Université Grenoble Alpes
2017

Cours MAT406

# Mathématiques assistées par ordinateur

Romain Joly

# Table des matières

**Motivations :**

- Les seules opérations exactes que l'on peut faire avec des nombres sont les additions, les soustractions, les multiplications et les divisions euclidiennes. Calculer $\ln 2$ ne peut donc pas se faire de façon exacte et d'ailleurs la définition du log ne fournit pas forcément un algorithme de calcul. Comment fonctionne la touche ln de la calculette ?
- Bien que l'on voit de nombreuses méthodes de résolution d'équations, de calcul d'intégrales ou de résolution d'équations différentielles, la plupart de ces problèmes ne peuvent être résolus exactement en situation réelle : une primitive est rarement exprimable en terme de fonctions usuelles. Comment obtenir une valeur raisonnable d'une intégrale concrète ? Notons au passage qu'il existe un théorème d'incomplétude en maths qui dit que, même si on rajoute plein de fonctions usuelles, on pourra toujours former avec une fonction dont la primitive n'est pas exprimable avec les fonctions usuelles.
- Bien sûr, ce problème existait bien avant les calculateurs électroniques. On ne sera donc pas surpris de voir dans ce cours des noms comme ceux de Newton, Euler ou Gauss.
- Le nom de cette UE est trompeur : l'échange entre mathématiques et ordinateur est à double sens. D'une part, l'ordinateur va nous aider à comprendre des aspects des mathématiques ou à faire des calculs complexes. D'autre part, les algorithmes de ces ordinateurs ont été développés à l'aide d'outils mathématiques pointus.

# Chapitre 1 : Nombres et erreurs

## 1 Représentation des nombres

Il existe de nombreuses façons de représenter les nombres. Notre écriture est basée sur l'écriture indienne importée via les arabes au moyen-âge. C'est une écriture positionnelle en base 10. Mais il nous reste dans la vie courante des vestiges de la base 60 des babyloniens et de la base 20 des gaulois. L'écriture de type romaine n'est pas du tout favorable au calcul, mais de nombreux humains utilisent une numération autre que positionnelle : les chinois et japonais n'ont pas de zéro et combinent, en écriture traditionnelle, chaque chiffre avec une indication de la puissance (102=« une centaine et deux »).

### 1.1 Les flottants

L'écriture décimale comprend une virgule fixe. Le problème est que la place que prend un nombre dans un ordinateur est limitée. Si on ne s'autorise que 10 chiffres, par exemple 5 avant et 5 après la virgule, on ne peut pas noter les grands nombres. Par ailleurs, un nombre comme 0,0000536847 est noté 00000,00005 et perd quasiment tous ses chiffres significatifs. Il est donc plus approprié d'utiliser une notation scientifique. Ainsi, le nombre précédent peut-être noté $5{,}36847 \times 10^{-5}$ ou $536{,}847 \times 10^{-7}$, $0{,}536847 \times 10^{-4}$ etc. La *représentation flottante normalisée* est la première : la virgule est placée juste derrière le premier chiffre non nul. Dans la représentation $5{,}36847 \times 10^{-5}$, la partie 5,36847 est appelée *mantisse* et $-5$ est l'*exposant*.

### 1.2 Le système IEEE 754

Même si on représente les nombres en notations flottantes, il reste plein de paramètres à fixer : base, précision... Le format le plus standard est celui de la norme IEEE 754 en simple précision (32 bits) ou double précision (64 bits). Nous ne parlerons ici que de la simple précision. Le nombre est codé sur 32 bits :

- Le premier bit est le signe : 0 pour + et 1 pour −.
- Suivent 8 bits pour coder l'exposant en base 2. L'exposant est le nombre codé (qui doit être non nul) à qui on soustrait 127. Il peut donc aller de $-127$ à 128 mais on verra que $-127$ et 128 sont des codes réservés pour des cas particuliers. L'exposant standard doit donc être entre $-126$ et 127.
- Enfin 23 bits codent la mantisse en base 2. Plus précisément, ils codent les chiffres après la virgule puisque le chiffre avant est forcément 1 en notation flottante normalisée.

