \geq 0, \forall x$, il en résulte que A est définie positive.

Théorème 6.3.5 (Condition nécessaire de convergence)**.**
Si A est une matrice hermitienne définie positive alors la méthode de relaxation converge si $\omega \in]0, 2[$.

Preuve :
D'après le théorème précédent, si A est hermitienne définie positive et $M^* + N$ définie positive alors la méthode converge. Il suffit donc que $M^* + N$ soit définie positive.
Or $M^* + N = \dfrac{2-\omega}{\omega} D$ et par suite $M^* + N$ est définie positive si $\dfrac{2-\omega}{\omega} > 0$ c.à.d si $0 < \omega < 2$.

Théorème 6.3.6 (Kahan)**.**
Le rayon spectral de la matrice de relaxation vérifie toujours l'inégalité
$\rho(T_\omega) \geq |\omega - 1|, \omega \neq 0$ il s'ensuit que la méthode de relaxation ne peut converger que si $\omega \in]0, 2[$.

Preuve :

$$T_\omega = \left(\frac{1}{\omega} D - L\right)^{-1} \left(\frac{1-\omega}{\omega} D + U\right)$$

Si les valeurs propres de T_ω sont notées $\lambda_i(\omega)$ on a :

$$\det T_\omega = \prod_{i=1}^{n} \lambda_i(\omega) = \frac{\det\left(\dfrac{1-\omega}{\omega} D + U\right)}{\det\left(\dfrac{1}{\omega} D - L\right)} = (1-\omega)^n.$$

D'où $\rho(T_\omega) \geq [|1-\omega|^n]^{\dfrac{1}{n}} = |1-\omega|$.
Pour que la méthode converge, il est nécessaire d'avoir $\rho(T_\omega) < 1$ et par conséquent $|1-\omega| < 1$ d'où $\omega \in]0, 2[$.

Remarque 6.3.4. Le résultat reste valable si A est tridiagonale par blocs (voir Ciarlet(1984) ou Sibony(1986)).

6.4 Comparaison des méthodes classiques

6.4.1 Comparaison des méthodes de Jacobi et de Gauss-Seidel

Théorème 6.4.1. *Soit A une matrice tridiagonale. Alors les méthodes de Jacobi et de Gauss-Seidel convergent ou divergent simultanément ; lorsqu'elles convergent, la méthode de Gauss-Seidel est plus rapide que celle de Jacobi, plus précisément, on a* $\rho(T_{GS}) = (\rho(T_J))^2$ *où T_{GS} et T_J sont les matrices de Gauss-Seidel et Jacobi (respectivement).*

Preuve :

$$T_{GS} = (D-L)^{-1}U \text{ et } T_J = D^{-1}(L+U)$$

On est donc amené à chercher les valeurs propres de ces deux matrices.
1) Soit λ une valeur propre de T_J. Alors λ est racine du polynôme caractéristique $P_J(\lambda)$

$$\begin{aligned} P_J(\lambda) &= \det\left(D^{-1}(L+U) - \lambda I\right) \\ &= \det\left(-D^{-1}\right)\det(\lambda D - (L+U)) \\ &= K_1 \det(\lambda D - (L+U)) \end{aligned}$$

2) Soit α une valeur propre de T_{GS}. Alors α est racine du polynôme caractéristique $P_{GS}(\alpha)$.

$$\begin{aligned} P_{GS}(\alpha) &= \det\left((D-L)^{-1}(U) - \alpha I\right) \\ &= \det\left(-(D-L)^{-1}\right)\det(\alpha D - \alpha L - U) \\ &= K_2 det(\alpha D - \alpha L - U). \end{aligned}$$

Comme il s'agit de matrices tridiagonales, on a le résultat suivant (voir exercice (3.7.2))

$$P_{GS}(\alpha) = K_2 \det\left(\alpha D - \alpha\mu L - \mu^{-1}U\right) \text{ pour tout } \mu \neq 0.$$

En particulier pour $\mu = \alpha^{-1/2}$ on obtient

$$P_{GS}(\alpha) = K_2 \alpha^{n/2} \det\left(\alpha^{1/2}D - (L+U)\right).$$

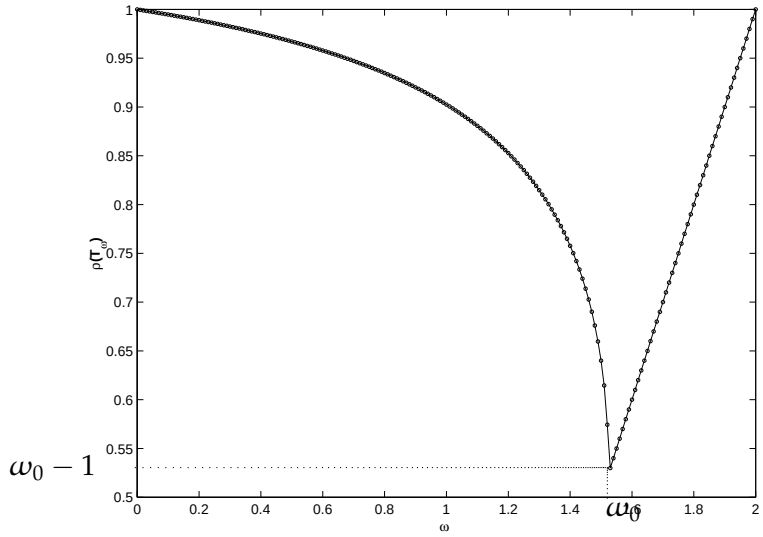
On voit bien donc que $P_{GS}(\alpha) = K_{12}\alpha^{n/2}P_J\left(\alpha^{1/2}\right)$. Par suite, si β est une valeur propre de T_J alors β^2 est une valeur propre de T_{GS}.
Réciproquement, si β^2 est une valeur propre non nulle de T_{GS} alors β et $-\beta$ sont des valeurs propres de T_J. En effet, en prenant $\mu = -1$, il vient que $\det(\beta D - L - U) = \det(\beta D + L + U) = 0$ donc $P_J(\beta) = 0$ et $P_J(-\beta) = 0$.

6.4.2 Comparaison des méthodes de Jacobi et de relaxation

Théorème 6.4.2. *Soit A une matrice tridiagonale, telle que toutes les valeurs propres de la matrice de Jacobi associée soient réelles. Alors les méthodes de Jacobi et de relaxation, pour $\omega \in]0,2[$, convergent où divergent simultanément. Lorsqu'elles convergent, on peut déterminer une valeur optimale ω_0 du paramètre ω telle que*

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(T_J)^2}}, \quad \rho(T_{\omega_0}) = \inf_{\omega\in]0,2[} \rho(T_\omega) = \omega_0 - 1.$$

Plus précisément, si on considère la fonction qui à ω fait correspondre $\rho(T_\omega)$ alors cette fonction a l'allure suivante



Remarques 6.4.1.

1. $\rho(T_{GS}) = \rho(T_1)$ est obtenue comme cas particulier avec $\omega = 1$ on sait d'après le théorème précédent que $\rho(T_{GS}) = (\rho(T_J))^2$.

2. On voit que pour la valeur ω_0 de ω on obtient

$$\rho(T_{\omega_0}) < \rho(T_{GS}) < \rho(T_J)$$

qui indique qu'avec un choix approprié de ω, la méthode de relaxation converge plus rapidement que celles de Gauss-Seidel et de Jacobi.

Preuve :

Notons α une racine de P_J et notons λ une racine de P_ω, On obtient alors

$$\begin{aligned}
P_J(\alpha) &= K_1 \det(\alpha D - L - U) \\
\text{et } P_\omega(\lambda) &= K_\omega \det\left(\frac{1-\omega-\lambda}{\omega}D + \lambda L + U\right) \\
&= (-1)^n K_\omega \det\left(\frac{\lambda+\omega-1}{\omega}D - \lambda L - U\right) \\
P_\omega(\lambda^2) &= \widetilde{K}_\omega \det\left(\frac{\lambda^2+\omega-1}{\omega}D - \lambda^2 L - U\right) \\
&= \widetilde{K}_\omega \det\left(\frac{\lambda^2+\omega-1}{\omega}D - \lambda L - \lambda U\right), \quad \lambda \neq 0
\end{aligned}$$

Soit encore :

$$P_\omega(\lambda^2) = \widetilde{K}_\omega \lambda^n \det\left(\frac{\lambda + \frac{(\omega-1)}{\lambda}}{\omega}D - L - U\right) \quad \text{pour tout} \quad \lambda \neq 0$$

Par identification, on voit que :

$$\begin{aligned}
P_\omega(\lambda^2) &= \widetilde{K}_{\omega,\lambda} P_J\left(\frac{\lambda + \frac{(\omega-1)}{\lambda}}{\omega}\right) \\
&= \widetilde{K}_{\omega,\lambda} P_J\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right)
\end{aligned}$$

Il s'ensuit qu'on peut déterminer une relation liant les valeurs propres non nulles de T_ω et celles de T_J.

Soit $\lambda^2 \neq 0$ une valeur propre de T_ω donc $\pm\dfrac{\lambda^2+\omega-1}{\lambda\omega}$ est une valeur propre de T_J. Inversement, si α est valeur propre de T_J, on sait que $-\alpha$ est aussi valeur propre de T_J et ceci implique que λ_1^2 et λ_2^2 sont valeurs propres de T_ω où λ_1 et λ_2 sont données par :

$$\begin{aligned}
\lambda_1^2 &= \frac{1}{2}\left(\alpha^2\omega^2 - 2\omega + 2\right) - \frac{\alpha\omega}{2}\sqrt{(\alpha^2\omega^2 + 4 - 4\omega)} \\
\lambda_2^2 &= \frac{1}{2}\left(\alpha^2\omega^2 - 2\omega + 2\right) + \frac{\alpha\omega}{2}\sqrt{(\alpha^2\omega^2 + 4 - 4\omega)}
\end{aligned}$$

λ_1 et λ_2 proviennent de la résolution de l'équation du second degré en λ, à savoir

$$\alpha = \frac{\lambda^2+\omega-1}{\lambda\omega}$$

ou encore

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0 \quad (*)$$

on a

$$\Delta(\alpha) = \alpha^2\omega^2 - 4\omega + 4 = \alpha^2\omega^2 - 4(\omega - 1) = \alpha^2\omega^2 + 4(1 - \omega)$$

1. Si $\alpha \geq 1$, on est dans le cas où la méthode de Jacobi diverge.
   Considérons alors $\Delta(\alpha)$ comme équation du second degré en ω : $\Delta(\alpha) = \alpha^2\omega^2 - 4\omega + 4$, son discriminant est $\Delta' = 4 - 4\alpha^2 < 0$ d' où $\Delta(\alpha)$ est toujours du signe de α^2 c'est à dire positif.
   $\Delta(\alpha) \geq 0$ implique que (*) admet deux solutions réelles

$$\lambda_1 = \frac{\alpha\omega - \sqrt{\Delta(\alpha)}}{2} \text{ et } \lambda_2 = \frac{\alpha\omega + \sqrt{\Delta(\alpha)}}{2}$$

   on en tire

$$\lambda_1^2 = \frac{1}{2}\left(\alpha^2\omega^2 - 2\omega + 2\right) - \frac{\alpha\omega}{2}\sqrt{(\alpha^2\omega^2 + 4 - 4\omega)}$$

$$\text{et} \quad \lambda_2^2 = \frac{1}{2}\left(\alpha^2\omega^2 - 2\omega + 2\right) + \frac{\alpha\omega}{2}\sqrt{(\alpha^2\omega^2 + 4 - 4\omega)}$$

   on a

$$\begin{aligned} \lambda_2^2 &\geq \frac{1}{2}\left(\omega^2 - 2\omega + 2\right) + \frac{\omega}{2}|\omega - 2| \text{ si } \alpha \geq 1 \\ &\geq \frac{1}{2}\left(\omega^2 - 2\omega + 2 + 2\omega - \omega^2\right) = 1 \end{aligned}$$

   Donc si la méthode de Jacobi diverge ($\alpha \geq 1$) alors on a aussi $\rho(T_\omega) \geq 1$, par conséquent la méthode de relaxation diverge aussi.

2. Si $|\alpha| < 1$
   On est amené à étudier

$$\rho(T_\omega) = \max_{\alpha \in Sp(T_J)} \left\{ \max\left(\left|\lambda_1^2(\alpha, \omega)\right|, \left|\lambda_2^2(\alpha, \omega)\right| \right) \right\}$$

   pour cela, considérons la fonction f définie par :

$$f : \mathbb{R}_+ \times]0, 2[\longrightarrow \mathbb{R}_+$$

$$(\alpha, \omega) \longrightarrow \max\left\{ \left|\lambda_1^2\right|, \left|\lambda_2^2\right| \right\}$$

   Remarque 6.4.2. on n 'a pas besoin de considérer $\alpha \in \mathbb{R}_-$ car
   $f(\alpha, \omega) = f(-\alpha, \omega)$.
   i) D'abord si $\alpha = 0$ alors $f(0, \omega) = |1 - \omega|$.

ii) Si $0 < \alpha < 1$ alors le trinôme $\omega \longrightarrow \alpha^2\omega^2 - 4\omega + 4 = \Delta(\alpha)$ considéré Comme équation de second degré en ω admet pour discriminant, $\Delta' = 4 - 4\alpha^2 > 0$ d' où deux racines réelles :

$$\omega_0(\alpha) = \frac{2 - \sqrt{4 - 4\alpha^2}}{\alpha^2} = \frac{4\alpha^2}{2\alpha^2\left(1 + \sqrt{1 - \alpha^2}\right)} = \frac{2}{1 + \sqrt{1 - \alpha^2}}$$

$$\omega_1(\alpha) = \frac{2 + \sqrt{4 - 4\alpha^2}}{\alpha^2} = \frac{4\alpha^2}{2\alpha^2\left(1 - \sqrt{1 - \alpha^2}\right)} = \frac{2}{1 - \sqrt{1 - \alpha^2}}$$

$\omega_0(\alpha)$ et $\omega_1(\alpha)$ satisfont les inégalités

$$1 < \omega_0(\alpha) = \frac{2}{1 + \sqrt{1 - \alpha^2}} < 2 < \omega_1(\alpha)$$

Donc pour α telle que $0 < \alpha < 1$ et pour ω compris entre ω_0 et ω_1 (en particulier $\omega_0 < \omega < 2$) $\Delta(\alpha, \omega)$ est négatif. Ce qui veut dire que pour ces valeurs, λ_1^2 et λ_2^2 sont des complexes conjugués.

Comme λ_1 et λ_2 vérifient $\lambda_1 + \lambda_2 = \alpha\omega$ et $\lambda_1.\lambda_2 = \omega - 1$.

Or $\lambda_1\lambda_2 = |\lambda_1|^2 = \left|\lambda_1^2\right| = \left|\lambda_2^2\right| = \omega - 1$ et par suite $f(\alpha, \omega) = \omega - 1$.

On vérifie que $f(\alpha, \omega) = \left|\lambda_2^2\right| = \lambda_2^2$ car $\Delta(\alpha) > 0$ et on a deux racines réelles λ_1 et λ_2 et on est amené à étudier l'allure des courbes $f(\alpha, \omega)$ on obtient $\dfrac{\partial f}{\partial \omega} < 0$ pour $0 < \alpha < 1$ et $0 < \omega < \omega_0(\alpha)$ donc la fonction $\omega \longrightarrow f(\alpha, \omega)$ est décroissante.

(i) Si $0 < \omega < \omega_0(\alpha)$.

Puisque $0 < \omega < \omega_0(\alpha)$ alors $f(\alpha, \omega) = \left|\lambda_2^2\right| = \lambda_2^2$.

$\Delta(\alpha) > 0$ donc on a deux racines réelles

$$\frac{\partial f}{\partial \omega} = \left(\frac{\alpha}{2} + \frac{\omega\alpha^2 - 2}{2\sqrt{\Delta(\alpha)}}\right) 2\lambda_2$$

en utilisant $2\lambda_2 = \alpha\omega + \sqrt{\Delta(\alpha)}$ donc $\sqrt{\Delta(\alpha)} = 2\lambda_2 - \alpha\omega$ d'où

$$\begin{aligned}
\frac{\partial f}{\partial \omega} &= \left(\frac{\alpha}{2} + \frac{\omega\alpha^2 - 2}{2(2\lambda - \alpha\omega)}\right) 2\lambda_2 \\
&= \left(\frac{2\alpha\lambda - \omega\alpha^2 + \omega\alpha^2 - 2}{2(2\lambda - \alpha\omega)}\right) 2\lambda \\
&= \left(\frac{\alpha\lambda - 1}{2\lambda - \alpha\omega}\right) 2\lambda_2 \\
&= 2\lambda_2\left(\frac{\lambda_2\alpha - 1}{2\lambda_2 - \alpha\omega}\right)
\end{aligned}$$

Donc $\frac{\partial f}{\partial \omega} < 0$ car $\alpha\lambda_2 < \lambda_2 < 1$ pour $0 < \alpha < 1$ en effet

$$\begin{aligned}
\lambda_2^2 &= \frac{\alpha^2\omega^2 - 2\omega + 2}{2} + \frac{\alpha\omega}{2}\sqrt{(\alpha^2\omega^2 - 4\omega + 4)} \\
&< \frac{\omega^2 - 2\omega + 2}{2} + \frac{\omega}{2}\sqrt{(\omega^2 - 4\omega + 2)} \\
&< \frac{\omega^2 - 2\omega + 2}{2} + \frac{\omega}{2}(2 - \omega) \\
&< 1
\end{aligned}$$

Par ailleurs

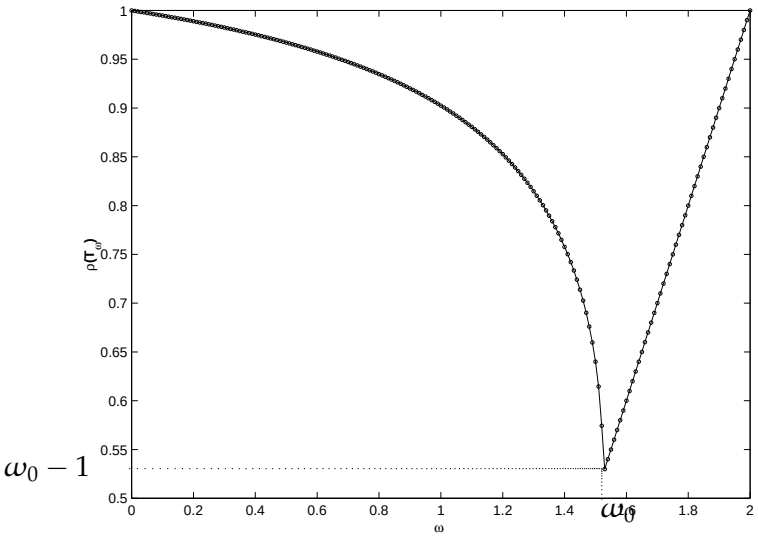
1- $\frac{\partial f}{\partial \omega}(0) = 0$ donc $\Delta(\alpha)(\omega = 0) = 4$

et $\lambda_2(\omega = 0) = \frac{\sqrt{\Delta(\alpha)(\omega = 0)}}{2} = \frac{\sqrt{4}}{2} = 1$par conséquent

$$\begin{aligned}
\frac{\partial f}{\partial \omega}(\omega = 0) &= 2\lambda_2 \left(\lambda_2 + \frac{\omega\lambda^2 - 2}{\sqrt{\Delta(\alpha)(\omega = 0)}} \right) \\
&= 2\lambda_2(0) \left(\lambda_2(0) - \frac{2}{\sqrt{\Delta(\alpha)(\omega = 0)}} \right) \\
&= 2(1 - \frac{2}{2}) \\
&= 0
\end{aligned}$$

2- $\frac{\partial f}{\partial \omega}(\omega_0) = \infty$ car ω_0 est racine de $\Delta(\alpha)$ d'où $\frac{1}{\Delta(\alpha)}$ tend vers l'infini quand ω tend vers ω_0.

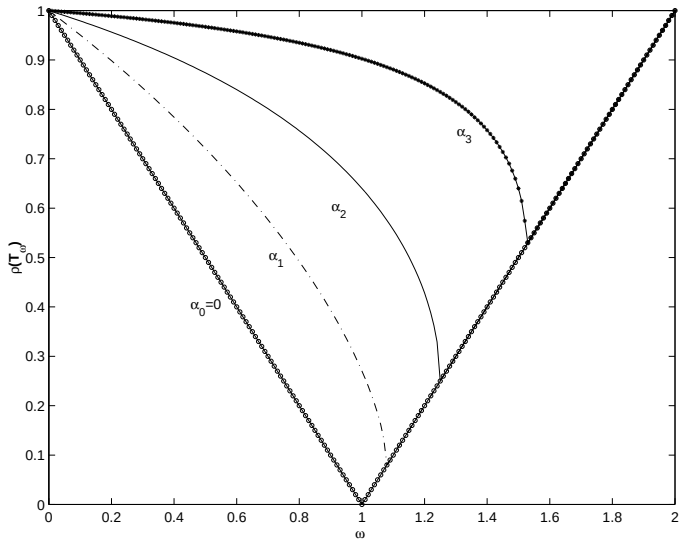
En conclusion, pour $0 < \alpha < 1$ et $0 < \omega < \omega_0$ la courbe de f comme fonction de ω a l'allure suivante



Enfin, remarquons que la fonction :

$$\alpha \longrightarrow \frac{2}{1+\sqrt{1-\alpha^2}} \quad \text{est croissante en fonction de } \alpha$$

donc le maximum est atteint pour $\rho(T_J)$



6.5 Méthodes semi-itératives

Dans la première partie de ce chapitre, nous avons considéré les méthodes itératives du type $x^{(k+1)} = Tx^{(k)} + C$, impliquant deux itérés successifs $x^{(k)}$ et $x^{(k+1)}$. A présent, nous considérons des méthodes semi-itératives permettant d'exprimer $y^{(k)}$ comme combinaison algébrique des k itérés précédents $x^{(0)}, x^{(1)}, \cdots, x^{(k-1)}$. Nous obtenons l'expression :

$$y^{(k)} = \sum_{j=0}^{k} \theta_j(k) x^{(j)}, \quad k \geq 0,$$

avec $\sum_{j=0}^{k} \theta_j(k) = 1, \quad k \geq 0.$

Si $e^{(j)} = x^{(j)} - x$ est l'erreur de la j^{eme} itération et $\varepsilon^{(j)} = y^{(j)} - x$ est l'erreur de la méthode semi-itérative alors on a :

$$\varepsilon^{(k)} = \sum_{j=0}^{k} \theta_j(k) e^{(j)},$$

et comme $e^{(k)} = T^k e^{(0)}$, il vient :

$$\varepsilon^{(k)} = \left(\sum_{j=0}^{k} \theta_j(k) T^j \right) e^{(0)},$$

ou encore

$$\varepsilon^{(k)} = Q_k(T) e^{(0)}, \tag{6.5.1}$$

où $Q_k(x)$ est le polynôme $\quad Q_k(x) = \sum_{j=0}^{k} \theta_j(k) x^j.$

Remarques 6.5.1.

1. Si on prend $\theta_j(k) = \dfrac{1}{k+1}$ pour $j = 0, 1, \cdots, k$, on retrouve $y^{(k)}$ comme moyenne arithmétique des $x^{(k)}$.
2. La méthode de Richardson donnée à l'exercice 3.7.6 peut être considérée comme un cas particulier des méthodes semi-itératives.

Théorème 6.5.1. *Si les coefficients des polunômes $Q_k(x)$ sont réels et si la matrice T est hermitienne et ses valeurs propres vérifient $-1 \leq a \leq \lambda_1 \leq \cdots \leq \lambda_n \leq b \leq 1$, alors*

$$\|Q_k(T)\|_2 = \max_{\lambda_i \in Sp(T)} |Q_k(\lambda_i)| \leq \max_{a \leq \lambda \leq b} |Q_k(\lambda)|.$$

En revenant à l'équation (6.5.1), nous sommes donc amenés au problème de minimisation suivant :

$$\min_{Q_k(1)=1} |Q_k(x)|.$$

Ceci conduit au problème bien connu de *minmax* de Chebyshev :

$$\min_{Q_k(1)=1} \max_{-1\leq a\leq x\leq b\leq 1} |Q_k(x)|,$$

dont la solution est donnée par les polynômes de Chebyshev (voir Varga(2000), Golub(1989))

$$C_k(x) = \begin{cases} \cos(k\cos^{-1}(x)), -1 \leq x \leq 1, \ \ k \geq 0 \\ \cosh(k\cosh^{-1}(x)), x \geq 1, \ \ k \geq 0 \end{cases}$$

6.6 Décomposition des matrices positives

Théorème 6.6.1. *Soit B une matrice carrée ; si $B \geq 0$ alors les assertions suivantes sont équivalentes :*

i) $\beta > \rho(B)$.

ii) $\beta I - B$ *est inversible et on a :*

$$(\beta I - B)^{-1} \geq 0$$

Preuve :

Supposons que $\beta > \rho(B)$.

En posant $M = (\beta I - B) = \beta(I - B_\beta)$ où $B_\beta = \dfrac{1}{\beta}B$, il est évident que $\rho(B_\beta) < 1$ et il découle du théorème **??** que B est inversible et $(\beta I - B)^{-1} \geq 0$.

Réciproquement, en supposant ii) et en utilisant le théorème **??** avec $B \geq 0$, si $v \geq 0$ est un vecteur propre de B associé à la valeur propre $\rho(B)$ alors v est aussi vecteur propre de $(\beta I - B)$ associé à la valerur propre $(\beta - \rho(B))$.

De l'inversibilité de $(\beta I - B)$ il s'ensuit que $(\beta - \rho(B)) \neq 0$ et par suite

$(\beta I - B)^{-1}v = \dfrac{1}{(\beta - \rho(B))}v$.

Enfin, v étant un vecteur propre ≥ 0 (non identiquement nul) et $(\beta I - B)^{-1} \geq 0$ entraîne que $(\beta - \rho(B)) > 0$.

Théorème 6.6.2. *Soit A une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $D = (a_{ii})$, on a équivalence entre :*

i) A *est inversible et* $A^{-1} \geq 0$

et

ii) $D > 0$, *et en posant* $B = I - D^{-1}A$ *on a* $B \geq 0$ *et* B *est convergente.*

Preuve :
Supposons que A est inversible et $A^{-1} = (\alpha_{ij})$ avec $\alpha_{ij} \geq 0 \ \forall \ i, j = 1, n$, en explicitant $A^{-1}A = I$, il vient

$$\alpha_{ii}a_{ii} + \sum_{j \neq i} \alpha_{ij}a_{ji} = 1 \quad \text{pour tout} \quad i = 1, \cdots, n$$

et par suite $\alpha_{ii}a_{ii} = 1 - \sum_{j \neq i} \alpha_{ij}a_{ji} \geq 1 \ \forall \ i = 1, \cdots, n$ du fait que $a_{ij} \leq 0$ et $\alpha_{ij} \geq 0$ pour tout $i \neq j$.
On a donc $a_{ii} > 0 \ \forall \ i = 1, \cdots, n$ et la matrice D est inversible avec $D^{-1} > 0$, d'où $B = I - D^{-1}A \geq 0$ et $I - B = D^{-1}A$ est inversible comme produit de deux matrices inversibles.
Enfin, la convergence de B ($\rho(B) < 1$) s'obtient en remarquant que
$(I - B)^{-1} = A^{-1}D \geq 0$ puis en appliquant le théorème 6.6.1 avec $\beta = 1$.
Réciproquement, toujours d'après le théorème 6.6.1, on a $I - B$ inversible et
$(I - B)^{-1} \geq 0$ ce qui entraîne que $A^{-1}D \geq 0$ et $A^{-1} \geq 0$.
Autres résultats
Avec des hypothèses suppémentaires, notamment d'irréductibilité, on obtient les résultats suivants (voir Varga [**?**]).

Théorème 6.6.3. *Si $A \geq 0$ alors les assertions suivantes sont équivalentes :*
i) $\alpha > \rho(A)$ et A est irreductible.
ii) $\alpha I - A$ est inversible et on a :

$$(\alpha I - A)^{-1} > 0$$

Théorème 6.6.4. *Soit A une matrice réelle avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $D = (a_{ii})$, on a équivalence entre :*
i) A est inversible et $A^{-1} > 0$ et
ii) $D > 0$ et en posant $B = I - D^{-1}A$ on a $B \geq 0$ et B est irreductible et convergente.

Théorème 6.6.5. *Soit A une matrice réelle irreductible à diagonale dominante avec $a_{ij} \leq 0$ pour tout $i \neq j$ et $a_{ii} > 0 \quad \forall i = 1, \cdots, n$ alors $A^{-1} > 0$.*

Théorème 6.6.6. *Soit A une matrice réelle symétrique inversible et irreductible avec $a_{ij} \leq 0$ pour tout $i \neq j$, alors $A^{-1} > 0$ si et seulement si A est définie positive.*

6.6.1 Décomposition régulière des matrices

Définition 6.6.1. Soient A, M et N des matrices carrées d'ordre n. Une décomposition $A = M - N$ est dite régulière si : M est inversible avec $M^{-1} \geq 0$ et $N \geq 0$.

La décomposition est dite régulière faible si : M est inversible avec $M^{-1} \geq 0$ et $M^{-1}N \geq 0$.

Théorème 6.6.7. *Soit $A = M - N$ une décomposition régulière de A. Alors A est inversible avec $A^{-1} \geq 0$ si et seulement si $\rho(M^{-1}N) < 1$ où*

$$\rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1$$

et par conséquent, si A est inversible avec $A^{-1} \geq 0$ alors la méthode $x^{(k+1)} = M^{-1}Nx^{(k)} + c$ est convergente pour n'importe quel choix initial.

Preuve :

Supposons A monotone.

En posant $B = A^{-1}N$, on peut aisement voir que

- $M^{-1}N = (I + B)^{-1}B$,
- si $M^{-1}Nv = \lambda v$ alors $Bv = \mu v$ avec $\mu = \dfrac{\lambda}{1-\lambda}$ et $\lambda = \dfrac{\mu}{1+\mu}$,
- $M^{-1}N \geq 0$ et $B \geq 0$,
- la fonction $\mu \longrightarrow \dfrac{\mu}{1+\mu}$ est strictement croissante pour $\mu \geq 0$ d'où $\rho(M^{-1}N) = \dfrac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1$ et par conséquent la méthode converge.

Réciproquement, si $\rho(M^{-1}N) < 1$ alors$(I - M^{-1}N)$ est inversible et on a successivement

$$M^{-1}A = I - M^{-1}N \quad \text{et } A^{-1} = (I - M^{-1}N)^{-1}M^{-1},$$

comme $M^{-1} \geq 0$ et que le théorème **??** donne $(I - M^{-1}N)^{-1} \geq 0$ on en déduit que $A^{-1} \geq 0$.

Théorème 6.6.8. *Soient $A = M_1 - N_1 = M_2 - N_2$ deux décompositions régulières de A avec $A^{-1} \geq 0$.*
Si $0 \leq N_1 \leq N_2$ alors $0 \leq \rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) \leq 1$.
Si de plus $A^{-1} > 0$ et si $0 < N_1 < N_2$ alors $0 < \rho(M_1^{-1}N_1) < \rho(M_2^{-1}N_2) < 1$.

Preuve :

On a $A^{-1}N_1 \leq A^{-1}N_2$ et par suite $\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) \leq 1$.

6.7 Comparaison des méthodes classiques dans le cas des matrices positives

Soit la décomposition $A = D - E - F$ où $D = (a_{ii})_{i=1,\cdots,n}$ et E et$F$ sont respectivement triangulaire inférieure et supérieure. En supposant D et $(D - E)$ inversibles

et en posant $L = D^{-1}E$ et $U = D^{-1}F$, on retrouve les matrices de Jacobi $T_J = L + U$ et de Gauss-Seidel $T_{GS} = (I - L)^{-1}U$ qu'on peut comparer dans certains cas particuliers.

Théorème 6.7.1 (Stein & Rosenberg)**.**
Soit $T_J = L + U$ la matrice de Jacobi supposée positive avec diagonale nulle et $T_{GS} = (I - L)^{-1}U$ la matrice de Gauss-Seidel. Alors une et une seule des relations suivantes a lieu

1. $\rho(T_J) = \rho(T_{GS}) = 0$,
2. $0 < \rho(T_{GS}) < \rho(T_J) < 1$,
3. $\rho(T_J) = \rho(T_{GS}) = 1$,
4. $1 < \rho(T_J) < \rho(T_{GS})$.

Corollaire 6.7.1. *Si la matrice de Jacobi est positive avec $\rho(T_J) < 1$, alors*

$$R(T_J) < R(T_{GS}).$$

Remarques 6.7.1.

1. Le théorème de Stein & Rosenberg affirme donc que les matrices de Jacobi et de Gauss-Seidel convergent ou divergent simultanément. Lorsqu'elles convergent, la matrice de Gauss-Seidel converge asymptotiquement plus rapidement.
2. En prenant $L + U$ positive et en posant

   $$M_1 = I; N_1 = L + U,$$

   $$M_2 = I - L; N_2 = U.$$

   Le théorème 6.6.8 permet d'obtenir directement le point 2 du théorème de Stein & Rosenberg.

Théorème 6.7.2. *Soit $A = I - B$ où $B = L + U$ est positive, irreductible et convergente avec L et U, respectivement, triangulaire strictement inférieure et triangulaire strictement supérieure alors la matrice de relaxation définie par*

$$M_\omega = (I - \omega L)^{-1}(\omega U + (1 - \omega)I)$$

est convergente pour $0 < \omega \leq 1$. De plus, si $0 < \omega_1 < \omega_2 \leq 1$, alors

$$0 < \rho(M_{\omega_2}) < \rho(M_{\omega_1}) < 1$$

et par conséquent $R(M_{\omega_1}) < R(M_{\omega_2})$.

6.8 Complément bibliographique

Selon le type de matrice à savoir, positive, définie positive, symétrique, hermitienne, à diagonale dominante, tridiagonale ou autre, plusieurs auteurs se sont intéressés aux méthodes itératives, à leur convergence et leurs comparaisons. Là encore, il serait vain de vouloir faire la review de tous les papiers traitant ces sujets. A titre indicatif, nous donnons une chronologie de quelques travaux tout en renvoyant à Ostrowski(1956), Collatz(1950), Bonnet & Meurant(1979), Stoer & Bulrisch(1983), Golub(1989,), Berman & Plemmons(1994) et Varga(2000) pour plus de détails.

Les méthodes classiques de Jacobi et de Gauss-Seidel ont été considérées par différents auteurs dans différentes situations. Ainsi, Mehmke(1892) et von Mises et al.(1929) figurent parmi les premiers à avoir considéré les conditions suffisantes de convergence de ces méthodes prises individuellement ou simultanément. Les résultats concernant les matrices à diagonale strictement dominante ont été démontr-és par Collatz(1942) et autres. Stein & Rosenberg(1948) ont considéré les matrices positives. Le traitement des matrices à inverses positives a débuté avec Stieljes(1887) et Frobenius(1912) mais les concepts de M-matrice et H-matrice sont attribués à Ostrowski (1937,1956) . D'après Varga(2000), un moyen de mesurer l'importance de ce concept est l'existence d'au moins 70 définitions équivalentes de M-matrice dans la litterature dont 50 ont été tabulées dans le livre de Berman & Plemmons(1994).

Pour les méthodes iteratives avec relaxation, semi-iteratives, d'extrapolation et d'inverses généralisés, nous renvoyons aux contributions de Frankel(1950), Young(1950,1971), Ostrowski(1954), Householder(1958), Keller(1958), Golub& Varga (1961), Varga(1959,1979, 2000) et Golub & Van Loan(1989) parmi tant d'autres.

Il faut signaler aussi qu'avec l'avènement du calcul parallèle, les méthodes de relaxation et de décomposition en général, ont connu une nouvelle relance. Voir à ce propos Evans(1984), Plemmons(1986), White(1989) et Golub & Van Loan(1989).

6.9 Exercices

Exercice 6.9.1. Soit B une approximation de l'inverse A^{-1} d'une matrice carrée A et $\hat{x}$ une solution approchée du système $Ax = b$.
En posant $R = I - BA$ et $r = b - A\hat{x}$, montrer que si $\|R\| < 1$ alors on a :

i) $\left\|A^{-1}\right\| \leq \dfrac{\|B\|}{1 - \|R\|}$

ii) $\left\|A^{-1} - B\right\| \leq \dfrac{\|B\|.\|R\|}{1 - \|R\|}$

iii) $\|x - \hat{x}\| \leq \dfrac{\|B\|.\|r\|}{1 - \|R\|}$

Exercice 6.9.2. Soit θ un scalaire non nul et $A(\theta)$ la matrice tridiagonale suivante :

$$A(\theta) = \begin{pmatrix} a_1 & b_1\theta^{-1} & & & & \\ c_2\theta & a_2 & b_2\theta^{-1} & & & \\ & c_3\theta & . & & & \\ & & & . & & \\ & & & & . & b_{n-1}\theta^{-1} \\ & & & & c_n\theta & a_n \end{pmatrix}$$

montrer que $\det A(\theta) = \det A(1)$.

Exercice 6.9.3. Soit E un espace vectoriel normé complet et A un opérateur de $E \longrightarrow E$ vérifiant :

$$\|Ax - Ay\| \leq \alpha\|x - y\| \ \ \forall x \in E, \forall y \in E;$$

avec $0 \leq \alpha < 1$.

1. Soit $x_0 \in E$ donné, montrer que l'itération $x_{n+1} = Ax_n$ définit une suite de Cauchy dans E. On notera x^* la limite de cette suite .
2. On considère une itération approchée donnée par : $y_{n+1} = Ay_n + \sigma_n, \ \ y_0 \in E$ donné et $\|\sigma_n\| \leq \varepsilon$ pour tout $n \geq 0$, prouver que
$$\|x^* - y_n\| \leq \frac{\alpha}{1-\alpha}\|y_n - y_{n-1}\| + \frac{\varepsilon}{1-\alpha}$$
3. Soit $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, M = \dfrac{1}{3}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ et $b(x) = \dfrac{1}{4}\begin{pmatrix} \cos x_1 \\ \sin x_2 \end{pmatrix}$

   a) Montrer que l'équation $x = Mx + b(x)$ admet une solution unique x^*

b) Montrer que si on utilise l'itération $x_{n+1} = Mx_n + b(x_n)$ alors on obtient

$$\|x^* - x_{n+1}\|_\infty \leq k\|x_{n+1} - x_n\|_\infty,$$

où k est une constante à déterminer.

c) Comment peut-on améliorer la vitesse de convergence de l'itération proposée ci-dessus ?

Exercice 6.9.4. On considère les deux matrices suivantes :

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ et } B = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

Chercher les matrices de Jacobi et de Gauss-Seidel associées aux matrices A et B et comparer les rayons spectraux.

Exercice 6.9.5. Soit A une matrice symétrique définie positive et T la matrice définie par : $T = 2D^{-1} - D^{-1}AD^{-1}$. On suppose que $2D - A$ est définie positive.

1. Montrer que la méthode de Jacobi appliquée au système $Ax = b$ converge .
2. Soit la méthode itérative suivante :

$$\begin{cases} x^{(0)} \text{ donné} \\ x^{(n+1)} = x^{(n)} + T\left(b - Ax^{(n)}\right) \end{cases}$$

Montrer que cette méthode converge et comparer sa vitesse de convergence à celle de la la méthode de Jacobi.

Exercice 6.9.6. Soit A une matrice carrée d'ordre n à coefficients dans $\mathbb{C}^n$. Pour chercher la solution du système $Ax = b$ on considère le schéma itératif suivant :

$$(*) \quad \begin{cases} x^{(n+1)} = (I - \alpha A)x^{(n)} + \alpha b \\ x^{(0)} \in \mathbb{C}^n \text{ donné} \end{cases}$$

avec $\alpha > 0$.

I- On suppose que A est à diagonale strictement dominante en lignes.
Montrer que si $0 < \alpha \leq \dfrac{1}{\max\limits_i(a_{ii})}$ alors $(*)$ converge.

II- On suppose que A est hermitienne définie positive dont les valeurs propres sont rangées par ordre décroissant : $0 \leq \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$.