Regardons un exemple :

$$\underbrace{0}_{\text{signe }+} \;\Big|\; \underbrace{10000011}_{\text{exposant } 131-127=4} \;\Big|\; \underbrace{10110000000000000000000}_{\text{mantisse 1,1011 en base 2}}$$

Il s'agit donc du nombre $1{,}1011 \times 2^4$ en base 2, c'est-à-dire 27 en base 10. Il existe aussi quelques codes particuliers :

- si l'exposant est nul, alors le nombre est dénormalisé et vaut $0{,}m \times 2^{-126}$ où $m$ est la mantisse. Ainsi, 0 est codé par que des zéros (sauf pour $-0$ qui a un 1).
- si l'exposant est 255 et la mantisse est nulle, alors le nombre code pour l'infini `Inf` ou `- Inf`.
- si l'exposant est 255 et la mantisse est non nulle, le nombre n'est pas correct et code `NaN` (« not a number »).

On peut donc coder les nombres entre environ $3{,}4 \times 10^{38}$ et $1{,}4 \times 10^{-45}$ avec une précision de l'ordre de 8 chiffres en base 10. On va voir plus bas qu'être conscient de ce fait peut être très important. Au passage, notons que le logiciel Maple code les nombres en base 10 et doit donc les convertir à chaque calcul envoyé au processeur. Cela ralentit grandement le logiciel.

Notons juste que le système en double précision (64 bits) est similaire mais avec respectivement 11 et 52 bits pour l'exposant et la mantisse.

# 2 Les erreurs

Les erreurs dans les calculs peuvent avoir plusieurs sources :

- la représentation des nombres en machine qui tronque le nombre à une certaine précision dans sa représentation en base 2,
- la méthode de calcul qui ne donne qu'une valeur approchée, par exemple quand on calcule une fonction comme l'exponentielle.

Si le nombre exact est $x$ et le nombre obtenu par le calcul $\tilde{x} = x + \delta x$, alors $|\delta x| = |x - \tilde{x}|$ est appelée *erreur absolue*. En règle générale, on est plutôt intéressé par l'*erreur relative* qui est $\left|\frac{\delta x}{x}\right| = \left|\frac{x-\tilde{x}}{x}\right|$. Quelques règles :

- Additionner ou soustraire ajoute les erreurs absolues. Multiplier ou diviser ajoute les erreurs relatives.
- On fera attention que faire la différence de deux nombres proches donne une grande erreur relative :

$$(1{,}001 \pm 10^{-4}) - (1{,}000 \pm 10^{-4}) = 0{,}001 \pm 2.10^{-4} \ .$$

- Les erreurs s'accumulent et sont multipliées si on les multiplie par de grands nombres.
- Les problèmes proches de problèmes singuliers, même s'ils sont théoriquement possibles, donnent de mauvais résultats par le calcul. Par exemple, inverser une matrice de déterminant quasi nul donnera un résultat avec de grandes erreurs. On parle de *mauvais conditionnement*.

On va regarder ci-dessous quelques problèmes classiques à garder en tête.

## 2.1 Sommation d'une série

La somme dans un ordinateur n'est pas une opération commutative ! Cela est dû à la troncation. Regardons un exemple en Xcas.

```
1 Digits:=4;
```
$$4$$
```
2 0.0005+0.0005
```
$$0.001$$
```
3 0.001+1
```
$$1.001$$
```
4 1+0.0005
```
$$1.0$$
```
5 1.0+0.0005
```
$$1.0$$

Nous comprenons donc qu'il ne faut pas sommer de petits nombres avec les grands mais d'abord les petits nombres entre eux puis les grands. L'expérience suivante montre que le résultat est bien différent suivant l'ordre de sommation.

```
1 a:=0; pour j de 1 jusque 100000 faire a:=evalf(a+1/j,5); ffaire;
```
$$(0.0,10.0)$$
```
2 a:=0; pour j de 100000 jusque 1 pas -1 faire a:=evalf(a+1/j,5);
ffaire;
```
$$(0.0,11.75)$$

On peut aussi noter que la série $\sum 1/n$ converge sur un ordi, ce qui paraît contradictoire avec la divergence en mathématique. Mais le résultat du calcul va dépendre de la précision de la machine et de l'ordre de sommation, ce n'est donc pas une somme qui aura un sens véritable même si la sommation nous renvoie un nombre.

## 2.2 Calcul de $\pi$

Inspiré de *Analyse numérique et équations différentielles* de Jean-Pierre Demailly.

Pour calculer $\pi$, on approche le cercle par des polygones à $2^{n+1}$ côtés. Pour calculer le périmètre de ces polygones, il suffit de calculer $\sin(\pi/2^n)$. Bien entendu, ce serait triché que d'utiliser une valeur de $\pi$ pour cela. On va donc calculer $\sin(\pi/2^n)$ par la formule

$$\sin\theta = \sqrt{\frac{1-\cos(2\theta)}{2}} = \sqrt{\frac{1-\sqrt{1-\sin(2\theta)^2}}{2}}\ .$$

On regarde donc la suite

$$x_1 = 1 \qquad x_{n+1} = \sqrt{\frac{1-\sqrt{1-x_n^2}}{2}}$$

et on doit avoir $\pi \simeq 2^n x_n$.

```
3 f(x):=sqrt((1-sqrt(1-x^2))/2)

4 a:=1;n:=5; pour j de 1 jusque (n-1) faire a:=f(a); ffaire;
2^n*a; evalf(pi)-2^n*a;
```

1.0,5.0,0.0980171403296,3.13654849055,0.00504416304371

```
5 a:=1;n:=10; pour j de 1 jusque (n-1) faire a:=f(a); ffaire;
2^n*a; evalf(pi)-2^n*a;
```

1.0,10.0,0.00306795676313,3.14158772545,$4.92814307052e-06$

```
6 a:=1;n:=20; pour j de 1 jusque (n-1) faire a:=f(a); ffaire;
2^n*a; evalf(pi)-2^n*a;
```

1.0,20.0,$2.99628227414e-06$,3.14182968189, − 0.000237028299409

```
7 a:=1;n:=30; pour j de 1 jusque (n-1) faire a:=f(a); ffaire;
2^n*a; evalf(pi)-2^n*a;
```

1.0,30.0,$4.21468485109e-08$,45.2548339959, − 42.1132413423

On voit que l'erreur devient finalement grande pour beaucoup d'itérations. La raison est que le calcul de $f(x)$ pour $x$ petit fait intervenir une différence du type $1-(1+\varepsilon)$. Le problème est donc mal conditionné. Si on écrit plutôt

$$x_{n+1} = \sqrt{\frac{1-\sqrt{1-x_n^2}}{2}} = \sqrt{\frac{1-(1-x_n^2)}{2(1+\sqrt{1-x_n^2})}} = \frac{x_n}{\sqrt{2(1+\sqrt{1-x_n^2})}}$$

on obtient

```
8 f(x):=x/(sqrt(2*(1+sqrt(1-x^2))))

9 a:=1;n:=30; pour j de 1 jusque (n-1) faire a:=f(a); ffaire;
2^n*a; evalf(pi)-2^n*a;
```

1.0,30.0,$2.92583615853e-09$,3.14159265359,$1.42108547152e-14$

## 2.3 Une troncation qui s'accumule

Inspiré de *Weisstein, Eric W. "Roundoff Error" From MathWorld–A Wolfram Web Resource.*

En 1982, la Bourse de Vancouver a créé un nouvel indice avec une valeur nominale de 1000. L'indice est mis à jour à chaque opération et tronqué au troisième chiffre après la virgule. Au bout de 22 mois, l'indice chiffre 524,881 alors qu'il aurait dû être de 1009,811. La raison est que la troncature engendre une erreur qui est toujours dans le même sens. L'accumulation de tant d'erreurs a conduit à une forte sous-évaluation. Notons que si on avait arrondi au plus proche, les erreurs aurait été dans les deux sens et le résultat moins

erroné. Pour avoir accumuler 500 en erreurs d'environ 0,0005, il faut $10^6$ erreurs. Si on utilise une marche aléatoire de $10^6$ pas de $\pm 0{,}00025$, on obtient un écart-type de 0,25.

## 2.4 Ariane V

Inspiré de *Weisstein, Eric W. "Roundoff Error" From MathWorld–A Wolfram Web Resource.*

Le 4 juin 1996, la première fusée Ariane V est lancée. Une partie de son système été repris sur Ariane IV et utilisait des entiers 16−bits. Mais la fusée Ariane V était bien plus puissante et les données dépassaient celles de Ariane IV. Après 37 secondes de vol, le système d'Ariane V convertit un entier 64−bits en un de 16−bits et produisit un *overflow*. Le code de l'overflow a été compris comme un entier et le système a modifié la trajectoire de la fusée, entraînant sa destruction.