1. Quelle condition doit vérifier α pour que la méthode converge ?
2. Déterminer la valeur optimale de α.

Exercice 6.9.7. Soit $A(h)$ la matrice tridiagonale donnée par :

$$A(h) = \frac{1}{h^2}\begin{pmatrix} 2+h^2a_1 & -1 & & & & \\ -1 & 2+h^2a_2 & -1 & & & \\ & -1 & . & & & \\ & & & . & & \\ & & & & . & -1 \\ & & & & -1 & 2+h^2a_n \end{pmatrix} \quad \text{avec} \quad a_i > 0, i = 1, \cdots, n.$$

1. Montrer que $A(h)$ est définie positive
2. Donner un algorithme de décomposition $A = LU$ où A est donnée par :

$$A = \begin{pmatrix} a_1 & d_2 & & & & \\ c_1 & a_2 & d_3 & & & \\ & c_2 & . & & & \\ & & & . & & \\ & & & & . & d_n \\ & & & & c_{n-1} & a_n \end{pmatrix}\begin{pmatrix} 1 & & & & & \\ l_1 & 1 & & & & \\ & l_2 & . & & & \\ & & & . & & \\ & & & & . & \\ & & & & l_{n-1} & 1 \end{pmatrix}\begin{pmatrix} u_1 & d_2 & & & & \\ & u_2 & d_3 & & & \\ & & . & & & \\ & & & . & & \\ & & & & . & d_n \\ & & & & & u_n \end{pmatrix}$$

en posant : $c_{n+1} = d_1 = l_1 = 0$ et en supposant que $|a_i| > |c_{i+1}| + |d_i|$
Montrer que : $|l_i| < 1$ et $|u_i| > |c_{i+1}|$

Exercice 6.9.8. Soit A la matrice donnée par $A = \begin{pmatrix} a & b & b \\ b & a & b \\ b & b & a \end{pmatrix}$ avec $a > b > .0$

1. Ecrire les matrices de Jacobi et de relaxation associées à la matrice A.
2. Donner une condition sur a et b pour que la méthode de Jacobi soit convergente.
3. On suppose que $a > 2b$, montrer que A est inversible que le conditionnement de A pour la norme $\| \ \|_\infty$vérifie $C(A) \leq \dfrac{\alpha a + \beta b}{\alpha a - \beta b}$ où α et β sont des constantes à déterminer.

Exercice 6.9.9. Soit A une matrice carrée hermitienne et définie positive, pour résoudre le systéme $Ax = b$, on pose $A = D - E - F$, $L = D^{-1}E$ et $U = D^{-1}F$ et on considère le procédé itératif $B(\omega)x^{(k+1)} = (B(\omega) - A)x^{(k)} + b$ avec $B(\omega) = \dfrac{1}{\omega}D(I - \omega L)$.

1. Ecrire l'itération sous la forme $x^{(k+1)} = T(\omega)x^{(k)} + c \quad (*)$ et montrer que $T(\omega) = (I - \omega L)^{-1}((1 - \omega)I + \omega U)$.
2. Montrer que la matrice $B(\omega) + B^H(\omega) - A$ est définie positive pour $0 < \omega < 2$.

3. Montrer que si λ est valeur propre de $Q(\omega) = A^{-1}(2B(\omega) - A)$ alors $Re\lambda > 0$.
4. Vérifier que $Q(\omega) + I$ est inversible et que $(Q(\omega) - I)(Q(\omega) + I)^{-1} = T(\omega)$.
5. Trouver une relation entre les valeurs propres μ de $T(\omega)$ et les λ.
6. En écrivant μ en fonction de λ, prouver que le carré du module de μ peut s'écrire $|\mu|^2 = \dfrac{|\lambda|^2 + \delta - 2Re\lambda}{|\lambda|^2 + \delta + 2Re\lambda}$ où δ est un nombre strictement positif à exprimer.
7. En déduire que l'itération $(*)$ converge pour $0 < \omega < 2$.

Exercice 6.9.10. Examen d'Analyse Numérique **Faculté des sciences Oujda**

Exercice 1. Soit $A = (a_{ij})$ une matrice carrée inversible dont les éléments diagonaux sont non nuls. A est écrite sous la forme $A = D - L - U$ où D est une matrice diagonale et L (respectivement U) est triangulaire inférieure (respectivement supérieure). Pour résoudre le système $Ax = b$, on utilise la méthode itétative suivante :

$$\begin{aligned} a_{ii}x_i^{(k+1)} &= a_{ii}x_i^{(k)} - \theta\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - (1-\theta)\omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} \\ &+ (1-\omega)\theta\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \omega\sum_{j=i}^{n} a_{ij}x_j^{(k)} + \omega b_i \end{aligned}$$

où θ et ω sont des réels fixés (ω non nul) et $k = 0, 1, \cdots$

1. (a) Montrer que la méthode proposée peut s'écrire sous la forme matricielle :

$$x^{(k+1)} = M(\theta, \omega)x^{(k)} + \mathbf{c}(\theta, \omega) \quad (*)$$

   (b) Pour quelles valeurs de θ et ω obtient-on les méthodes de Jacobi et Gauss-Seidel ?

2. On prend $\theta = 1$
   (a) Trouver une relation entre les valeurs propres de $M(\theta, \omega)$ et celles de la matrice de Gauss-Seidel associée à A.
   (b) Peut-on comparer la vitesse de convergence de la méthode $(*)$ à celle de Gauss-Seidel ?
3. On prend $U = L^t$ et $\theta = 0$
   (a) Trouver une relation entre les valeurs propres de $M(\theta, \omega)$ et celles de la matrice de Jacobi associée à A.

(b) En supposant que la méthode de Jacobi converge, montrer que la méthode $(*)$ avec $\theta = 0$ converge pour $0 < \omega < \dfrac{2}{1-\mu_1}$ où $\mu_1$est la plus petite valeur propre de la matrice de Jacobi associée à A.

4. En supposant que A est une matrice tridiagonale, trouver une relation entre les valeurs propres $M(0,\omega)$ et celles de $M(1,\omega)$.

5. En supposant que A est définie positive et symétrique peut-on comparer les vitesses de convergences des méthodes de Jacobi, Gauss-Seidel, $M(0,\omega)$ et $M(1,\omega)$.

Exercice 6.9.11. $\|.\|$ une norme matricielle subordonnée.

1. On considère le système linéaire : $(I+E)x = b$ et le système perturbé

$$(I+E+F)(x+\delta x) = b$$

où I est la matrice carrée identité d'ordre n, E et F deux matrices carrées d'ordre n ; b et x sont des vecteurs à n composantes. On suppose que $\|E\| = \dfrac{1}{2}$ et $\|F\| = \varepsilon < \dfrac{1}{2}$. Montrer que $\|\delta x\| \leq \dfrac{a+b\varepsilon}{c+d\varepsilon}\|b\|$ où a, b, c et d sont des constantes à déterminer.

2. Soit A une matrice diagonalisable admettant $\lambda_1, \lambda_2, \cdots, \lambda_n$ pour valeurs propres, Montrer que si μ est une valeur propre de la matrice perturbée $A + \delta A$ alors $\min\limits_{i=1}^{i=n} |\lambda_i - \mu| \leq C(P)\|\delta A\|$ où $C(P)$ désigne le conditionnement de la matrice P telle que $P^{-1}AP = diag(\lambda_i)$.

Exercice 6.9.12. Soit $A = (a_{ij})$ une matrice carrée inversible dont les éléments diagonaux sont non nuls. A est écrite sous la forme $A = D - L - U$ où D est une matrice diagonale et L (respectivement U) est triangulaire inférieure (respectivement supérieure). Pour résoudre le système $Ax = b$, on utilise la méthode itérative suivante : $a_{ii}x_i^{(k+1)} = a_{ii}x_i^{(k)} + \omega\left(b_i - \sum\limits_{j=1}^{n} a_{ij}x_j^{(k)}\right) + r\sum\limits_{j=1}^{i-1} a_{ij}\left(x_j^{(k)} - x_j^{(k+1)}\right)$, où r et ω sont des réels fixés (ω non nul) et $k = 0, 1, \cdots$

1. Montrer que la méthode proposée peut s'écrire sous la forme matricielle : $x^{(k+1)} = M(r,\omega)x^{(k)} + c$ avec : $M(r,\omega) = (D - rL)^{-1}(aD + bL + eU)$ où a , b sont des réels qu'on exprimera en fonction de r et/ou de ω.

2. Vérifier que cette méthode permet d'obtenir les méthodes de Jacobi, Gauss-Seidel et de relaxation pour des choix appropriés de r et ω.

3. Montrer que les valeurs propres de $M(r,\omega)$ sont les racines de l'équation : $\det(\alpha D - \beta L - \omega U)$ avec $\alpha = \lambda + \omega - 1$ et $\beta = (\lambda - 1)r + \omega$.

4. En supposant que A est une matrice tridiagonale, montrer que les valeurs propres μ de $M(0,1)$ sont liées aux valeurs propres λ de la matrice générale $M(r,\omega)$ par la relation $(\lambda+\omega-1)^2 = \omega\mu^2((\lambda-1)r+\omega)$.
5. Comparer la vitesse de convergence des méthodes classiques en discutant selon les valeurs des paramètres.

Exercice 6.9.13. 1. Soit H une matrice hermitienne définie positive.

a- Montrer que pour tout $r > 0$, la matrice $rI + H$ est inversible.

b- Montrer que la matrice $(rI - H)(rI + H)^{-1}$ est hermitienne et en déduire que : $\left\|(rI - H)(rI + H)^{-1}\right\|_2 = \max_j \left|\frac{r-\lambda_j}{r+\lambda_j}\right|$ où les λ_j sont les valeurs propres de H.

2. Soient H_1 et H_2 deux matrices Hermitiennes définies positives.
Etant donné un vecteur initial $x^{(0)}$ on définit la suite des vecteurs $x^{(k)}$ par :

$$\left(\begin{array}{l} (rI + H_1)x^{(k+\frac{1}{2})} = (rI - H_2)x^{(k)} + b \\ (rI + H_2)x^{(k+1)} = (rI - H_1)x^{(k+\frac{1}{2})} + b \end{array}\right)$$

où b est un vecteur et r un scalaire strictement positif.

a- Ecrire le vecteur $x^{(k+1)}$ sous la forme : $x^{(k+1)} = Tx^{(k)} + c$

b- Montrer que $\rho(T) < 1$ (on pourra considérer la matrice $\widetilde{T} = (rI - H_2)T(rI + H_2)^{-1}$

c- Montrer que la suite $\left(x^{(k)}\right)$ converge vers x^*, solution d'un système linéaire que l'on déterminera.

c- On suppose que les valeurs propres de H_1 et de H_2 sont dans l'intervalle $[a,b], a > 0$, trouver la valeur de r qui rend minimum la quantité :

$$\left\|(rI - H_1)(rI + H_1)^{-1}\right\|_2 \left\|(rI - H_2)(rI + H_2)^{-1}\right\|_2$$

(On pourra montrer que $\min_{r\geq 0} \max_{0<a\leq x\leq b} \left|\frac{r-x}{r+x}\right|^2 = \left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\right)^2$ et qu'il est atteint pour $r = \sqrt{ab}$)

Exercice 6.9.14. Soit A une matrice carrée autoadjointe. On suppose que A est symétrique.

1. En décomposant A sous la forme $A = M - N$, établir les relations suivantes :
   (a) $AM^{-1}N = N - NM^{-1}N$

(b) $\left(M^{-1}N\right)^{\top} AM^{-1}N = AM^{-1}N - \left(I - (M^{-1}N)^{\top}\right) N \left(I - M^{-1}N\right)$

(c) $A = \left(I - (M^{-1}N)^{\top}\right)M^{\top}$

(d) $A - \left(M^{-1}N\right)^{\top} A \left(M^{-1}N\right) = \left(I - M^{-1}N\right)^{\top} \left(M^{\top} + N\right) \left(I - M^{-1}N\right)$

(e) En déduire que si $\mathbf{x}$ est un vecteur propre de $M^{-1}N$ associé à la valeur propre λ alors

$$\left(1 - |\lambda|^2\right) \langle A\mathbf{x}, \mathbf{x}\rangle = |1 - \lambda|^2 \left\langle \left(M^{\top} + N\right)\mathbf{x}, \mathbf{x}\right\rangle$$

et que si $|\lambda| < 1$ alors $\langle Ax, x\rangle$ et$\left\langle \left(M^{\top} + N\right)x, x\right\rangle$ sont de même signe.

2. Théorème d'Ostrowski.
En supposant A définie positive et en la décomposant sous la forme $A = D - L - L^{\top}$

(a) En notant T la matrice de Gauss-Seidel associée à A et en posant $T_1 = D^{1/2}TD^{-1/2}$ et $L_1 = D^{-1/2}LD^{-1/2}$ montrer que $T_1 = (I - L_1)^{-1}L_1^{\top}$

(b) Montrer que T_1 et T ont mêmes valeurs propres.

(c) Soit λ une valeur propre de T_1 associée au vecteur propre x tel que $x^H x = 1$ en posant : $x^H L_1 x = a + ib$, montrer que : $|\lambda|^2 = \dfrac{a^2 + b^2}{1 - 2a + a^2 + b^2}$.

(d) Montrer que $1 - 2a > 0$ et déduire que $\rho(T) < 1$.

(e) Conclure.

Problème 6.9.1. A désigne toujours une matrice carrée d'ordre n.

1. Montrer que si la matrice $A = M - N$ est singulière alors on ne peut pas avoir $\rho(M^{-1}N) < 1$ même si M est régulière.

2. Soit A une matrice hermitienne mise sous la forme $A = M - N$ où M est inversible. on note $B = I - (M^{-1}A)$ la matrice de l'itération associée à $Ax = b$. On suppose que $M + M^* - A$ est définie positive.
Montrer que si x est un vecteur et $y = Bx$ alors :

$$\langle x, Ax\rangle - \langle y, Ay\rangle = \langle x - y, (M + M^* - A)(x - y)\rangle.$$

En déduire que $\rho(B) < 1$ si et seulement si A est définie positive.

3. Soit $a \in \mathbb{R}$; la matrice A est définie de la manière suivante :

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

(a) Pour quelles valeurs de a de la matrice A est-elle définie positive. Peut-on en déduire les valeurs de a pour lesquelles la méthode de Gauss-Seidel est convergente.

(b) Ecrire la matrice J de l'itération de Jacobi. Pour quelles valeurs de a la matrice de Jacobi converge t-elle.

(c) Ecrire la matrice G de l'itération de Gauss-Seidel. Calculer $\rho(G)$. Pour quelles valeurs de a la matrice de gauss-Seidel converge t-elle plus vite que la méthode de Jacobi.

4. On donne deux matrices réelles régulières A et B et $a, b \in \mathbb{R}^n$.

(a) On construit les deux itérations suivantes :

$$\begin{cases} x_0, y_0 \in \mathbb{R}^n \\ x_{k+1} = Bx_k + a \\ y_{k+1} = Ay_k + b \end{cases} \tag{6.9.1}$$

Donner une condition nécessaire et suffisante de convergence des deux suites x_k et y_k.

(b) On pose $z_k = (x_k, y_k)^\top$. Expliciter les matrices C et c telles que $z_{k+1} = Cz_k + c$ et comparer $\rho(C)$ et $\rho(AB)$.
On considère les deux itérations

$$\begin{cases} x_{k+1} = By_k + a \\ y_{k+1} = Ax_{k+1} + b \end{cases} \tag{6.9.2}$$

(c) Donner une condition nécessaire et suffisante de convergence de (6.9.2).

(d) Mettre (6.9.2) sous la forme $z_{k+1} = Dz_k + d$. Expliciter D et d et comparer $\rho(D)$ et $\rho(AB)$.

(e) Comparer les taux de convergence des algorithmes (6.9.1)et (6.9.2).

Chapitre 7

Analyse numérique des équations differentielles ordinaires (e.d.o)

7.1 Rappels sur les équations differentielles ordinaires (e.d.o)

On considère l'e.d.o du premier ordre

$$y^{'}(x) = f(x,y),\ \ f : \mathbb{R} \times \mathbb{R}^m \longrightarrow \mathbb{R}^m, x \in \mathbb{R},\ \ y \in \mathbb{R}^m \text{ où } \ \ y = (y_1, y_2, \cdots, y_m)^\top \tag{7.1.1}$$

L'équation (7.1.1) peut encore s'écrire sous la forme d'un système d'e.d.o :

$$\begin{array}{rcl} y_1^{'} & = & f_1(x, y_1, \cdots, y_m) \\ y_2^{'} & = & f_2(x, y_1, \cdots, y_m) \\ \vdots & \vdots & \vdots \\ y_m^{'} & = & f_m(x, y_1, \cdots, y_m) \end{array} \tag{7.1.2}$$

Lorsque les conditions initiales sont précisées, on obtient un problème de condition initiale (p.c.i) encore appelé problème de Cauchy :

$$\begin{array}{rcl} y^{'}(x) & = & f(x,y) \\ y(a) & = & \alpha \ \text{ avec } \ \alpha = (\alpha_1, \cdots, \alpha_m)^\top \text{ donné} \end{array} \tag{7.1.3}$$

L'existence et l'unicité de la solution du p.c.i (7.1.3) sont données par le théorème suivant

Théorème 7.1.1. *Soit $f : \mathbb{R} \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$ une fonction définie et continue pour tout couple $(x,y) \in D$ où $D = \{(x,y); a \leq x \leq b, -\infty < y_i < \infty\}$, avec a et b finis. On suppose qu'il existe une constante L telle que :*

$$\|f(x,y) - f(x,y^*)\| \leq L\|y - y^*\| \text{ pour tous } (x,y) \text{ et } (x,y^*) \text{ appartenant à } D \tag{7.1.4}$$

Alors pour tout $\alpha \in \mathbb{R}^m$, il existe une solution unique $y(x)$ du problème (7.1.3), où y est continue differentiable pour tout couple $(x, y) \in D$.

Remarque 7.1.1. Un système differentiel d'ordre q peut toujours être ramené à un système du premier ordre du type (7.1.3)

$$y^{(q)} = \varphi\left(x, y^{(0)}, \cdots, y^{(q-1)}\right);\ \varphi : \mathbb{R} \times \mathbb{R}^m \times \cdots \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$$

En posant : $Y_1 = y, Y_2 = Y_1^{'}(= y^{'}), \cdots, Y_q = Y_{q-1}^{'}(= y^{(q-1)})$ on obtient
$\mathbf{Y}^{'} = F(x, \mathbf{Y})$ avec $\mathbf{Y} = \left(\mathbf{Y}_1^{\top}, \mathbf{Y}_2^{\top}, \cdots, \mathbf{Y}_q^{\top}\right)^{\top} \in \mathbb{R}^{qm}$

$$F = \left(\mathbf{Y}_2^{\top}, \mathbf{Y}_3^{\top}, \cdots, \mathbf{Y}_q^{\top}, \varphi^{\top}\right)^{\top} \in \mathbb{R}^{qm}$$

où $y^{(r)}(a) = \alpha_{r+1}, r = 0, 1, \cdots, q-1$.
On obtient

$$\mathbf{Y}^{'} = F(x, \mathbf{Y}); \mathbf{Y}(\mathbf{a}) = \alpha$$

Exemple 7.1.1.

$$y^{(iv)} = f(x, y), \quad f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$$

$$Y_1 = y, Y_2 = Y_1^{'} = y^{'}, Y_3 = Y_2^{'} = y^{''}, Y_4 = Y_3^{'} = y^{'''}, Y_4^{'} = y^{(iv)}$$

si on pose $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_4)^{\top}$ on obtient $\mathbf{Y} = \mathbf{F}(x, \mathbf{Y})$.

7.2 Systèmes linéaires

Le système

$$y^{'}(x) = f(x, y), \quad f : \mathbb{R} \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$$

est dit linéaire si f est de la forme

$$f(x, y) = A(x)y + \psi(x) \tag{7.2.1}$$

où $A(x)$ est une matrice d'ordre m et $\psi \in \mathbb{R}^m$.

Si de plus $A(x) = A$ est indépendante de x, on obtient un système linéaire à coefficients constants de la forme :

$$y^{'}(x) = Ay + \psi(x) \tag{7.2.2}$$

Si $\psi(x) = 0$, le système est dit homogène.

$$y^{'} = Ay \tag{7.2.3}$$

Si les valeurs propres de A sont distinctes, la solution du système homogène est de la forme

$$y(x) = \sum_{j=1}^{k} C_j \exp(\lambda_j x) V_j + \varphi(x) \tag{7.2.4}$$

où λ_j est valeur propre de A associée au vecteur propre V_j et les C_j sont des constantes arbitraires et $\varphi(x)$ est une solution particulière de l'équation (7.2.2).

Exemple 7.2.1.

$$y' = Ay + \psi(x) \ , y(0) = \alpha$$

avec $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, $\psi(x) = (2x-1, x+1)^\top$, $\alpha = (0,0)^\top$, les valeurs propres de A sont $\lambda_1 = 3$ et $\lambda_2 = -1$.
Les vecteurs propres de A sont $V_1 = (1,1)^\top$ et $V_2 = (1,-1)^\top$.

En cherchant une solution particulière de la forme : $\varphi(x) = \begin{pmatrix} ax+b \\ cx+d \end{pmatrix}$, on obtient

$$y(x) = C_1 \exp(3x) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2 \exp(-x) \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -5/3 \\ -x + 4/3 \end{pmatrix}.$$

Enfin, en considérant la condition initiale on obtient $C_1 = 1/6$ et $C_2 = 3/2$,

$$\text{d'où } y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix} = \begin{pmatrix} \frac{1}{6}\exp(3x) + \frac{3}{2}\exp(-x) - \frac{5}{3} \\ \frac{1}{6}\exp(3x) - \frac{3}{2}\exp(-x) - x + \frac{4}{3} \end{pmatrix}.$$

7.3 Notions de stabilité

Considérons le système nonlinéaire suivant

$$\frac{dX(t)}{dt} = F(X(t)) \quad (**)$$

où $X(t) = (X_1, \cdots, X_n)^\top$ est un vecteur de $\mathbb{R}^n$ et $F = (f_1, \cdots, f_n)^\top$ une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$ suffisamment régulière.
Si F est lineaire le système est dit *linéaire*

Définition 7.3.1 (Point d'équilibre)**.** Un vecteur $\bar{X}$ est un point équilibre du système $(**)$ si à l' instant t_0 l'état du système est égal à $\bar{X}$ et il restera égal à $\bar{X}$ dans le futur c.à.d :

$$X(t_0) = \bar{X} \ \text{ et } \ \forall \ t > t_0 \ \ X(t) = \bar{X}.$$

On parle aussi de point stationnaire, état stationnaire et solution stationnaire qui désignent la même notion.

Définition 7.3.2.

1. Un point d'équilibre $\bar{X}$ est dit stable si :
   $\exists R_0 > 0$ et $\forall R < R_0 \quad \exists r \quad 0 < r < R$ tel que si $X(t_0) \in B(\bar{X}, r)$ alors
   $X(t) \in B(\bar{X}, R) \quad \forall \quad t > t_0, B(\bar{X}, r)$:boule de centre $\bar{X}$ et de rayon r ;
2. Un point d'équilibre est dit asymptotiquement stable s'il est stable et en plus $\exists R_0$ tel que pour tout $X(t_0) \in B(\bar{X}, R_0)$ on a $\lim_{t \longrightarrow +\infty} X(t) = \bar{X}$;
3. Un point d'équilibre est dit marginalement stable s' il est stable et non asymptotiquement stable ;
4. Un point est instable si il n' est pas stable ;
5. Un point est globalement asymptotiquement stable si pour tout $X(t_0)$ on a $\lim_{t \longrightarrow +\infty} X(t) = \bar{X}$.

Théorème 7.3.1. *Si F est linéaire, une condition nécessaire et suffisante pour que le système $(**)$ admette 0 comme point d'équilibre asymptotiquement stable est que $Re(\lambda) < 0$, pour toute valeur propre λ de F.*
Si au moins une des valeurs propres de F vérifie $Re(\lambda) > 0$ alors le point d'équilibre 0 est instable .

Définition 7.3.3. Si le système non linéaire $(**)$ admet un point d'équilibre $\bar{X}$, on appelle matrice de linéarisation du système la matrice

$$\mathcal{A} = \begin{pmatrix} \frac{\partial f_1}{\partial X_1} & \cdots & \frac{\partial f_1}{\partial X_n} \\ \frac{\partial f_2}{\partial X_1} & \cdots & \frac{\partial f_2}{\partial X_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial X_1} & \cdots & \frac{\partial f_n}{\partial X_n} \end{pmatrix}_{\bar{X}},$$

le système $\frac{dY(t)}{dt} = \mathcal{A}Y(t)$ est dit système linearisé obtenu à partir du système $(**)$.

Théorème 7.3.2. *Si le système non linéaire $(**)$ admet l'origine comme point fixe unique alors dans un voisinage de l'origine, les comportements du système non linéaire et du système linearisé sont équivalents à condition que le système n'admette pas de point centre (valeurs propres imaginaires pures).*

Remarque 7.3.1. Le théorème peut s'appliquer aux points d'équilibre qui sont distincts de l'origine, il suffit d'introduire des coordonnées locales.

Théorème 7.3.3 (Théorème de Liapunov). *Soient $\bar{X}$ un point d'équilibre de $(**)$ et V une fonction définie de U dans $\mathbb{R}$ de classe C^1, U un voisinage de $\bar{X}$ tel que :*

i) $V(\bar{X}) = 0$ *et* $V(X) > 0$ *si* $X \neq \bar{X}$,

ii) $\dot{V}(X) \leq 0 \quad \forall X \in U \setminus \{\bar{X}\}$.

Alors $\bar{X}$ est stable. Si de plus

iii) $\dot{V}(X) < 0 \quad \forall X \in U \setminus \{\bar{X}\}$

Alors $\bar{X}$ est asymptotiquement stable. $V(X)$ est dite fonction de Liapunov.

Pour les preuves des théorèmes 7.3.1, 7.3.2 et 7.3.3, voir ([**?**, **?**, **?**]).

7.4 Système d'équations aux differences linéaires avec coéfficients constants

Soit $\{y_n, n = N_0, N_0 + 1, \cdots, \}$ une suite de vecteurs de $\mathbb{R}^m$ et (α_n) une suite de réels. On appelle système d'équations aux differences linéaires d'ordre k le système d'équations suivantes

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \psi_n, n = N_0, N_0 + 1, \cdots \text{ avec} \qquad \psi_n \in \mathbb{R}^m \tag{7.4.1}$$

Si de plus $\psi_n = 0$, le système est dit homogène :

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0, n = N_0, N_0 + 1, \cdots \tag{7.4.2}$$

La solution générale du système (7.4.1) s'obtient de façon similaire à celle qui donne la solution du système d'équations differentielles linéaires (7.2.1). Elle est donnée par la somme d'une solution (Y_n) du système homogène (7.4.2) et d'une solution particulière φ_n du système (7.4.1) $(y_n = Y_n + \varphi_n)$.

Définition 7.4.1. On appelle polynôme caractéristique de l'équation aux differences linéaires le polynôme défini par $\pi(r) = \sum_{j=0}^{k} \alpha_j r^j$. Si les racines r_j de $\pi(r)$ sont simples alors la solution générale du système (7.4.1) est donnée sous la forme $y_n = \sum_{j=1}^{k} \theta_j r_j^n + \varphi_n$, où φ_n est une solution particulière du système (7.4.1).

Si r_1 est une racine de $\pi(r)$ de multiplicité m, la solution générale du système (7.4.1) est donnée sous la forme : $y_n = \sum_{j=1}^{m} n^{j-1} \theta_j r_1^n + \sum_{j=m+1}^{k} \theta_j r_j^n + \varphi_n$.

Exemple 7.4.1. $y_{n+2} - y_{n+1} + \frac{1}{2}y_n = 1$, $K = 2$ est une solution particulière

$$\pi(r) = r^2 - r + \frac{1}{2},\ \ \Delta = 1 - 2 = i^2.$$

Les racines sont complexes conjugués $r_1 = (1+i)/2$ et $r_2 = (1-i)/2$. La solution générale est donnée par $y_n = \theta_1(1+i)^n/2^n + \theta_2(1-i)^n/2^n + 2$.

7.5 Méthodes numériques pour les problèmes de condition initiale

Considérons le problème de condition initiale

$$y^{'}(x) = f(x,y), \quad y(a) = \alpha. \tag{7.5.1}$$

On suppose que ce problème admet une solution unique dans l'intervalle $[a,b]$.

Les méthodes numériques utilisent la discretisation de l'intervalle $[a,b]$ en posant

$$x_i = a + ih,\ \ i = 0, 1, \cdots, N$$

où $h = \dfrac{b-a}{N}$ est le pas de discrétisation (ici le pas est supposé constant mais on peut envisager des pas h_i variables).

La solution exacte au point x_i est notée $y(x_i)$, la solution approchée est notée y_i ($y(x_i) \simeq y_i$), une méthode numérique est un système d'équations aux differences impliquant un certain nombre d'approximations successives $y_n, y_{n+1}, \cdots, y_{n+k}$ où k désigne le nombre de pas de la méthode.

Si $k = 1$ on parle de méthode à un pas.

Si $k > 1$ la méthode est dite à pas multiples ou multi-pas.

Exemple 7.5.1.

$$y_{n+1} - y_n = hf_n$$

$$y_{n+2} + y_{n+1} - 2y_n = \frac{h}{4}(f_{n+2} + 8f_{n+1} + 3f_n)$$

$$y_{n+2} - y_{n+1} = \frac{h}{3}(3f_{n+1} - 2f_n)$$

$$y_{n+1} - y_n = \frac{h}{4}(k_1 + k_2)$$

avec

$$k_1 = f_n = f(x_n, y_n) \ \text{ et } \ k_2 = f\left(x_n + h, y_n + \frac{1}{2}hk_1 + \frac{1}{2}hk_2\right).$$

Ces exemples peuvent être donnés sous la forme générale

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \phi_f(y_{n+k}, \cdots y_n, x_n; h). \tag{7.5.2}$$

7.5.1 Convergence

Considérons une méthode numérique de type (7.5.2) avec des valeurs initiales appropriées :

$$\left\{ \begin{array}{l} \sum_{j=0}^{k} \alpha_j y_{n+j} = \phi_f(y_{n+k}, \cdots y_n, x_n; h) \\ y_\varkappa = \gamma_\varkappa(h)\ ,\ \varkappa = 0, 1, \cdots, k-1 \end{array} \right\} \tag{7.5.3}$$

Définition 7.5.1. La méthode (7.5.3) est dite convergente si pour tout problème de condition initiale vérifiant les hypothèses du théorème 7.1.1 (existence et unicité) on a

$$\max_{0\leq n\leq N} \|y(x_n) - y_n\| = 0 \quad \text{quand } h \longrightarrow 0$$

Définition 7.5.2. On appelle erreur de troncature locale de la méthode numérique le résidu $R_{n+k}(h)$ défini par

$$R_{n+k}(h) = \sum_{j=0}^{k} \alpha_j y(x_{n+j}) - h\phi_f(y(x_{n+k}), y(x_{n+k-1}), \cdots, y(x_n), x_n, h) \tag{7.5.4}$$

Remarque 7.5.1. Parfois on utilise aussi $\tau_n(h) = \frac{1}{h} R_{n+k}(h)$ comme définition d'erreur de troncature locale, mais dans le cadre de ce chapitre c'est la première définition qui est adoptée.

Définition 7.5.3. La méthode (7.5.3) est dite d'ordre p si $R_{n+k} = O\left(h^{p+1}\right)$.

7.5.2 Consistance

La méthode numérique est dite consistante si son ordre est au moins 1 ou encore, pour tout p.c.i vérifiant les hypothèses du théorème 7.1.1 d'existence et d'unicité on a

$$\lim_{h\longrightarrow 0} \frac{1}{h} R_{n+k}(h) = 0, \quad x = a + nh.$$

Lemme 7.5.1. *La méthode numérique (7.5.2) est consistante si et seulement si*

i) $\sum_{j=0}^{k} \alpha_j = 0.$

ii) $(\phi_f(y(x_n), y(x_n) \cdots y(x_n), x_n, 0))/\sum_{j=0}^{k} j\alpha_j = f(x_n, y(x_n)).$

7.5.3 Stabilité

Considérons le problème de condition initiale

$$z' = f(x,z),\ \ z(a) = \alpha,\ \ x \in [a,b]$$

Avant de définir la stabilité de la méthode numérique, on pourrait se poser la question de savoir comment réagirait la solution $z(x)$ de ce problème à une perturbation des conditions initiales et/ou de f ? Soit donc z^* la solution du problème perturbé :

$$z' = f(x,z) + \delta(x),\ \ z(a) = \alpha + \delta,\ \ x \in [a,b]\,.$$

Définition 7.5.4 (Hahn,Stetter)**.** Si ($\delta(x)$, δ) et ($\delta^*(x)$, δ^*) sont deux perturbations et $z(x)$ et $z^*(x)$ les solutions qui en résultent et s'il existe une constante S telle que

$$\forall x \in [a,b]\,, \|z(x) - z^*(x)\| < S\varepsilon \text{ si } \|\delta(x) - \delta^*(x)\| < \varepsilon \text{ et } \|\delta - \delta^*\| < \varepsilon$$

Alors le p.c.i est dit totalement stable.

Remarque 7.5.2. Dans cette définition de stabilité, on exige simplement l'existence d'une constante S finie (mais pas nécessairement petite) et on montre que les hypothèses du théorème 7.1.1 sont suffisantes pour que le p.c.i soit totalement stable. On dit aussi que le problème est bien posé.

Si maintenant on considère la méthode numérique, on peut se demander quel effet aurait une perturbation de l'équations aux differences sur la solution numérique y_n. On a alors la définition suivante :

Définition 7.5.5 (Lambert)**.** Soient $(\delta_n, n = 0,1,\cdots N)$ et $(\delta_n^*, n = 0,1,\cdots N)$ deux perturbations de l'équation aux differences (7.1.1)et $(z_n, n = 0,1,\cdots N)$ et $(z_n^*, n = 0,1,\cdots N)$ les solutions qui en résultent. On dira que la méthode numérique est zéro-stable si il existe deux constantes S et h_0 telles que pour tout $h \in [0,h_0]$, si $\|\delta_n - \delta_n^*\| < \varepsilon\ 0 \le n \le N$ alors $\|z_n - z_n^*\| < S\varepsilon, \quad 0 \le n \le N$.

Remarque 7.5.3. La zéro-stabilité est encore appelée stabilité au sens de Dahlquist.

Définition 7.5.6. On appelle 1^{er} polynôme caractéristique de la méthode numérique le polynôme $\rho(t)$ défini par $\rho(t) = \sum_{j=0}^{k} \alpha_j t^j$.

Définition 7.5.7. On dit que la méthode numérique satisfait les conditions aux racines si tous les zéros du 1^{er} polynôme catactérisitique sont de module inférieur ou égal à 1 et ceux ayant un module égal à 1 sont des zéros simples.

Théorème 7.5.1. *Une condition nécessaire et suffisante pour que la méthode soit zéro-stable et que la méthode vérifie la condition aux racines.*

Théorème 7.5.2. *Une condition suffisante et nécessaire pour que la méthode (7.5.2) soit convergente est qu'elle soit zéro- stable et consistante.*

Pour les preuves des théorèmes 7.5.1 et 7.5.2, voir ([**?**, **?**]).

7.5.4 Méthode d'Euler

La plus simple et la plus connue des méthodes d'approximation des solutions des e.d.o est la méthode d'Euler donnée par :

$$y_{n+1} - y_n = hf_n, \quad y_0 = \alpha \tag{7.5.5}$$

En considérant

$$y(x_n + h) - y(x_n) - hf(x_n, y(x_n)) = \frac{1}{2}h^2 y''(\theta_n) \text{ avec } x_n \leq \theta_n \leq x_{n+1} \tag{7.5.6}$$

On voit que la méthode d'Euler est une méthode explicite d'ordre 1 (donc consistante).

Lemme 7.5.2. *i) Pour tout $x \geq -1$ et toute constante positive m on a*

$$0 \leq (1+x)^m \leq \exp(mx)$$

ii) Si s et t sont des réels positifs et $(z_n)_{n=0}^{n=k}$ une suite vérifiant

$$z_0 \geq -\frac{t}{s} \text{ et } z_{n+1} \leq (1+s)z_n + t \quad \forall n = 1, \cdots, k$$

Alors on a

$$z_{n+1} \leq \exp((n+1)s)\left(\frac{t}{s} + z_0\right) - \frac{t}{s}.$$

Théorème 7.5.3. *Soit $f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue et vérifiant la condition de Lipschitz pour y sur $D = \{(x, y); a \leq x \leq b, -\infty < y < \infty\}$ avec a et b finis. On suppose qu'il existe une constante M telle que $|y''(x)| \leq M$ pour tout $x \in [a, b]$. Soit $y(x)$ la solution unique du p.c.i $y'(x) = f(x, y)$, $y(a) = \alpha$ et $(y_n)_{n=0}^{n=N}$ la suite des approximations générée par la méthode d'Euler. Alors on a :*

$$|y(x_n) - y_n| \leq \frac{hM}{2L}(\exp(L(x_n - a)) - 1) \text{ pour chaque } n = 0, 1, \cdots, N$$

Preuve.

Pour $n = 0$, l'inégalité est vérifiée puisque $y(x_0) = y_0 = \alpha$.

Les équations (7.5.5) et (7.5.6) donnent :

$$|y(x_{n+1}) - y_{n+1}| \leq |y(x_n) - y_n| + h|f(x_n, y(x_n) - f(x_n, y_n)| + \frac{1}{2}h^2|y''(\theta_n)| \quad (7.5.7)$$

Les hypothèses du théorème conduisent alors à la mojoration

$$|y(x_{n+1}) - y_{n+1}| \leq |y(x_n) - y_n|(1 + hL) + \frac{h^2M}{2}$$

En appliquant le lemme avec $z_n = |y(x_n) - y_n|, s = hL$ et $t = \frac{h^2M}{2}$ il vient :

$$|y(x_{n+1}) - y_{n+1}| \leq \exp(((n+1)hL)\left(|y(x_0) - y_0| + \frac{h^2M}{2hL}\right) - \frac{h^2M}{2hL}$$

et comme $|y(x_0) - y_0| = 0$ et $(n+1)h = x_{n+1} - a$, on obtient :

$$|y(x_{n+1}) - y_{n+1}| \leq \frac{hM}{2L}(\exp(L(x_{n+1} - a)) - 1)$$

D'après le théorème 7.5.3, l'erreur de la méthode d'Euler est majorée par une fonction linéaire en h. Ceci laisse comprendre que plus on diminue h plus on réduit l'erreur. Cependant, le corollaire qui va suivre indique autre chose.

Corollaire 7.5.1. *Soit $y(x)$ la solution unique du p.c.i $y'(x) = f(x, y)$, $a \leq x \leq b$, $y(a) = \alpha$ et $w_0, w_1, \cdots, w_N$ les approximations vérifiant $w_0 = \alpha + \delta_0$ et $w_{n+1} = w_n + hf(x_n, w_n) + \delta_{n+1}; n = 0, \cdots, N-1$. Si $|\delta_n| < \delta \ \forall n = 0, \cdots, N$ et les hypothèses du théorème précedent sont vérifiées, alors*

$$|y(x_n) - w_n| \leq \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)(\exp(L(x_n - a)) - 1) + |\delta_0| \exp L(x_n - a)$$

Preuve. Identique à celle du théorème avec $y(x_0) - w_0 = \delta_0$ et $t = \frac{h^2M}{2} + |\delta_n|$.

Remarque 7.5.4. La simplicité de la méthode d'Euler en fait un exemple pédagogique d'introduction aux autres méthodes plus complexes. Cependant, un des critères principaux de l'analyse numérique consiste à chercher des méthodes ayant l'ordre de précision le plus élevé possible et comme la méthode d'Euler est d'ordre 1, son utilisation se trouve limitée en pratique et on est amené à considérer des méthodes plus précises. Trois directions principales permettent d'obtenir des méthodes d'ordres élevés.

La première direction consiste à travailler avec des méthodes à un pas mais en cherchant à atteindre des ordres élevés en utilisant un développement de Taylor

et en négligeant le terme d'erreur mais ces méthodes ont un handicap à cause des dérivées susccessives de f.

Une deuxième possibilité est donnée par des choix appropriés de $\phi_f(y_{n+k}, \cdots y_n, x_n; h)$ dans l'équation (7.5.1), les méthodes de Runge-Kutta sont la meilleure illustration de cette direction.

Enfin, une troisième direction est offerte par les Méthodes Linéaires à Pas Multiples (MLPM).

En se basant sur le critère de précision, on voit qu'on est obligé de chercher des méthodes dont la performance est supérieure à celle d'Euler. On fait donc appel à d'autres méthodes plus précises.