## 2.5 Les missiles Patriot

Tiré de *Weisstein, Eric W. "Roundoff Error" From MathWorld–A Wolfram Web Resource.*

En 1991, pendant la Guerre du Golfe, une batterie américaine de missiles Patriot a échoué dans linterception dun missile Scud irackien causant 28 morts. Le problème est qu'une horloge comptait le temps en dixièmes de seconde et que 1/10 n'est pas un nombre rond en base 2. Dans un système 24−bits, 1/10 est codé avec une erreur d'environ $9{,}54 \times 10^{-8}$ secondes. La batterie ayant tourné pendant 100 heures, on obtient

$$9{,}54 \times 10^{-8} \times 100 \times 3600 \times 10 \simeq 0{,}34 \text{ secondes.}$$

A la vitesse d'un Scud, il s'agit d'une erreur de 500 mètres. Pour corriger cela, il a été décidé de réinitialiser régulièrement les batteries.

# 3 Séries et développements limités

Les seules fonctions qui se calculent se façon exacte sont les fractions rationnelles (quotient de polynômes). Pour les fonctions usuelles comme l'exponentielle ou le log, il faut faire un calcul approché, typiquement en utilisant le développement de Taylor.

**Théorème 1.1. Développement de Taylor avec reste intégral**
*Soit $I$ un intervalle de $\mathbb{R}$ et $f$ une fonction de classe $\mathcal{C}^{K+1}(I,\mathbb{R})$, alors pour tout $x_0 \in I$,*

$$\begin{aligned}\forall x \in I\ ,\ \ f(x) =& f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \dots \\ &+ \frac{f^{(K)}(x_0)}{k!}(x-x_0)^k + \int_{x_0}^{x} \frac{f^{(K+1)}(t)}{K!}(x-t)^K dt\ .\end{aligned}$$

***Démonstration :*** On fait une récurrence sur $k$ en utilisant que

$$\int_{x_0}^{x} \frac{f^{(k)}(t)}{(k-1)!}(x-t)^{k-1}dt = \left[-\frac{f^{(k)}(t)}{k!}(x-t)^k\right]_{x_0}^{x} - \int_{x_0}^{x} -\frac{f^{(k+1)}(t)}{k!}(x-t)^k dt\ .$$

□

Si on sait montrer que le reste tend vers 0 quand $K$ tend vers $+\infty$, alors on peut écrire

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \ .$$

On parle alors de *développement en série entière*. Pour utiliser correctement ce développement, il faut savoir quand arrêter la sommation, car on ne peut ajouter une infinité de nombres dans la vraie vie. Pour cela, on peut utiliser le reste intégral ci-dessus ou bien les deux résultats simples suivants.

**Proposition 1.2. Reste d'une série sous-géométrique**
*Soit une série $\sum_k c_k$ telle qu'il existe $M$ et $\lambda \in [0,1[$ tels que $|c_k| \leq M\lambda^k$. Alors la série est bien convergente et le reste vérifie*

$$R_K = \left| \sum_{K+1}^{\infty} c_k \right| = \left| \sum_{k=0}^{\infty} c_k - \sum_{k=0}^{K} c_k \right| \ \leq \ M\frac{\lambda^{K+1}}{1-\lambda} \ .$$

***Démonstration :*** La convergence vient de la comparaison de $|c_k|$ avec le terme générale d'une série géométrique convergente. Pour l'estimation du reste, on a

$$\left| \sum_{K+1}^{\infty} c_k \right| \leq \sum_{K+1}^{\infty} |c_k| \leq \sum_{K+1}^{\infty} M\lambda^k = M\frac{\lambda^{K+1}}{1-\lambda} \ .$$

□

**Proposition 1.3. Reste d'une série alternée**
*Soit $(c_k)$ une suite de nombres positifs décroissante et tendant vers 0. Alors $\sum_k (-1)^k c_k$ est une série convergente dont le reste vérifie*

$$\left| \sum_{k=K+1}^{\infty} (-1)^k c_k \right| \leq c_{K+1} \ .$$

***Démonstration :*** On vérifie facilement par récurrence que $u_n = \sum_{k=0}^{2n} (-1)^k c_k$ et $v_n = \sum_{k=0}^{2n+1} (-1)^k c_k$ sont deux suites adjacentes qui convergent donc vers une même limite. En outre, cette limite $\ell$ est coincée entre les deux suites et si $K = 2n$ est pair, on a $c_{K+1} = u_n - v_n$ et donc $|u_n - \ell| \leq c_{K+1}$ (et idem si $K$ impair). □