7.5.5 Méthodes de Taylor dans le cas scalaire

Supposons que la solution $y(x)$ du p.c.i

$$y^{'}(x) = f(x, y), \quad a \leq x \leq b, \quad y(a) = \alpha \tag{7.5.8}$$

est de classe $C^{(n+1)}$. Alors en écrivant le développement de Taylor au point $x_{n+1} = x_n + h$ on obtient

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(x_n) + \cdots + \frac{h^n}{n!}y^{(n)}(x_n) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\zeta_n),$$

$$x_n < \zeta_n < x_{n+1}$$

En remplaçant $y^{'}(x_n)$ par $f(x_n, y_n)$ ainsi que les dérivées supérieures de $y^{'}$ par celles de f puis en laissant tomber le terme $\frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\zeta_n)$, on obtient la méthode numérique :

$$\left\{ \begin{array}{c} y_0 = \alpha \\ y_{n+1} = y_n + hT_f^n(x_n, y_n, h), \\ T_f^n(x_n, y_n, h) = f(x_n, y_n) + \frac{h}{2!}f'(x_n, y_n) + \cdots + \frac{h^{n-1}}{n!}f^{(n-1)}(x_n, y_n) \end{array} \quad n = 0, 1, \cdots, N-1 \right\} \tag{7.5.9}$$

Exemple 7.5.2. La méthode d'Euler fait partie des méthodes de Taylor.

Exercice 7.5.1. En faisant un développement de Taylor de $y(x)$ à l'ordre 3 puis en remplaçant $y''(x_n)$ par $\frac{1}{h}(y'(x_{n+1}) - y'(x_n)) + O(h)$, montrer qu'on obtient la méthode d'Euler modifiée $y_{n+1} - y_n = \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_{n+1}))$.

Remarque 7.5.5. Au vu du critère de précision et bien que les méthodes de Taylor paraissent faciles dans leur écriture, elles sont rarement utilisées dans la pratique à cause des difficultés engendrées par le calcul des dérivées successives de f comme fonction de deux variables. C'est pour cette raison qu'on cherche des méthodes permettant d'atteindre un ordre élevé tout en évitant le calcul des dérivées successives de f. Au critère de précision s'ajoute le critère de coût.

7.5.6 Méthodes de Runge-Kutta (R.K) dans le cas scalaire

Revenons au p.c.i (7.1.3), les méthodes de R.K se présentent sous la forme :

$$y_{n+1} - y_n = h \sum_{i=1}^{l} b_i k_i$$

avec $k_i = f\left(x_n + c_i h, y_n + h \sum_{j=1}^{l} a_{ij} k_j\right), i, = 1, 2, \cdots, l$ et on suppose que $c_i = \sum_{j=1}^{l} a_{ij}, i, = 1, 2, \cdots, l$

Si $a_{ij} = 0$ pour $j > i$ alors :

$$k_i = f\left(x_n + c_i h, y_n + h \sum_{j=1}^{i} a_{ij} k_j\right) \qquad i, = 1, 2, \cdots, l$$

on obtient ainsi :

$$k_1 = f(x_n, y_n)$$

$$k_2 = f(x_n + c_2 h, y_n + c_2 h k_1)$$

$$k_3 = f(x_n + c_3 h, y_n + (c_3 - a_{32}) h k_1 + a_{32} h k_2)$$

Remarques 7.5.6.

1. Les méthodes de R.K sont des méthodes à un pas, elles peuvent-être écrite sous la forme générale $y_{n+1} - y_n = h\phi_f(x_n, y_n, h)$, où $\phi_f(x, y, h) = \sum_{i=1}^{l} b_i k_i$.
2. Les méthodes de R.K satisfont la condition aux racines, elles sont donc zéro-stables. Par conséquent, pour étudier la convergence il suffit d'étudier la consistance.

7.5.7 Méthodes de Runge-Kutta explicites

Les méthodes explicites de R-K d'ordre 1,2 ,3 et 4 sont obtenues en considérant

$$y_{n+1} = y_n + h(b_1 k_1 + b_2 k_2 + b_3 k_3 + b_4 k_4)$$

en déterminant pour chaque ordre les valeurs possibles des paramètres.

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + c_2 h, y_n + c_2 h k_1) \\ k_3 &= f(x_n + c_3 h, y_n + (c_3 - a_{32}) h k_1 + a_{32} h k_2) \\ \vdots &\ \vdots\ \ \vdots \end{aligned}$$

Les formes explicites des méthodes d'ordre 1,2 et 3 sont obtenues aprés differentiation et identification.

Méthode d'ordre 1

En considérant $y_{n+1} = y_n + h b_1 k_1$ avec $k_1 = f(x_n, y_n)$ on voit que le l'odre maximal est 1 obtenu avec le paramètre b_1 égal à 1 et qui donne la méthode d'Euler :

$$y_{n+1} - y_n = h f(x_n, y_n)$$

Méthodes d'ordre 2

En partant de $y_{n+1} = y_n + h b_1 k_1 + b_2 k_2$ avec $k_1 = f(x_n, y_n)$ et
$k_2 = f(x_n + c_2 h, y_n + c_2 k_1)$, $p = 2$ est l'ordre maximal possible qu'on peut atteindre si les paramètres b_1, b_2 et $c_2$vérifient les équations $b_1 + b_2 = 1$ et $b_2 c_2 = \frac{1}{2}$. Comme on a deux équations pour trois inconnues, il en résulte une infinité de méthodes explicites d'ordre 2 . On en donne quelques unes :

Méthode d'Euler modifiée : ($b_1 = 0, b_2 = 1$ et $c_2 = \frac{1}{2}$)

$$y_{n+1} - y_n = h f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right)$$

Méthode d'Euler améliorée : ($b_1 = b_2 = \frac{1}{2}$ et $c_2 = 1$)

$$y_{n+1} - y_n = \frac{h}{2}(f(x_n, y_n) + f(x_n + h, y_n + k_1))$$

Méthode de Heun d'ordre 2 ($b_1 = \frac{1}{4}$, $b_2 = \frac{3}{4}$ et $c_2 = \frac{2}{3}$)

$$y_{n+1} = y_n + \frac{h}{4}\left(f(x_n, y_n) + 3f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}k_1)\right)$$

$$k_1 = h f(x_n, y_n)$$

Méthodes d'ordre 3

En considérant $y_{n+1} = y_n + b_1 k_1 + b_2 k_2 + b_3 k_3$ avec $k_1 = h f(x_n, y_n)$,
$k_2 = h f(x_n + c_2 h, y_n + c_2 h k_1)$ et $k_3 = h f(x_n + c_3 h, y_n + (c_3 - a_{32}) k_1 + a_{32} k_2)$.

On obtient une famille de méthode d'ordre 3 si les paramètres vérifient les équations

$$\begin{aligned} b_1 + b_2 + b_3 &= 1 \\ b_2 c_2 + b_3 c_3 &= \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 &= \frac{1}{3} \\ b_3 c_2 a_{32} &= \frac{1}{6} \end{aligned}$$

Deux représentants de cette famille de méthode d'ordre 3 sont :

Méthode de Heun d'ordre 3 ($b_1 = \frac{1}{4}$, $b_2 = 0, b_3 = \frac{3}{4}$ et $c_2 = \frac{1}{3}$ et $c_3 = \frac{2}{3}, a_{32} = \frac{2}{3}$)

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}(k_1 + 3k_3) \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{3}h, y_n + \frac{1}{3}k_1\right) \\ k_3 &= hf\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}k_2\right) \end{aligned}$$

Méthode de Kutta d'ordre 3 ($b_1 = \frac{1}{6}$, $b_2 = \frac{2}{3}, b_3 = \frac{1}{6}$ et $c_2 = \frac{1}{2}$et $c_3 = 1, a_{32} = 2$)

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right) \\ k_3 &= hf(x_n + h, y_n - k_1 + 2k_2) \end{aligned}$$

Méthode d'ordre 4

Méthode de Runge-Kutta d'ordre 4

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right) \\ k_3 &= hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2\right) \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned}$$

Méthode de Runge-Kutta d'ordre 4

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{8}(k_1 + 3k_2 + 3k_3 + k_4) \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{3}h, y_n + \frac{1}{3}k_1\right) \\ k_3 &= hf\left(x_n + \frac{2}{3}h, y_n - \frac{1}{3}k_1 + k_2\right) \\ k_4 &= hf(x_n + h, y_n + k_1 - k_2 + k_3) \end{aligned}$$

Méthodes de Runge-Kutta d'ordre supérieur et contrôle de l'erreur

Des méthodes de R-K d'ordres successifs peuvent être combinées pour le contrôle de l'erreur. Deux types de cette utilisation sont illustrées **[?]** par :

1. Méthode de Runge-Kutta-Fehlberg. Elle consiste à utiliser une méthode de R-K d'ordre 5

   $$\widetilde{y}_{n+1} = y_n + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6$$

   pour estimer l'erreur de troncature locale de la méthde de R-K d'ordre 4

   $$y_{n+1} = y_n + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5$$

   avec

   $$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{4}h, y_n + \frac{1}{4}k_1\right) \\ k_3 &= hf\left(x_n + \frac{3}{8}h, y_n + \frac{3}{32}k_1 + \frac{9}{32}k_2\right) \\ k_4 &= hf\left(x_n + \frac{12}{13}h, y_n + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right) \\ k_5 &= hf\left(x_n + h, y_n + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right) \\ k_6 &= hf\left(x_n + \frac{1}{2}h, y_n - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right) \end{aligned}$$

2. Méthode de Runge-Kutta-Verner. Elle consiste à utiliser une méthode de R-K d'ordre 6 :

   $$\widetilde{y}_{n+1} = y_n + \frac{3}{40}k_1 + \frac{875}{2244}k_3 + \frac{23}{72}k_4 + \frac{264}{1955}k_5 + \frac{125}{11592}k_7 + \frac{43}{616}k_8$$

pour estimer l'erreur de troncature locale de la métohde de R-K d'ordre 5 :

$$y_{n+1} = y_n + \frac{13}{160}k_1 + \frac{2375}{5984}k_3 + \frac{5}{16}k_4 + \frac{12}{85}k_5 + \frac{3}{44}k_6$$

avec

$$\begin{aligned}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf\left(x_n + \frac{1}{6}h, y_n + \frac{1}{6}k_1\right) \\
k_3 &= hf\left(x_n + \frac{4}{15}h, y_n + \frac{4}{75}k_1 + \frac{16}{75}k_2\right) \\
k_4 &= hf\left(x_n + \frac{2}{3}h, y_n + \frac{5}{6}k_1 - \frac{8}{3}k_2 + \frac{5}{2}k_3\right) \\
k_5 &= hf\left(x_n + \frac{5}{6}h, y_n - \frac{165}{64}k_1 + \frac{55}{6}k_2 - \frac{425}{64}k_3 + \frac{85}{96}k_4\right) \\
k_6 &= hf\left(x_n + h, y_n + \frac{12}{5}k_1 - 8k_2 + \frac{4015}{612}k_3 - \frac{11}{36}k_4 + \frac{88}{255}k_5\right) \\
k_7 &= hf\left(x_n + \frac{1}{15}h, y_n - \frac{8263}{15000}k_1 + \frac{124}{75}k_2 - \frac{643}{680}k_3 - \frac{81}{250}k_4 + \frac{2484}{10625}k_5\right) \\
k_8 &= hf\left(x_n + h, y_n + \frac{3501}{1720}k_1 - \frac{300}{43}k_2 + \frac{297275}{52632}k_3 - \frac{319}{2322}k_4 + \frac{24068}{84065}k_5 + \frac{3850}{26703}k_7\right)
\end{aligned}$$

Méthodes Linéaires à Pas Multiples (MLPM)

En posant $f_n = f(x_n, y_n)$, on définit une MLPM par $\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}$, avec α_j et β_j des constantes vérifiant les conditions $\alpha_k = 1$ et $|\alpha_0| + |\beta_0| \neq 0$.

Exemple 7.5.3. $y_{n+2} - y_{n+1} = hf_{n+1}$ (ou de façon équivalente $y_{n+1} - y_n = hf_n$) est la méthode d'Euler à un pas.

Définition 7.5.8. On appelle 2^{eme} polynôme caractéristique de la méthode, le polynôme défini par $\sigma(t) = \sum_{j=0}^{k} \beta_j t^j$.
L'opérateur de difference linéaire est défini par

$$L(z(x); h) := \sum_{j=0}^{k} (\alpha_j z(x + jh) - h\beta_j z^{'}(x + jh))$$

Si on suppose que z est une fonction qu'on peut dériver autant de fois qu'on veut, on obtient :

$$L(z(x); h) = C_0 z(x) + C_1 h z'(x) + \cdots + C_q h^q z^{(q)}(x) + \cdots$$

Définition 7.5.9. La MLPM est dite d'ordre p si

$$C_0 = C_1 = \cdots = C_p = 0 \text{ et } C_{p+1} \neq 0.$$

C_{p+1} est appelée constante d'erreur de la MLPM

7.6 Applications

La modélisation et la simulation par ordinateur ont fait de l'outil mathématique un moyen important pour la compréhension, l'analyse et le contrôle de certaines maladies. La formulation d'un modèle permet de tester des hypothèses, d'estimer des paramètres, de construire des théories, de discuter des conjonctures, de formuler des scénarios prédictifs, de visualiser certaines sensibilités et de remédier à l'impossibilité de certaines expériences pour coût élevé ou danger opérationnel. La modélisation mathématique permet de mieux saisir les concepts d'épidémie, de seuil et des nombres de contacts, de replacements ou de reproduction. Le modèle mathématique fournit également le support nécessaire à la réalisation d'appareils et d'équipements d'analyse et de contrôle dans le domaine médical. A ce propos, nous renvoyons à une review récente de Hethcote [**?**] et aux 200 références citées par l'auteur.

Dans ce chapitre, nous considérons trois types d'applications, la première traite les maladies d'infection transmise de façon directe, la deuxième est consacrée à la transmission par vecteur et la troisième traite l'effet de l'effort physique sur les dynamiques d'insuline et de glucose.

Pour les modèles à transmission directe, nous nous limiterons aux cas discrets.

Application 7.6.1. Modèles épidémiologiques
Pour les maladies infectieuses, les modèles mathématiques utilisés sont en général des modèles à compartiments c à d que la population est subdivisée en sous classes :

- $M(t)$: nombre d'individus avec immunité passive à l'instant t ;
- $S(t)$: nombre des susceptibles à l'instant t (on désigne par susceptibles, les individus qui peuvent avoir la maladie) ;
- $E(t)$: nombre des exposés à l'instant t (on désigne par exposés, les individus qui sont infectés mais ne sont pas infectieux) ;
- $I(t)$: nombre des infectieux à l'instant t(les personnes qui sont déjà touchées par la maladie et peuvent transmettre la maladie) ;
- $R(t)$: nombre des résistants à l'instant t (les personnes qui sont guéries avec immunisation)

Chacune des classes précédentes est dite compartiment, et suivant les caractéristiques de chaque maladie on a les types de modèles suivants : SI, SIS, SIR, SEIR, SEIRS,$\cdots$)

Remarque 7.6.1.
- En général les classes M et E sont souvent omises.
- une population est dite fermée si on néglige l'émigration et l'immigration,
- une population est dite homogène si la population se mixe de façon homogène c à d si on néglige, les détails associés à l'âge, la location géographique (ville, village), les facteurs socio-culturels et l'hétérogénéité génétique,
- une population est de taille fixe si le taux de naissance est égal au taux de mortalité.

Exemple 7.6.1. Modèle SIS

C'est un modèle à deux compartiments utilisé essentiellement pour décrire les maladies sexuellement transmissibles. Les guéris ne développent pas une immunité et redeviennent susceptibles à la maladie : deux classes de sous populations susceptibles et infectieux. La dynamique des sous populations est régie par le système d'équations suivant :

$$\begin{cases} S_{n+1} & = S_n(1-\lambda.\Delta t.I_n)+\gamma.\Delta t.I_n+\mu.\Delta t(1-S_n) \\ I_{n+1} & = I_n(1-\gamma.\Delta t-\mu.\Delta t+\lambda.\Delta t.S_n) \\ I_0+S_0 & = 1 \text{ avec } S_0 \geq 0 \text{ et } I_0 > 0 \end{cases}$$

où S_n et I_n sont les proportions d'individus susceptibles et infectieux, respectivement.

Comme $S_n + I_n = 1$ alors $I_{n+1} = I_n(1-(\gamma+\mu).\Delta t+\lambda.\Delta t-\lambda.\Delta t.I_n)$.
- si $(\lambda-\gamma)\Delta t < 2$ alors on a convergence vers le point fixe ;
- si $2 < (\lambda-\gamma)\Delta t < 2.449$ alors on a 2_points cycles ;
- si $(\lambda-\gamma)\Delta t$ est assez grand alors on a comportement chaotique.

$N = 150, I_0 = 2, \gamma = 1.5, \Delta t = .5$

$\lambda = 4.5 \implies$ convergence voir figure 7.1 ;

$\lambda = 6.0 \implies$ 2_points cycles voir figure 7.1 ;

$\lambda = 6.5 \implies$ 4_points cycles voir figure 7.2 ;

$\lambda = 7.0 \implies$ comportemet chaotique voir figure 7.2.

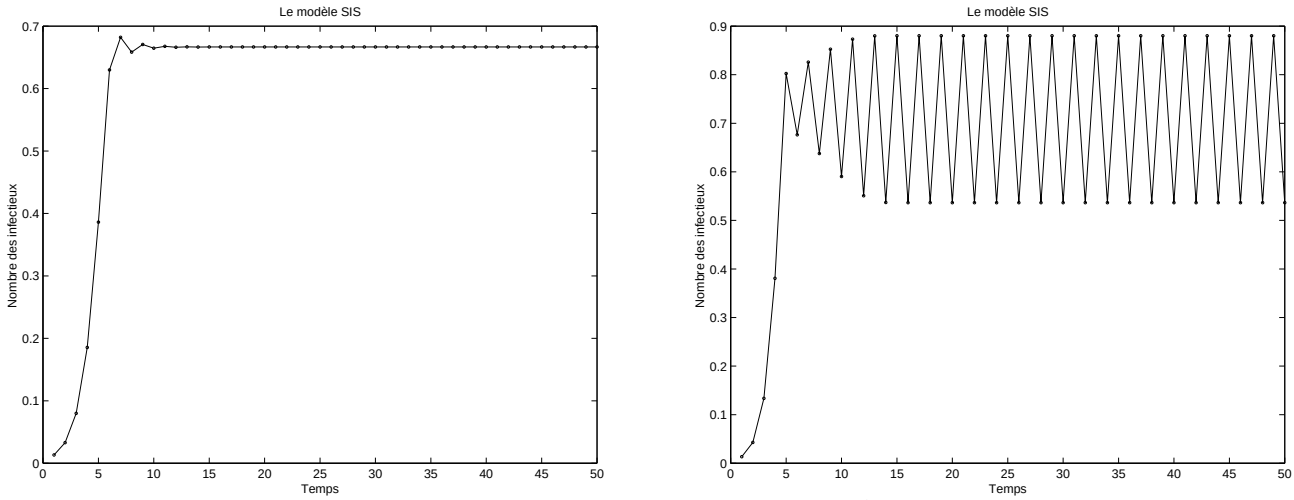


FIG. 7.1 – Convergence $\lambda = 4.5$,2_cycles $\lambda = 6$

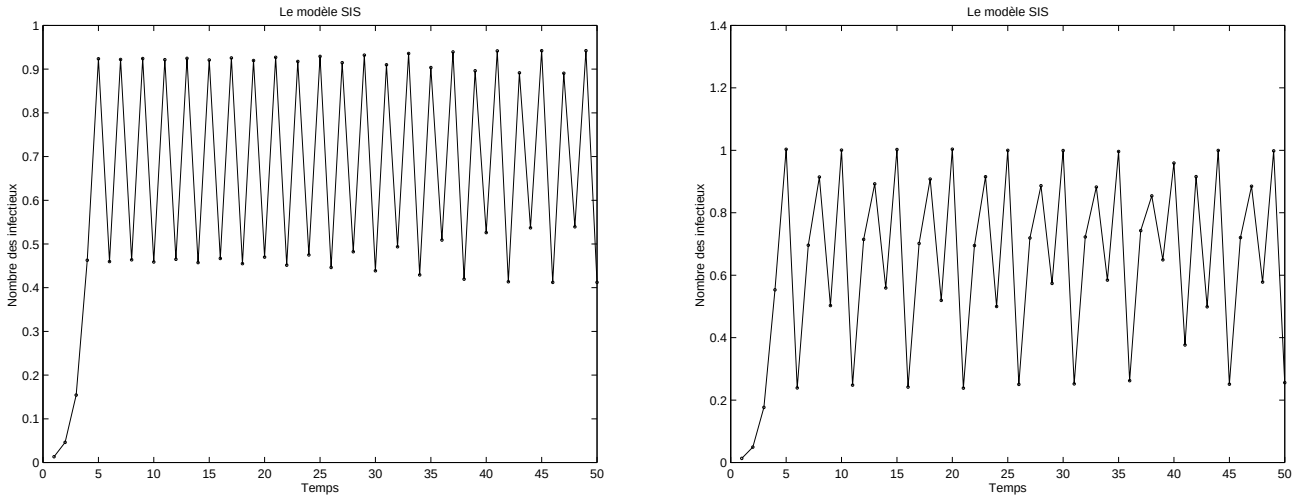


FIG. 7.2 – 4_cycles $\lambda = 6.5$,Chaos $\lambda = 7$

Exemple 7.6.2. Modèle SIR

Le modèle SIR est un modèle simple de transmission des maladies par contact entre les individus ,ces derniers deviennent immunisés après une seule infection, une troixième classe de sous populations est considérée : la classe R des isolés ou enlevés "removed" ce sont les personnes qui, une fois infectés soit deviennent immunisés soit ils meurent. Le modèle SIR est plus adapté pour les maladies d'enfants comme la rougeole. Les équations qui régissent ce modèle sont :

$$\begin{cases} S_{n+1} & = S_n(1-\lambda.\Delta t.I_n)+\mu.\Delta t(1-S_n) \\ I_{n+1} & = I_n(1-\gamma.\Delta t-\mu.\Delta t+\lambda.\Delta t.S_n) \\ R_{n+1} & = R_n(1-\mu.\Delta t)+\gamma.\Delta t.I_n \\ S_0+I_0+R_0 & = 1 \text{ avec } S_0>0,\, I_0>0 \text{ et } R_0\geq 0 \end{cases}$$

$N = 200$,$I_0 = .115$,$S_0 = .885$ et $\Delta t = 0.2$:

1. $\lambda = 3.5$,$\gamma = 2$ voir figure 7.3 ;
2. $\lambda = 5.5$,$\gamma = 2.5$ voir figure 7.3.

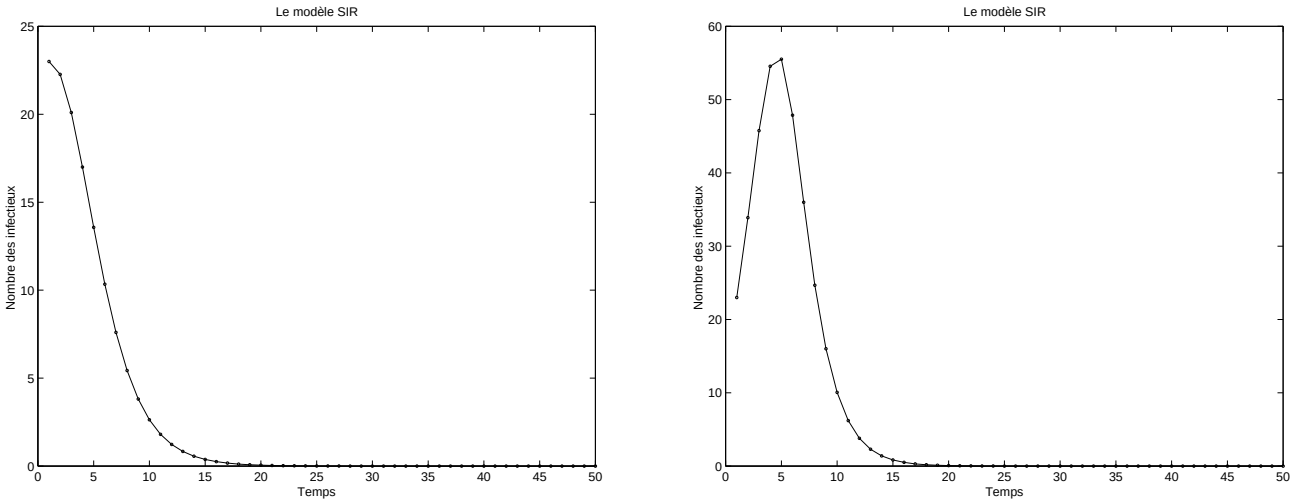


FIG. 7.3 – $\lambda = 3.5$,$\gamma = 2$-$\lambda = 5.5$,$\gamma = 2.5$

(**SIR à deux populations**)

i) $\Delta t = 0.25, \alpha_{11} = 2, \alpha_{12} = 0.5, \alpha_{21} = 4, \alpha_{22} = 2, \gamma_1 = 2, \gamma_2 = 1, N^1 = 100, N^2 = 200, I_0^1 = 10, I_0^2 = 50.$

ii) $\Delta t = 0.25, \alpha_{11} = 2, \alpha_{12} = 0.5, \alpha_{21} = 4, \alpha_{22} = 2, \gamma_1 = 2, \gamma_2 = 1, N^1 = 100, N^2 = 200, I_0^1 = 10, I_0^2 = 150.$

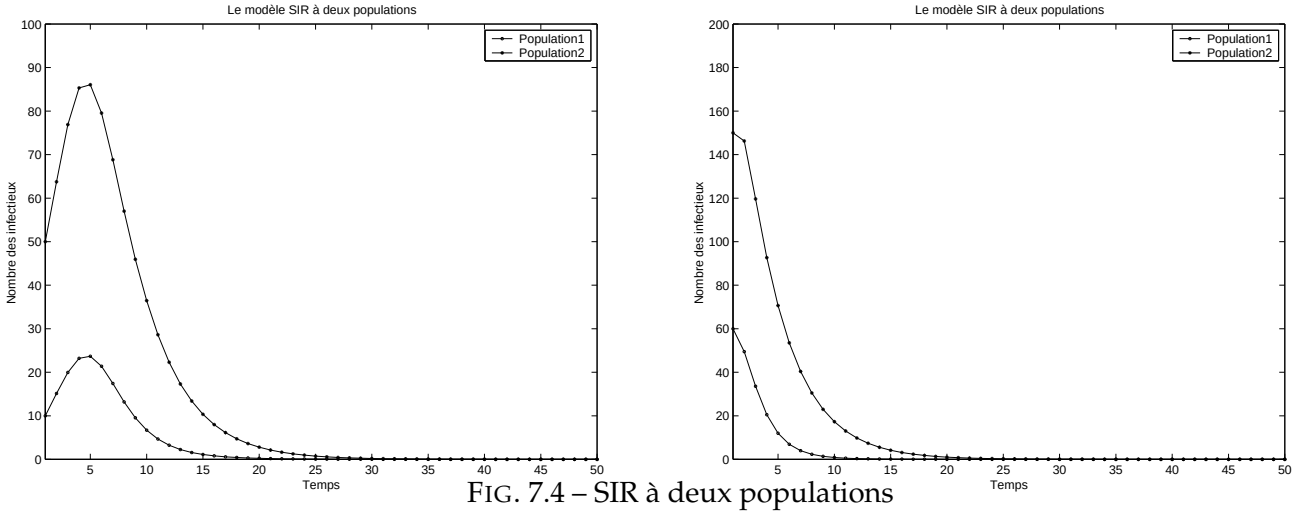


FIG. 7.4 – SIR à deux populations

Exemple 7.6.3. Modèle SEIR

Modèle SEIR avec λ indépendante du temps

Le modèle SEIR est une variante du modèle SIR où la période de latence est prise en compte c.à.d une classe de la population est intermédiaire entre la classe des susceptibles et les infectieux : la classe des exposés qu' un susceptible traverse pour passer à l 'état d'infection. Un autre paramètre est introduit qu'on note β et on parle de période de *latence* $\frac{1}{\beta}$. Les équations régissant ce modèle sont :

$$\begin{cases} S_{n+1} = S_n(1 - \lambda.I_n.\Delta t) + \mu.\Delta t(1 - S_n) \\ E_{n+1} = E_n + S_n\lambda.I_n.\Delta t - (\beta + \mu).\Delta t.E_n) \\ I_{n+1} = I_n(1 - \gamma.\Delta t - \mu.\Delta t) + \beta.\Delta t.E_n \\ R_{n+1} = R_n(1 - \mu.\Delta t) + \gamma.\Delta t.I_n \\ S_0 + I_0 + R_0 + E_0 = 1 \end{cases}$$

On suppose toujours que la taille de la population est constante.
Pour $\lambda = 10^{-6}$ $\beta = 45.6$, $\gamma = 73$, $\mu = 0.02$, $I_0 = 0.0006$, $E_0 = 0.001$ et $S_0 = 0.25$, le modèle SEIR a le comportement suivant :

i) $\Delta t = 0.01 \longrightarrow$ convergence voir figure 7.5.

ii) $\Delta t = 0.018 \longrightarrow$ oscillation puis convergence voir figure 7.5.

iii) $\Delta t = 0.027 \longrightarrow$ oscillation voir figure 7.6.

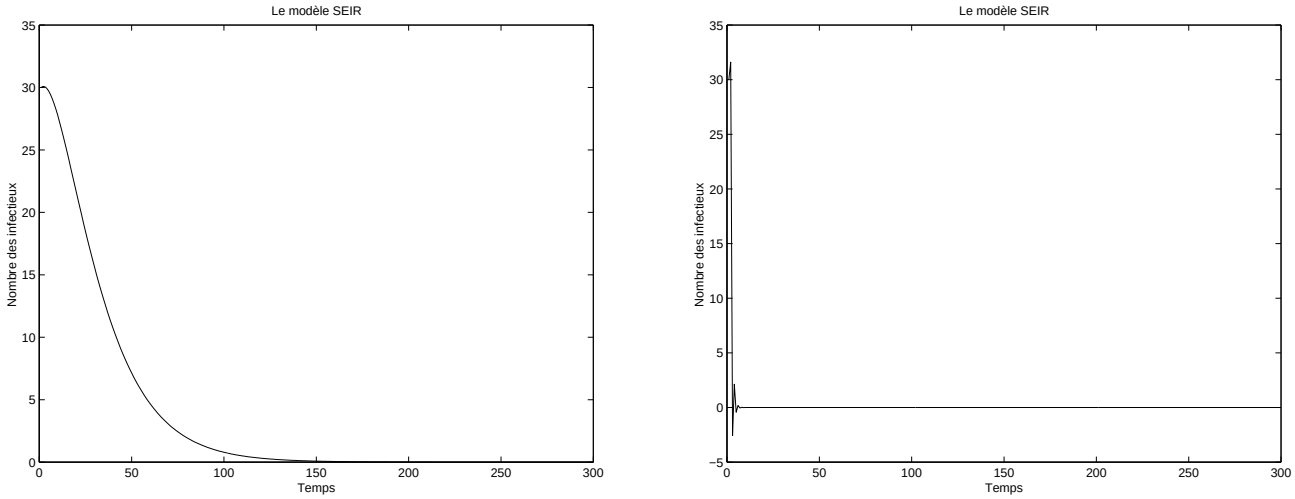


FIG. 7.5 – Convergence $\Delta t = 0.01$, Convergence oscillatoire $\Delta t = 0.018$

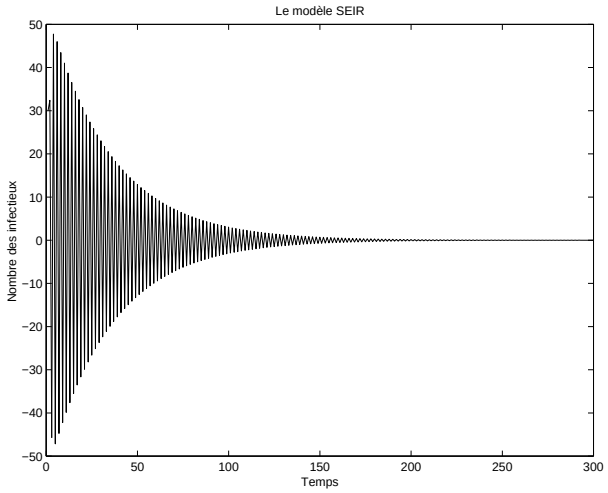


FIG. 7.6 – Oscillation $\Delta t = 0.027$

Exemple 7.6.4. Modèle SEIR saisonnier

Le modèle saisonier SEIR tient compte de la variation du taux d'infections selon les saisons donc on a le système suivant :

$$\begin{cases} S_{n+1} = S_n(1 - \lambda_n.I_n.\Delta t) + \mu.\Delta t(1 - S_n) \\ E_{n+1} = E_n + S_n\lambda_n.I_n.\Delta t - (\beta + \mu).\Delta t.E_n) \\ I_{n+1} = I_n(1 - \gamma.\Delta t - \mu.\Delta t) + \beta.\Delta t.E_n \\ R_{n+1} = R_n(1 - \mu.\Delta t) + \gamma.\Delta t.I_n \\ S_0 + I_0 + R_0 + E_0 = 1 \end{cases}$$

avec $\lambda_n = \lambda(n.\Delta t) = c_0 + c_1(1 + \cos(2\pi n\Delta t))$ (dû à Bolker et Grenfell)

i) Pour $\beta = 45$, $\gamma = 78$, $\mu = 0.01$, $\Delta t = 0.001$, $I_0 = 0.006$, $E_0 = 0.03$ et $S_0 = 0.25$ il y'a une épidémie voir figure 7.7.

ii) Pour $\beta = 45$, $\gamma = 73$, $\mu = 0.02$, $\Delta t = 0.027$, $I_0 = 0.001$, $E_0 = 0.01$ et $S_0 = 0.25$ il y'a oscillation voir figure 7.7.

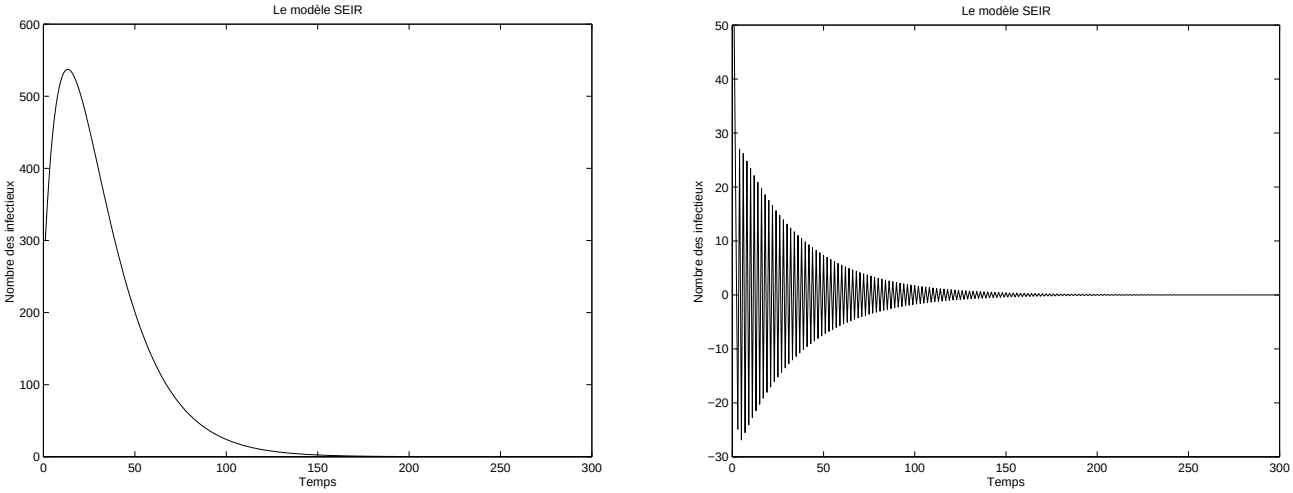


FIG. 7.7 – Convergence-Oscillation

Application 7.6.2. Transmission indirecte par vecteur

Nous considérons le cas de la maladie infectueuse *DENGUE* transmise à l'homme par le mosquito (Aedes) et qui est actuellement endemique dans plus de cent pays d'Afrique, d'Amérique et d'Asie [**?**].

Le modèle comporte un ensemble d'équations pour l'homme et un autre pour le vecteur (mosquito) exprimés à partir du diagramme de la figure 7.8.

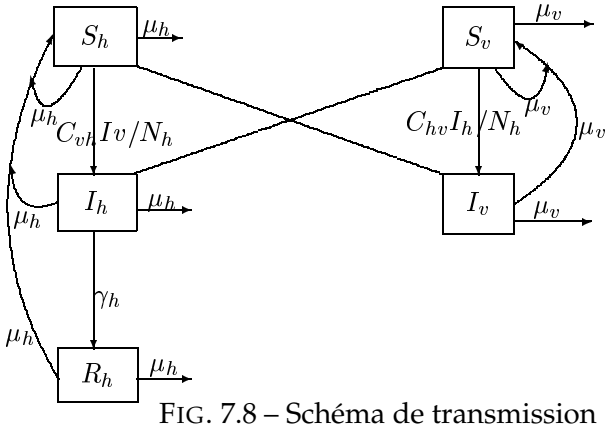


FIG. 7.8 – Schéma de transmission

On suppose qu'on dispose d'une population humaine (respectivement de vecteurs mosquitos) de taille N_h(resp. N_v) composée de Susceptibles S_h, d'infectieux I_h et de guéris R_h (resp. S_v et I_v).

Le modèle suppose un mixage homogène d'humain et de vecteur de telle sorte

que chaque piqure a la même probabilité d'être appliquée à un humain particulier. En notant b_s le taux moyen de piqures d'un vecteur susceptible, p_{hv} la probabilité moyenne de transmission d'un infectieux humain à vecteur susceptible, le taux d'exposition pour les vecteurs est donné par : $(p_{hv} I_h b_s)/N_h$.
en notant p_{vh} la probabilité moyenne de transmission d'un infectieux vecteur à un humain et I_v le nombre de vecteurs infectieux, le taux d'exposition pour les humains est donné par : $(p_{vh} I_v b_i)/N_h$ par conséquent :

- Le taux de contact adéquat d'humain à vecteurs est donné par : $C_{hv} = p_{hv} b_s$
- Le taux de contact adéquat de vecteurs à humain est donné par :
  $C_{vh} = p_{vh} b_i$.

La durée de vie humaine est prise égale à 25 000 jours (68.5 ans), et celle des vecteurs est de : 4 jours. Les autres paramètres sont donnés dans le tableau 7.1 d'après **[?]**.

Nom du paramètre	Notation	valeur de base
probabilité de transmission de vecteur à humain	p_{hv}	0.75
probabilité de transmission d'humain à vecteur	p_{vh}	0.75
piqure par susceptible mosquito par jour	b_s	0.5
piqure par infectieux mosquito par jour	b_i	1.0
taux de contact effectif humain à vecteur	C_{hv}	0.375
taux de contact effectif, vecteur à humain	C_{vh}	0.75
Durée de vie des humains	$\frac{1}{\mu_h}$	25000 jours
Durée de vie des vecteurs	$\frac{1}{\mu_v}$	4 jours
durée de l'infection	$\frac{1}{\mu_h + \gamma_h}$	3 jours

TAB. 7.1 – définitions et valeurs des paramètres

Les équations qui régissent le modèle sont données par
Population humaine

$$\begin{cases} \frac{dS_h}{dt} = & \mu_h N_h - (\mu_h + p + C_{vh} I_v/N_h) S_h \\ \frac{dI_h}{dt} = & (C_{vh} I_v/N_h) S_h - (\mu_h + \gamma_h) I_h \\ \frac{dR_h}{dt} = & p S_h + \gamma_h I_h - \mu_h R_h \end{cases}$$

Population vecteur

$$\begin{cases} \frac{dS_v}{dt} = & \mu_v N_v - (\mu_v + C_{hv} I_h / N_h) S_v \\ \frac{dI_v}{dt} = & (C_{hv} I_h / N_h) S_v - \mu_v I_v \end{cases}$$

Avec les conditions $S_h + I_h + R_h = N_h$, $S_v + I_v = N_v$, $R_h = N_h - S_h - I_h$ et $S_v = N_v - I_v$

Le système s'écrit

$$\begin{cases} \frac{dS_h}{dt} = & \mu_h N_h - (\mu_h + p + C_{vh} I_v / N_h) S_h \\ \frac{dI_h}{dt} = & (C_{vh} I_v / N_h) S_h - (\mu_h + \gamma_h) I_h \\ \frac{dI_v}{dt} = & C_{hv} I_h / N_h (N_v - I_v) - \mu_v I_v \end{cases}$$

Pour l'étude de stabilité et d'autres détails concernant les paramètres du modèle, le lecture pourra consulter le papier cité en réferences. Dans le cadre restreint de ce chapitre, nous discutons très brièvement l'output du modèle.