## 3.1 Application au calcul de $\ln 2$

On a

$$\ln(1+x) = \sum_{k=0}^{K} \frac{(-1)^{k+1}}{k} x^k + (-1)^K \int_0^x \frac{(x-t)^K}{(1+t)^{K+1}} \, dt \ .$$

On pourrait obtenir que

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

mais la convergence est lente puisqu'il s'agit d'une série alternée : pour avoir le résultat à $10^{-8}$ près, il faudrait $10^8$ termes. Il paraît plus judicieux d'utiliser que $\ln 2 = -\ln(1/2)$. On a alors un reste vérifiant

$$\left|(-1)^K \int_0^{-1/2} \frac{(-1/2-t)^K}{(1+t)^{K+1}}\,dt\right| \le 2\int_0^{1/2} \left|\frac{(1/2-s)}{(1-s)}\right|^K ds \le 2\int_0^{1/2} \left|1-\frac{1}{2(1-s)}\right|^K ds$$
$$\le \frac{1}{2^{K-1}} .$$

On obtient alors un calcul de $\ln 2$ à $10^{-8}$ près si on a $2^{K-1} \ge 10^8$, c'est-à-dire $K \ge 8\ln 10/\ln 2 + 1 \simeq 27{,}5$. On a donc

$$\ln 2 = \sum_{k=0}^{28} \frac{1}{k2^k} \quad \text{à } 10^{-8} \text{ près.}$$

On peut faire le test sous Xcas

```
10 a:=0.;pour j de 1 jusque 28 faire a:=a+1/(j*2^j); ffaire;
print(a); print(evalf(ln(2)));
a :0.693147180435
0.69314718056
```

## 3.2 Application au calcul de $e^{-10}$

Pour calculer la fonction exponentielle, on peut utiliser la formule de Taylor

$$e^x = \sum_{k=0}^{K} \frac{x^k}{k!} + \int_0^x \frac{e^t}{K!}(x-t)^K\,dt \ . \tag{1.1}$$

La série entière converge bien puisque

$$\left|\int_0^x \frac{e^t}{K!}(x-t)^K\,dt\right| \le e^{|x|}\frac{|x|^K}{K!} \xrightarrow[K\to+\infty]{} 0 \ .$$

En théorie, pour tout $x \in \mathbb{R}$, il suffit de tronquer la série au moment où le reste est suffisamment petit. Prenons par exemple $x = -10$. On a une série alternée convergente et donc le reste est de l'ordre du premier terme omis, soit $10^{N+1}/(N+1)!$. Pour que ce terme soit plus petit que $10^{-8}$, il faut aller jusqu'à environ $N = 40$. On obtient un résultat de l'ordre de $4.10^{-5}$ en ayant ajouté et soustrait des termes allant jusqu'à plus de 2755. On a vu que ce genre de calculs est très mauvais pour la précision et même si tout se passe bien, on a seulement 3 chiffres significatifs pour du 32−bits. On peut donc déjà dire qu'il vaut mieux utiliser $e^{-10} = 1/e^{10}$ et calculer $e^{10}$. Le problème est que le reste est de l'ordre de $10^K/K!$ qui met du temps à tendre vers 0. Pour calculer $e^{10}$, on peut aussi dire qu'il suffit de calculer $e^1$ puis de le mettre à la puissance 10. Cette puissance augmente de 10 l'erreur relative, donc on va chercher $e$ avec une précision d'environ $10^{-9}$, ce qui

est atteint quand $K! \geq 10^9$ avec $K = 13$ termes. Comparons la méthode naïve avec la méthode améliorée.

```
11 exp(-10)
```

$$4.53999297625e-05$$

```
12 a:=1; pour j de 1 jusque 100 faire a:=a+(-10)^j/j!; ffaire;
```

$$1.0,4.5399924373e-05$$

```
13 a:=1; pour j de 1 jusque 13 faire a:=a+1/j!; ffaire; a:=1/a^10;
```

$$1.0,2.71828182829,4.53999297645e-05$$

# Chapitre 2 : Équations non linéaires

De nombreux problèmes conduisent à résoudre une équation non-linéaire du type

$$f(x) = 0 \ .$$

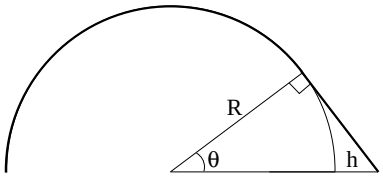
Prenons comme exemple dans ce chapitre les deux problèmes suivants.