La diffuculté principale de la dengue provient du fait qu'elle est causée par quatre virus différents et que l'ummunité provisoire acquise contre un virus ne protège pas contre les autres virus, au contraire, un individu attaqué par un deuxième virus court le danger d'évoluer vers le stade de Dengue Hemorragique.

La recherche de vaccin se complique par le fait que ce vaccin doit couvrir le spectre des quatre virus et ceci constitue un problème.

Le modèle montre que la réduction du nombre de mosquitos (par insecticides ou autre) n'est pas suffisante pour éradiquer l'épidémie de la Dengue. Elle peut tout au plus retarder l'apprition de l'épidémie. Comme la découverte d'un vaccin global n'est pas attendue dans le court terme, les auteurs suggèrent la combinaison du contrôle des facteurs environnementaux et de vaccins partiels contre chaque virus.

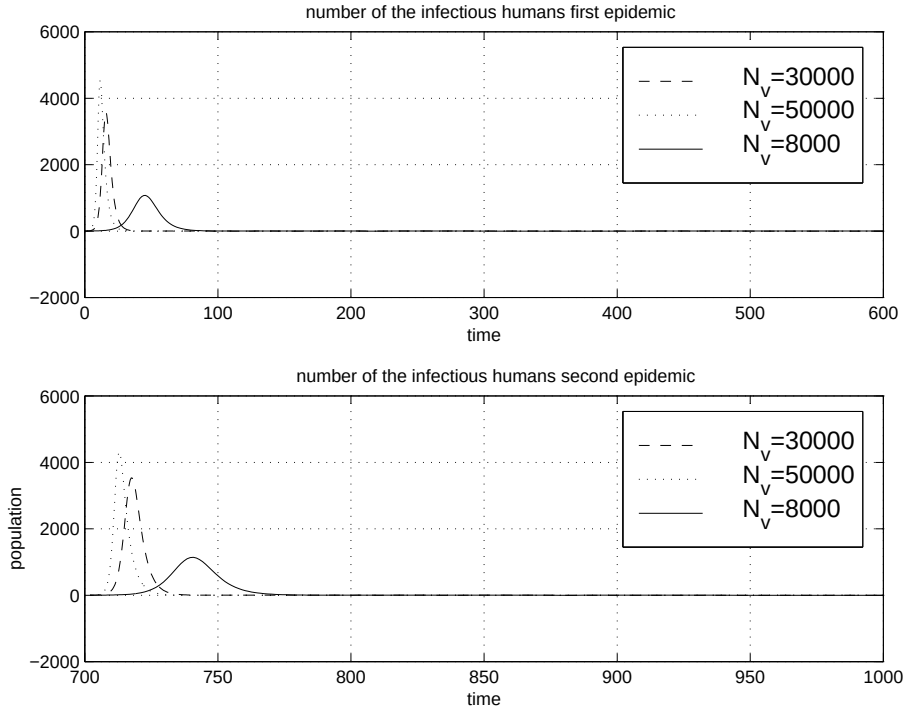


FIG. 7.9 –

Application 7.6.3. (**Diabète et effort physique[?]**)
Pour un diabétique, l'effort physique figure au même niveau que le traitement medical et la diétitique. Il peut l'aider à contrôler la quantité de sucre dans le sang de façon directe ou indirecte en améliorant la sensibilité de l'insuline et la réponse des muscles et des cellules ou encore en combattant l'excès de poids qui constitue un facteur à risque. S'il est bien connu que les spécialistes du diabètes insistent toujours sur le rôle de l'effort physique en prenant en compte les capacités et les données de chaque individu, il est aussi intéressant de voir comment les modèles mathématiques sont utilisés dans ce domaine.

En 1939, Himsworth et Ker ont introduit la première approche de mesure de sensibilité d'insuline in vivo. Les modèles mathématiques ont été utilisés pour la dynamique de glucose et d'insuline. Le pionnier dans ce domaine est Bolie(1961) qui a proposé un modèle simple supposant que la disparition du glucose est une fonction linéaire du glucose et d'insuline , la secretion d'insuline est proportionnelle au glucose et elle disparait proportionnellement à la concentration d'insuline dans le plasma. Ce modèle qui sera utilisé avec quelques modifications par d'autres auteurs (Akerman et al (1965), Della et al (1970), Serge et al. (1973)) peut-être formulé par le système différentiel suivant :

$$\begin{cases} \dfrac{dG(t)}{dt} = -a_1 G - a_2 I + p \\ \dfrac{dI(t)}{dt} = -a_3 G - a_4 I \end{cases}$$

où $G = G(t)$ représente la concentration de glucose et $I = I(t)$ représente l'insuline, p, a_1, a_2, a_3, a_4 sont des paramètres.

Durant les dernières décennies, une littérature abondante a été consacrée à ce sujet, le lecteur interessé pourra consulter les deux reviews récentes par Bellazi et al (2001) et Parker et al. (2001).

Incorporant l'effet de l'effort physique sur la dynamique du glucose et de l'insuline, Derouich et Boutayeb (2002) ont proposé le modèle suivant

$$\begin{cases} \dfrac{dG(t)}{dt} = -(1+q_2)X(t)G(t) + (p_1+q_1)(G_b - G(t)) \\ \dfrac{dI(t)}{dt} = (p_3+q_3)(I(t)-I_b) + p_2X(t) \end{cases}$$

où $(I(t) - I_b)$ représente la différence entre l'insuline dans le plasma et l'insuline de base $(G_b - G(t))$ représente la différence entre la concentration de glucose dans le plasma et le glucose de base $X(t)$ est l'insuline interstitiale, q_1, q_2,q_3 sont des paramètres liés à l'effort physique.

Les auteurs ont discuté les résultats de ce modèle dans trois cas différents :

- cas normal(non diabétique),
- cas diabétique insulino-dépendant,
- cas diabétique non insulino-dépendant.

Dans les trois cas, le modèle met en évidence l'intérêt de l'effort physique à améliorer la sensibilité de l'insuline mais le dernier cas reste le plus illustratif comme l'indique la figure 7.10 où on voit qu'un diabétique non insulino-dépendant pourrait étre amené à vivre continuellement avec 2g/ml de glucose dans le sang sans se rendre compte des conséquences à long terme de cette overdose. Avec un effort physique, le diabétique peut ramener la courbe de glucose au voisinage de la concentration normale de 1g/ml.

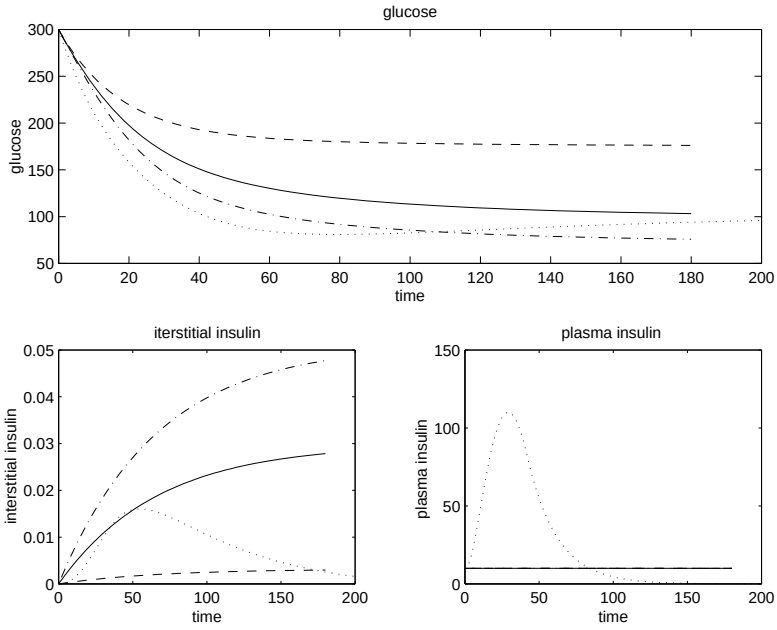


FIG. 7.10 – Effet de l'effort physique

7.7 Complément bibliographique

La modélisation épidémiologique remonte au moins à 1760 lorsque Bernouilli proposa un modèle mathématique pour la variole.

En 1906 un modèle discret pour l'épidémie de la rougeole a été considéré par Haner.

Ross utilisa les équations differentielles pour des modèles hôte-vecteur de Malaria en 1911.

D'autres modèles déterministes ont été proposés par Ross, Hudson et Lotka (1922).

A partir de 1926 la contribution de Kermack et McKendrick restera marquée par le seuil endémique que la densité des susceptibles doit dépasser pour que l'épidémie se propage .

Le livre de Bailey publié en 1957 sur la théorie mathématique des maladies infectieuses marquera la lancée d'une nouvelle vague de la deuxième moitié du 20^{eme} siécle.

Bartlett (1960) est parmi les precurseurs des modèles épidémiologiques stochastiques même si ceux-ci ont été utilisés auparavant par Yule(1924) et d'autres.

Les années 60 verront aussi le débat s'animer entre stochastique et determinisme. Certains écologistes vont soutenir que les modèles stochastiques sont les mieux adaptés à décrire la nature avec des facteurs imprédictibles et des événe-

ments aléatoires conduisant à des séries chronolgiques. D'autres vont défendre les modéles déterministes sous-jaçant l'idée du chaos apparaissant comme du stochastique ou un bruit blanc dans les séries chronologiques. Les travaux de May (1974, 75) et plus particulièrement son excitant papier publié dans la revue Nature en 1976 sous le titre de "Simple mathematical models with very complicated dynamics" constitueront des réferences pour les écologistes. Par la suite, Anderson et May vont publier une multitude de papiers et de livres dévoués à l'épidémiologie des maladies infectieuses.

Hoppensteadt (1975) est considéré par certains auteurs comme le premier à avoir proposé une analyse mathématique des modèles avec structure d'âge.

Durant les dernières décennies, la littérature consacrée à la modélisation épidémiologique a connu un essort considérable comme en atteste la dizaine de reviews (Becker, 1978 ; Castillo-Chavez, 1989 ; Dietz, 1967-75-85 ; Wickwire, 1977 ; Hethcote, 1981-94-2000 ; Isham, 1988) et la trentaine de livres (Anderson & May, 1982-91 ; Bailey, 1975-82 ; Bartlett, 1960 ; Castillo-Chavez, 1989 ; Diekman, 2000 ; Frauenthal, 1980 ; Grenfell and Dobson, 1995 ; Hethcote, 1984-92 ; Isham & Medeley, 1996 ; Kranz, 1990 ; Lauwerier, 1981 ; Ludwig & Cooke, 1975 ; Mollison, 1991 ; Scott & Smith, 1994). D'autres ouvrages et reviews traitent les modèles épidémiologiques comme partie des modèles écologiques (Dietz K. (1975), Pielou (1977), Okubo (1980), Kot (2000)).

7.8 Exercices

Exercice 7.8.1. On considère le modèle discret suivant :

$$S_{n+1} = \exp(-aI_n)S_n$$

$$I_{n+1} = bI_n + (1 - \exp(-aI_n))S_n$$

$$R_{n+1} = (1-b)I_n + R_n$$

On suppose que $R_0 = 0$, montrer que

1. La population des susceptibles tend vers une limite S quand n tend vers l'infini.
2. Si $F = S/S_0$ alors

$$F = exp(-aS_0/(1-b))(1 + I/S_0 - F).$$

3. Expliquer comment F peut jouer un rôle de seuil.

Exercice 7.8.2. Chercher l'ordre et la constante d'erreur C_p pour la méthode suivante :

$$y_{n+2} - (1+\alpha)y_{n+1} + \alpha y_n = h((1+\beta)f_{n+2} - (\alpha + \beta + \alpha\beta)f_{n+1} + \alpha\beta f_n)$$

Exercice 7.8.3. Chercher les solutions des systèmes differentiels et aux differences suivants : $y^{'} = Ay + \psi(x) \qquad , y(0) = \alpha$ avec

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \psi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}, \qquad \alpha = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix}$$

$y_{n+4} - 6y_{n+3} + 14y_{n+2} - 16y_{n+1} + 8y_n = \phi_n$ avec $y_n \in \mathbb{R}^2$ et $\phi_n = \begin{pmatrix} n \\ 1 \end{pmatrix}$.

Exercice 7.8.4. Considérons le système aux differences :

$$y_{n+2} - 2\mu y_{n+1} + \mu y_n = c, \;\; n = 0, 1, \cdots$$

avec $y_n, c \in \mathbb{R}^m$ et $0 \leq \mu \leq 1$.

Montrer que y_n converge vers $c/(1-\mu)$ quand $n \longrightarrow \infty$.

Exercice 7.8.5. A- Montrer que :

i) Pour tout $x \geq -1$ et toute constante positive m on a :

$$0 \leq (1+x)^m \leq \exp(mx)$$

ii) Si s et t sont des réels positifs et $(z_n)_{n=0}^{n=k}$ une suite vérifiant :

$$z_0 \geq -\frac{t}{s} \text{ et } z_{n+1} \leq (1+s)z_n + t \ \ \forall\, n = 1, \cdots, k$$

Alors on a :

$$z_{n+1} \leq \exp((n+1)(1+s)) \left(\frac{t}{s} + z_0\right) - \frac{t}{s}.$$

B- Soit $y(x)$ la solution unique du p.c.i $y'(x) = f(x,y)$,$a \leq x \leq b$, $y(a) = \alpha$ et $w_0, w_1, \cdots, w_N$, les approximations vérifiant : $w_0 = \alpha + \delta_0$ et $w_{n+1} = w_n + hf(x_n, w_n) + \delta_{n+1}; n = 0, ..., N-1$

1. On suppose que : $D = \{(x,y); a \leq x \leq b, -\infty < y < \infty\}$ avec a et b finis
   i) il existe une constante L positive telle que :
   $$|f(x,y) - f(x,y^*)| \leq L|y - y^*| \forall y, y^* \in D$$
   ii) il existe une constante M telle que :$|y''(x)| \leq M$ pour tout $x \in [a,b]$.
   iii) les perturbations verifient : $|\delta_n| < \delta \ \forall n = 0, \cdots, N$
   Montrer que :
   $$|y(x_n) - w_n| \leq \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)(\exp(L(x_n - a)) - 1) + |\delta_0| \exp L(x_n - a)$$
   (Preuve : identique à celle du théorème du cours avec $y(x_0) - w_0 = \delta_0$ et $t = \frac{h^2 M}{2} + |\delta_n|$)
2. En posant $\varepsilon(h) = \frac{hM}{2} + \frac{\delta}{h}$ et en remarquant que $\lim_{h \to 0} \varepsilon(h) = \infty$, montrer qu'on peut déterminer une limite inférieure h_0 de h qui rend minimum $\varepsilon(h)$.

Exercice 7.8.6. Soit A une matrice diagonalisable admettant les vecteurs propres V_j associés aux valeurs propres λ_j avec $Re(\lambda_j) < 0$ pour tout j.

On considère les deux schémas d'Euler donnés par :

$$E1 \ : \ u_{n+1} = u_n + hAu_n, \quad u_0 = \eta$$
$$E2 \ : \ u_{n+1} = u_n + hAu_{n+1}, \quad u_0 = \eta$$

Montrer que les solutions de ces deux schémas sont données par :

$$S1 \ : \ u_n = \sum_{j=1}^{k} c_j (1 + h\lambda_j)^n V_j$$
$$S2 \ : \ u_n = \sum_{j=1}^{k} c_j (1 - h\lambda_j)^{-n} V_j$$

Quelles conditions doit-on imposer à h pour que les solutions $S1$ et $S2$ tendent vers zéro ou restent bornées quand $n \longrightarrow \infty$

Exercice 7.8.7. En posant $f_n = f(x_n, y_n)$, on définit une MLPM par :

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}$$

avec α_j et β_j des constantes vérifiant les conditions :

$$\alpha_k = 1 \text{ et } |\alpha_0| + |\beta_0| \neq 0$$

On appelle 1 er (resp. 2 eme) polynôme caractéristique de la MLPM le polynôme $\rho(t)$(resp. $\sigma(t)$) :

$$\rho(t) = \sum_{j=0}^{k} \alpha_j t^j, \quad \sigma(t) = \sum_{j=0}^{k} \beta_j t^j.$$

Si l'opérateur de difference linéaire est donné par :

$$L(z(x); h) := \sum_{j=0}^{k} (\alpha_j z(x + jh) - h\beta_j z^{'}(x + jh)) = C_0 z(x) + C_1 h z'(x) + \cdots + C_q h^q z^{(q)}(x) + \cdots$$

1. Montrer que :

$$C_0 = \sum_{j=0}^{k} \alpha_j = \rho(1)$$

$$C_1 = \sum_{j=0}^{k} (j\alpha_j - \beta_j) = \rho'(1) - \sigma(1)$$

$$C_q = \sum_{j=0}^{k} \left(\frac{1}{q} j^q \alpha_j - \frac{1}{(q-1)} j^{q-1} \beta_j \right), \quad q = 2, 3, \cdots$$

2. La MLPM est dite consistante si son ordre est supérieur ou égal à 1 $(p \geq 1)$
   Montrer que la MLPM est consistante si et seulement si :

$$\rho(1) = 0 \text{ et } \rho'(1) = \sigma(1)$$

Exercice 7.8.8. Chercher l'ordre des méthodes numériques suivantes :

1) $y_{n+2} = y_n + \frac{h}{6}(f_n + 4f_{n+1} + f_{n+2})$
2) $y_{n+4} = y_n + \frac{4h}{3}(2f_{n+1} - f_{n+2} + 2f_{n+3})$
3) $y_{n+1} = y_n + \frac{h}{4}\left(f(x_n, y_n) + 3f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}k_1) \right), k_1 = hf(x_n, y_n)$

4) $y_{n+1} = y_n + \frac{h}{6}(k_1 + 4k_2 + k_3)$

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\right) \\ k_3 &= hf(x_n + h, y_n - k_1 + 2k_2) \end{aligned}$$

Exercice 7.8.9. Considérons la méthode numérique

$$y_{n+2} - (1+\alpha)y_{n+1} + \alpha y_n = \frac{h}{2}((3-\alpha)f_{n+1} - (1+\alpha)f_n)$$

où $f_n = f(x_n, y_n)$ et $-1 \leq \alpha \leq 1$.

1. Donner l'erreur de troncature locale de la méthode et l'ordre de la méthode en fonction de α.
2. Cette méthode est utilisée numériquement pour résoudre l'équation scalaire

$$\begin{cases} y'(x) = y(x) \\ y(0) = 1 \end{cases}$$

$$\text{en supposant que} \quad \begin{aligned} y_0 &= 1 + \omega h^3 \\ y_1 &= \exp(h) + \theta h^3 \end{aligned}$$

Montrer que la solution approchée y_n peut-être donnée sous la forme

$$y_n = \frac{\Omega(r_2)r_1^n - \Omega(r_2)r_1^n}{r_1 - r_2}$$

où r_1 et r_2 sont les racines de l'équation caractéristique asociées à l'équation aux différences. et $\Omega(r) = \exp(h) - r + (\theta - r\omega)h^3$.

3. On suppose que r_1 est de la forme

$$r_1 = 1 + h + \frac{h^2}{2} + O(h^3)$$

   (a) Donner une expression analogue pour r_2.

   (b) Etudier la convergence de y_n dans le cas $\alpha = -1$.

Exercice 7.8.10. On considère la méthode numérique suivante

$$y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h(\beta_1 f_{n+1} + \beta_0 f_n) \quad (M)$$

où $f_n = f(x_n, y_n)$ et $f(x_n, y(x_n)) = y'(x_n)$.

1. Déterminer les constantes $\alpha_0, \alpha_1, \beta_0$ et β_1 pour que la méthode soit d'ordre maximum.
2. La méthode est-elle zéro-stable pour les valeurs des constantes trouvées ?
3. On utilise la méthode (M) avec $y_0 = 1$ et $y_1 = \exp(-h)$ pour approcher la solution du problème de condition initiale suivant

$$y'(x) = -y(x), y(0) = 1 \qquad (P)$$

Montrer qu'on obtient l'équation aux différences suivante

$$y_{n+2} + 4(1+h)y_{n+1} + (-5+2h)y_n = 0; n = 0, 1, \cdots \qquad (D)$$

4. Montrer que les racines r_1 et r_2 de l'équation caractéristique associée à (D) sont

$$r_1 = -2-2h+3\left(1+\frac{2}{3}h+\frac{4}{9}h^2\right)^{1/2} \text{ et } r_2 = -2-2h-3\left(1+\frac{2}{3}h+\frac{4}{9}h^2\right)^{1/2}.$$

5. Montrer que la solution de (D) est de la forme

$$y_n = C_1(r_1)^n + C_2(r_2)^n,$$

avec $C_1 = \dfrac{r_2 - \exp(-h)}{r_2 - r_1}$ et $C_2 = \dfrac{\exp(-h) - r_1}{r_2 - r_1}$.

6. On donne $\left(1+\frac{2}{3}h+\frac{4}{9}h^2\right)^{1/2} = 1+\frac{1}{3}h+\frac{1}{6}h^2-\frac{1}{18}h^3+\frac{1}{216}h^4+O(h^5)$.
En déduire que $r_1 = 1-h+\frac{1}{2}h^2-\frac{1}{6}h^3+\frac{1}{72}h^4+O(h^5)$, $r_2 = -5-3h+O(h^2)$.
7. Montrer alors que $C_1 = 1+O(h^2)$ et $C_2 = -\frac{1}{216}h^4+O(h^5)$.
8. En considérant pour x fixé, $nh = x$, montrer que $C_1(r_1)^n \longrightarrow \exp(-x)$ quand $n \longrightarrow \infty$.
9. Montrer que (y_n) diverge. Cette divergence était-elle attendue ? (justifiez).

7.9 Examen d'Analyse Numérique Session de juin

Exercice 7.9.1. : (6 points)

Soit $f(t) = \dfrac{2tLog(t)}{(1+t^2)^2}$ et $F(x) = \int_1^x f(t)dt$ avec $x \in]0, +\infty[$

1) Montrer que $F(x) = \dfrac{x^2Log(x)}{1+x^2} - \dfrac{1}{2}Log(1+x^2) + \dfrac{1}{2}Log(2)$

2) Déduire $\lim_{x\to 0} F(x)$

3) Montrer que $F(\frac{1}{x}) = F(x)$ et déduire $\lim_{x\to+\infty} F(x)$

4) Etudier la nature de $\int_0^{+\infty} f(t)dt$

Exercice 7.9.2. : (7points)

Soient f une fonction de classe $C^3([0,1])$ et $\varepsilon \in]0,1[$

Soit $P^\varepsilon(x)$ le polynôme interpolant f aux points : 0, ε et 1

1) Exprimer $P^\varepsilon(x)$ dans la base de Newton

2) Montrer que pour tout $x \in [0\ ,\ 1\ ]$, on a :

$\lim_{\varepsilon\to 0} P^\varepsilon(x) = P(x) = f(0) + xf'(0) + (f(1) - f(0) - f'(0))x^2 \qquad (*)$

3) Montrer que le polynôme $P(x)$ trouvé dans $(*)$ est l'unique polynôme qui vérifie :

$P(0) = f(0)\ ;\ P(1) = f(1)$ et $P'(0) = f'(0)$

4) Pour un x arbitraire dans $]0\ ,\ 1\ [$ on pose $K = \dfrac{f(x) - P(x)}{x^2(x-1)}$ et on considère la fonction :

$\Psi(t) = f(t) - P(t) - Kt^2(t-1) = f(t) - P(t) - \dfrac{f(x) - P(x)}{x^2(x-1)}t^2(t-1)$

a) Vérifier que $\Psi(0) = \Psi'(0) = \Psi(1) = \Psi(x) = 0$

b) Montrer que Ψ' s'annule en 3 points de $[0\ ,\ 1\ ]$ et Ψ'' s'annule en 2 points de $[0,1]$

c) Déduire qu'il exisite $\theta \in [0\ ,\ 1\ ]$ tel que $\Psi^{(3)}(\theta) = 0$ et conclure que :

$$f(x) - P(x) = \frac{f^{(3)}(\theta)}{6}x^2(x-1)$$

5) On pose $M_3 = \max_{0\le t\le 1}\left|f^{(3)}(t)\right|$,

a) Montrer que $|f(x) - P(x)| \le \dfrac{2M_3}{81} \quad \forall\, x \in [0,1]$

b) Montrer que : $\left|\int_0^1 (f(x) - P(x))dx\right| \le \dfrac{M_3}{72}$

Exercice 7.9.3. (6 points) :

A) On considère l'équation différentielle :

$(E1)$ $\qquad \dfrac{dP}{dt} = rP(t)(1 - \dfrac{P(t)}{K})$ où $K \in \mathbb{R}_+^*$ et $r \in \mathbb{R}$ avec $P(0) \neq 0$ et $P(0) \neq K$.

1. En faisant un changement de variables, montrer que la résolution de $(E1)$ revient à celle de $(E2)$ $\qquad \dfrac{dy}{dt} = ry(t)(1 - y(t))$ où $K \in \mathbb{R}_+^*$ et $r \in \mathbb{R}$
2. Intégrer l'équation différentielle $(E2)$
3. Déduire la solution de l'équation différentielle $(E1)$ qui vérifie $P(0) = P_0$ (P_0 donné)
4. Chercher $\lim_{t\to+\infty} P(t)$ selon les valeurs de r

B) La discrétisation de $(E2)$ conduit à la suite récurrente (u_n) définie par :

$0 < u_0 < 1$ et $u_{n+1} = h(u_n) = ru_n(1 - u_n) \quad \forall n \ge 0 \ \ r \in]0,4[$

1. Trouver les points fixes θ_1 et θ_2 de h
2. A quelles conditions sur r la suite (u_n) converge-t-elle vers les points fixes θ_1 et θ_2?

7.10 Examen Analyse Numérique Session de juin (rattrapage)

Exercice 7.10.1. : (4 points)

On définit la fonction $g(x) = \dfrac{1}{1-x}$ pour $x \in [0, 1[$

1) Calculer $F(x) = \int_0^x g(t)dt$ où $x < 1$ et donner la valeur $F(\frac{2}{3})$

2) Donner l'expression du polynôme de Lagrange $Q(x)$ interpolant g en : 0, $\frac{1}{3}$ et $\frac{2}{3}$

3) Trouver les coéfficients d_0 , d_1 et d_2 tels que pour tout polynôme P de degré inférieur ou égal à 2 $(P(x) = ax^2 + bx + c)$ on ait :

$\int_0^{\frac{2}{3}} P(t)dt = d_0P(0) + d_1\ P(\frac{1}{3}) + d_2\ P(\frac{2}{3})$ $(*)$

4) Utiliser l'égalité $(*)$ pour donner une valeur approchée de $Log(3)$

Exercice 7.10.2. : (4 points)

Soit h une fonction de classe $C^3(\mathbb{R})$ dont la dérivée et la dérivée seconde ne s'annulent pas au voisinage de a et telle que : $h(a) = 0$

On considère les suites (x_n) et (y_n) définies par la donnée de x_0 et les relations :

$y_n = x_n - \dfrac{h(x_n)}{h'(x_n)}$

$x_{n+1} = y_n - \dfrac{h(y_n)}{h'(x_n)}$

On suppose que ces suites convergent vers a et que $\forall n \in \mathbb{N}$, $x_n \neq a$ et $y_n \neq a$

1. Montrer que $\lim\limits_{n\to\infty} \dfrac{y_n - a}{(x_n - a)^2} = \dfrac{h''(a)}{2h'(a)}$
2. Calculer $\lim\limits_{n\to\infty} \dfrac{x_{n+1} - a}{(x_n - a)(y_n - a)}$ en fonction de $h'(a)$ et $h''(a)$
3. Démontrer que la suite (x_n) est d'ordre au moins 3
4. Trouver l'ordre de convergence de la suite (y_n) .

7.11 Corrigé(Session de Juin)

Exercice 7.11.1. : (6 points)

Soit $f(t) = \dfrac{2tLog(t)}{(1+t^2)^2}$ et $F(x) = \int_1^x f(t)dt$ avec $x \in]0, +\infty[$

1) Montrer que $F(x) = \dfrac{x^2Log(x)}{1+x^2} - \dfrac{1}{2}Log(1+x^2) + \dfrac{1}{2}Log(2)$

En faisant une intégration par partie où $U'(t) = \dfrac{2tLog(t)}{(1+t^2)^2}$ et $V(t) = Log(t)$ on a :

$$F(x) = \left[\frac{-Log(t)}{(1+t^2)}\right]_1^x + \int_1^x \frac{1}{t(1+t^2)}dt = \left[\frac{-Log(t)}{(1+t^2)}\right]_1^x + \left[Log(t) - \frac{1}{2}Log(1+t^2)\right]_1^x$$
$$= \frac{-Log(x)}{(1+x^2)} + Log(x) - \frac{1}{2}Log(1+x^2) + \frac{1}{2}Log(2) \quad \textbf{(2 points)}$$
$$= \frac{x^2Log(x)}{(1+x^2)} - \frac{1}{2}Log(1+x^2) + \frac{1}{2}Log(2) \quad (1)$$

2) Déduire $\lim\limits_{x\to 0} F(x)$ **(0.5 point)**

L'équation (1) nous donne $\lim\limits_{x\to 0} F(x) = \dfrac{1}{2}Log(2)$ (en utilisant $\lim\limits_{x\to 0} \dfrac{x^2Log(x)}{(1+x^2)} = 0$)

3) Montrer que $F(\dfrac{1}{x}) = F(x)$ et déduire $\lim_{x\to+\infty} F(x)$ **(1.5 point = 1+ 0.5)**

En faisant le changement de variable $u = \dfrac{1}{t}$ on obtient :

$$t = 1 \to u = 1 \qquad t = x \to u = \frac{1}{x} \quad dt = \frac{-du}{u^2} \text{ et } \frac{2tLog(t)}{(1+t^2)^2}dt = \frac{2uLog(u)}{(1+u^2)^2}du$$

D'où $F(x) = \int_1^x \dfrac{2tLog(t)}{(1+t^2)^2} = \int_1^{\frac{1}{x}} \dfrac{2uLog(u)}{(1+u^2)^2} = F(\dfrac{1}{x})$

On en déduit que $\lim_{x\to+\infty} F(x) = \lim_{x\to 0} F(\dfrac{1}{x}) = \lim_{x\to 0} F(x) = \dfrac{1}{2}Log(2)$

4) Etudier la nature de $\int_0^{+\infty} f(t)dt$ **(2 points = 1+ 1)**

On a $\int_0^{+\infty} f(t)dt = \int_0^1 f(t)dt + \int_1^{+\infty} f(t)dt = -\lim_{x\to 0}\int_1^x f(t)dt + \lim_{x\to+\infty}\int_1^x f(t)dt$
$$= -\lim_{x\to 0} F(x) + \lim_{x\to+\infty} F(x)$$

D'après la question précédente ces deux limites existent donc $\int_0^{+\infty} f(t)dt$ converge

On peut aussi le faire directement en décomposant et en intégrant par parties ou en utilisant les fonctions équivalentes.

Exercice 2 : (7 points)

Soient f une fonction de classe $C^3([0,1])$ et $\varepsilon \in]0,1[$

Soit $P^\varepsilon(x)$ le polynôme interpolant f aux points : 0, ε et 1

1) Exprimer $P^\varepsilon(x)$ dans la base de Newton **(1 point)**

$$P^\varepsilon(x) = f(0) + \frac{f(\varepsilon)-f(0)}{\varepsilon}x + \frac{\dfrac{f(1)-f(\varepsilon)}{1-\varepsilon} - \dfrac{f(\varepsilon)-f(0)}{\varepsilon}}{1-0}x(x-\varepsilon)$$

2) Montrer que pour tout $x \in [0\ ,\ 1\ ]$, on a : **(1 point)**

$$\lim_{\varepsilon\to 0} \frac{f(\varepsilon)-f(0)}{\varepsilon} = f'(0) \quad \textbf{(0.25 points)}$$

$$\lim_{\varepsilon \to 0} \frac{f(1)-f(\varepsilon)}{1-\varepsilon} = f(1)-f(0) \quad \textbf{(0.25 points)}$$

D'où

$$\lim_{\varepsilon \to 0} P^{\varepsilon}(x) = P(x) = f(0) + xf\,'(0) + (f(1)-f(0)-f\,'(0))x^2 \qquad (*)$$

3) Montrer que le polynôme $P(x)$ trouvé dans $(*)$ est l'unique polynôme qui vérifie :

$P(0) = f(0)\ ;\ P(1) = f(1)$ et $P\,'(0) = f\,'(0)$ **(0.25 point)**

En effet, on vérifie bien que : $P(0) = f(0)\ ;\ P(1) = f(1)$ et $P\,'(0) = f\,'(0)$

Par ailleurs, si $Q(x) = ax^2 + bx + c$ est un autre polyôme qui vérifie $(**)$ alors on a :

$Q(0) = c = f(0)$

$Q'(0) = b = f'(0)$

$Q(1) = a + b + c = f(1) \rightarrow a = f(1) - f(0) - f'0)$

c.a.d : $Q(x) = (f(1) - f(0) - f'0))x^2 + f'(0)x + f(0)$ **(0.75 point)**

4) Pour un x arbitraire dans $]0\ ,\ 1\ [$ on pose $K = \dfrac{f(x)-P(x)}{x^2(x-1)}$ et on considère la fonction :

$$\Psi(t) = f(t) - P(t) - Kt^2(t-1) = f(t) - P(t) - \frac{f(x)-P(x)}{x^2(x-1)}t^2(t-1)$$

a) Vérifier que $\Psi(0) = \Psi'(0) = \Psi(1) = \Psi(x) = 0$ **(0.5 point)**

La vérification est évidente

b) Montrer que Ψ' s'annule en 3 points de $[0\ ,\ 1\ ]$ et Ψ'' s'annule en 2 points de $[0\ ,\ 1\ ]$

En appliquant Rolle à $\Psi(t)$:

avec $\Psi(0) = \Psi(x) = 0 \rightarrow \exists\ c_1 \in\ ]0, x[$ tq $\Psi'(c_1) = 0$

et $\Psi(x) = \Psi(1) = 0 \rightarrow \exists\ c_2 \in\ ]x, 1[$ tq $\Psi'(c_2) = 0$

De plus $\Psi'(0) = 0$ **(0.5 point)**

En appliquant Rolle à $\Psi'(t)$:

$\exists\ d_1 \in\ ]0, c_1[$ tq $\Psi''(d_1) = 0$

$\exists\ d_2 \in\ ]c_1, c_2[$ tq $\Psi''(d_2) = 0$ **(0.5 point)**

c) Et finalement Rolle nous donne : $\exists\ \theta \in\ ]d_1, d_2[\subset [0,1]$ tq $\Psi^{(3)}(\theta) = 0$

et comme $P^{(3)}(x) = 0$ et $(t^2(t-1))^{(3)} = 6$ on obtient : **(0.5 point)**

$$f(x) - P(x) = \frac{f^{(3)}(\theta)}{6}x^2(x-1)$$

5) On pose $M_3 = \max_{0 \le t \le 1} \left|f^{(3)}(t)\right|$,

a) Montrer que $|f(x) - P(x)| \le \dfrac{2M_3}{81} \quad \forall\ x \in [0,1]$ **(1 point)**

Il suffit d'étudier la fonction $x \rightarrow g(x) = \left|x^2(x-1)\right| = x^2(1-x)\ ,\ x \in [0,1]$

Or $g'(x) = 2x - 3x^2$ donc $g'(x) = 0$ si et ssi : $x = 0$ ou $x = \dfrac{2}{3}$

Ce qui implique que g admet un maximum qui est $g(\frac{2}{3}) = \frac{4}{9}\frac{1}{3} = \frac{4}{27}$

et finalement : $|f(x) - P(x)| \leq \frac{M_3}{6}\frac{4}{27} = \frac{2M_3}{81} \ \forall\, x \in [0,1]$

b) Montrer que : $\left|\int_0^1 (f(x) - P(x))dx\right| \leq \frac{M_3}{72}$ **(1 point)**

On a : $\left|\int_0^1 (f(x) - P(x))dx\right| \leq \int_0^1 |f(x) - P(x)|\, dx \leq \frac{M_3}{6}\int_0^1 x^2(1-x)dx$

Soit : $\left|\int_0^1 (f(x) - P(x))dx\right| \leq \frac{M_3}{6}(\frac{1}{3} - \frac{1}{4}) = \frac{M_3}{72}$

Exercice 7.11.2. : (6 points) :

A) On considère l'équation différentielle :

$(E1)$ $\quad \frac{dP}{dt} = rP(t)(1 - \frac{P(t)}{K})$ où $K \in \mathbb{R}_+^*$ et $r \in \mathbb{R}$ avec $P(0) \neq 0$ et $P(0) \neq K$.

1) En faisant un changement de variables, montrer que la résolution de $(E1)$ revient à celle de

$(E2)$ $\quad \frac{dy}{dt} = ry(t)(1 - y(t))$ où $K \in \mathbb{R}_+^*$ et $r \in \mathbb{R}$ **(0.5 point)**

Il suffit de prendre $P(t) = Ky(t)$ pour obtenir $(E2)$

2) Intégrer l'équation différentielle $(E2)$ (en supposant $y \neq 0$ et $y \neq 1.$) **(1.5 point)**

On peut intégrer $(E2)$ soit comme une equation de Bernouilli soit directement comme suit :

$\int \frac{dy}{y(1-y)} = \int r dt$ qui donne :

$Log\left|\frac{y}{1-y}\right| = rt + \widetilde{C} \rightarrow \frac{y}{1-y} = Ce^{rt}$

ou encore : $y(t) = \frac{Ce^{rt}}{(1 + Ce^{rt})} = \frac{C}{(C + e^{-rt})}$

3) Déduire la solution de l'équation différentielle $(E1)$ qui vérifie $P(0) = P_0$ (P_0 donné)

En revenant à $P(t) = Ky(t) = \frac{CK}{(C + e^{-rt})}$, on cherche C en prenant $P(0) = P_0$

Il vient donc : $P(0) = P_0 = \frac{CK}{(C+1)} \rightarrow (C+1)P_0 = CK \rightarrow C(K - P_0) = P_0$

D'où $C = \frac{P_0}{(K - P_0)}$ et $P(t) = \frac{K\frac{P_0}{(K-P_0)}}{(\frac{P_0}{(K-P_0)} + e^{-rt})} = \frac{KP_0}{P_0 + (K-P_0)e^{-rt}} (***)$ **(1 point)**

4) Chercher $\lim_{t \to +\infty} P(t)$ selon les valeurs de r **(1 point)**

L'équation $(***)$ nous donne :

$\lim\limits_{t \to +\infty} P(t) = P_0 \quad$ si $r = 0$

$\lim\limits_{t \to +\infty} P(t) = 0 \quad$ si $r < 0$

$\lim_{t \to +\infty} P(t) = K \quad$ si $r > 0$

B) La discrétisation de $(E2)$ conduit à la suite récurrente (u_n) définie par :

$0 < u_0 < 1$ et $u_{n+1} = h(u_n) = ru_n(1 - u_n) \quad \forall n \geq 0 \;\; r \in]0, 4[$

1) Trouver les points fixes θ_1 et θ_2 de h **(0.5 point)**

Les points fixes sont les solutions de $\theta = r\theta(1 - \theta)$

Soit $\theta_1 = 0$ ou $1 = r(1 - \theta)$ qui donne $\theta_2 = \dfrac{r-1}{r} = 1 - \dfrac{1}{r}$

2) A quelles conditions sur r la suite (u_n) converge-t-elle vers les points fixes θ_1 et θ_2?

La suite converge vers θ_1 si $|h'(0)| < 1$ c.a.d : $r < 1$ **(0.5 point)**

La suite converge vers θ_2 si $\left| h'(1 - \dfrac{1}{r}) \right| < 1$ c.a.d : $-1 < 2 - r < 1$

ou encore : $1 < r < 3$ **(1 point)**