*1) Calculer $\sqrt{2}$, c'est-à-dire résoudre $x^2 - 2 = 0$.*

*2) On tend une corde tout autour de la circonférence de la Terre. Cette corde est parfaitement tendue et plaquée sur la Terre, supposée ronde. On rallonge la corde de $1\,m$ et on la tire vers le haut. De combien peut-on la soulever ?*

Un peu de trigonométrie nous amène à chercher l'angle $\theta$ de la figure ci-contre et on obtient l'équation

$$\tan\theta - \theta = \frac{1}{2R} \quad \text{avec } R = 6{,}37 \times 10^6 \ .$$




N.B. : dans ce chapitre, on ne parlera que de fonctions de $\mathbb{R}$ dans $\mathbb{R}$ mais plusieurs des résultats et méthodes ont une généralisation dans $\mathbb{R}^d$.

## 1 La méthode de dichotomie

La méthode de dichotomie consiste littéralement à successivement *couper en deux* un intervalle par des recherches à tâtons. Il est basé sur le théorème des valeurs intermédiaires.

**Théorème 2.1. Les valeurs intermédiaires**
*Soit $f$ une fonction continue sur un intervalle $]a,b[ \subset \mathbb{R}$ qui admet des limites, éventuellement infinies, en $a$ et $b$ (eux aussi éventuellement infinis). Soit $y$ un réel strictement compris entre $\lim_{x\to a}$ et $\lim_{x\to b}$, alors il existe $c \in ]a,b[$ tel que $f(c) = y$.*

**Théorème 2.2. Les valeurs intermédiaires : version courte**
*Soit $f \in \mathcal{C}^0([a,b],\mathbb{R})$ telle que $f(a)f(b) < 0$, alors il existe $c \in ]a,b[$ tel que $f(c) = 0$.*

La méthode de Dichotomie consiste en l'algorithme suivant. Soit $f \in \mathcal{C}^0([a,b],\mathbb{R})$ telle que $f(a)f(b) < 0$. On pose $x_0 = a$ et $y_0 = b$. On définit deux suites $(x_n)$ et $(y_n)$ par la récurrence suivante : soit $z_n = (x_n + y_n)/2$ ; si $f(z_n) = 0$ on a trouvé une solution et on peut s'arrêter ; si $f(x_n)f(z_n) < 0$ alors $x_{n+1} = x_n$ et $y_{n+1} = z_n$ ; si $f(y_n)f(z_n) < 0$, alors $x_{n+1} = z_n$ et $y_{n+1} = y_n$.

**Proposition 2.3.** *Si la condition d'arrêt $f(z_n) = 0$ n'arrive pas, les suites $(x_n)$ et $(y_n)$ sont bien définies et convergent vers un zéro de $f$.*

***Démonstration :*** On vérifie facilement par récurrence que $f(x_n)$ et $f(y_n)$ sont toujours de signes opposés et donc que si $f(z_n) \neq 0$, soit $f(x_n)$ soit $f(y_n)$ sera de signe opposé à $f(z_n)$. Par ailleurs $|x_n - y_n| = 2^{-n}|x_0 - y_0|$ et donc les deux suites sont adjacentes et convergent vers un même point $\ell$. Par ailleurs, par les valeurs intermédiaires, il existe une suite $(c_n)$ de zéros de $f$ telle que $x_n < c_n < y_n$. Donc cette suite converge aussi vers $\ell$ et par continuité, $f(\ell) = 0$. □

La vitesse de convergence de cette méthode est linéaire.

**Définition 2.4.** *Une suite $(x_n)$ converge linéairement vers une limite $\ell$ s'il existe $C > 0$ tel que*

$$\forall n \in \mathbb{N} \ , \ \ |x_{n+1} - \ell| \leq C|x_n - \ell| \ .$$

La méthode de dichotomie double la précision à chaque étape. Il faut donc 3 à 4 étapes pour gagner un chiffre décimal significatif. Par exemple, 20 étapes de dichotomie permettent d'obtenir $\sqrt{2}$ avec 5 décimales exactes.

```
1 a:=0.;b:=2.;

2 pour j de 1 jusque 20 faire c:=(a+b)/2; si
((c^2-2)*(a^2-2))<0 alors a:=a;b:=c; sinon a:=c;b:=b; fsi;ffaire;
afficher(a);afficher(b);evalf(sqrt(2));
a :1.41421318054
b :1.41421508789

                    1.41421508789,1,1,1.41421356237
```

Résolvons maintenant notre second problème avec la méthode de dichotomie.

```
3 f(x):=tan(x)-x-1/(2*6.371*10^6);

       (x)->tan(x)-x-1/(2*6.371*10^6)

4 a:=0;b:=1.5;

5 c:=(a+b)/2; si (f(c)*f(a))<0 alors a:=a;b:=c; sinon a:=c;b:=b;
fsi; print(a);print(b);
a :0
b :0.75

6 pour j de 1 jusque 20 faire c:=(a+b)/2; si (f(c)*f(a))<0 alors
a:=a;b:=c; sinon a:=c;b:=b; fsi;ffaire; print(a);print(b);
a :0.00617480278015
b :0.00617551803589
```

Il nous faut donc 20 itérations pour obtenir que l'on peut soulever la corde d'environ $0{,}00617\,R \simeq 39\,\text{km}$.

◇ Avantages : simple et fonctionne toujours si on a trouvé un intervalle de départ.

◇ Inconvénients : assez lent ; ne donne aucune info sur le nombre de solutions ; trouver un intervalle de départ peut être compliqué.

# 2 Méthodes itératives

## 2.1 Fonctions contractantes et point fixe

Un des théorèmes fondamentaux de l'analyse est le suivant.

**Théorème 2.5.** *Soit $I$ un intervalle fermé de $\mathbb{R}$, c'est-à-dire du type $[a,b]$, $]-\infty,b]$, $[a,+\infty[$ ou bien $\mathbb{R}$. Soit $f : I \to I$ une fonction contractante de $I$ dans lui-même, c'est-à-dire qu'il existe $K \in [0,1[$ tel que*

$$\forall (x,y) \in I^2 \ , \ \ |f(x) - f(y)| \leq K|x - y| \ .$$

*Alors $f$ admet un unique point fixe $x^*$ et toute suite itérative $x_{n+1} = f(x_n)$ converge linéairement vers $x^*$.*

*En outre, on a une estimation de l'erreur donnée par*

$$|x_n - x^*| \leq \frac{K}{1-K}|x_n - x_{n-1}| \ .$$

***Démonstration :*** Remarquons d'abord que s'il existe deux points fixes $x^*$ et $y^*$ alors on doit avoir

$$|x^* - y^*| = |f(x^*) - f(y^*)| \leq K|x^* - y^*|$$

avec $K < 1$ et donc $x^* = y^*$. Pour l'existence et la convergence, prenons une suite itérative quelconque $x_{n+1} = f(x_n)$. On a par récurrence que $|x_{n+1} - x_n| \leq K^n|x_1 - x_0|$ donc pour tout $q \geq p \geq N$, on a

$$|x_p - x_q| \leq \sum_{n=p}^{q-1} |x_{n+1} - x_n| \leq \frac{K^p - K^q}{1-K}|x_1 - x_0| \leq \frac{C}{K}^N \ .$$

Comme $|K| < 1$, la suite est de Cauchy et converge. Notons bien que la limite $x^*$ est dans $I$ car $I$ est fermé. On a $x_{n+1} - x_n \to 0$ et donc $f(x_n) - x_n \to 0$. Par continuité de $f$ ($f$ est lipschitzienne), on a $f(x^*) = x^*$ et donc $x^*$ est un point fixe (qui est unique).

L'estimation de l'erreur vient de

$$|x_n - x^*| \leq \sum_{k=n}^{\infty} |x_k - x_{k+1}| \leq \sum_{k=1}^{\infty} K^k|x_n - x_{n-1}| = \frac{K}{1-K}|x_n - x_{n-1}| \ .$$

□

On peut utiliser ce théorème si on arrive à écrire notre équation sous la forme $g(x) = x$ avec $g$ contractant. Pour calculer $\sqrt{2}$, on peut par exemple chercher le point fixe de la fonction $f(x) = x - \frac{1}{10}(x^2 - 2)$ qui est contractante sur $[1,4]$.

```
7 a:=1.;g(x):=x-(x^2-2)/10; pour j de 1 jusque 20 faire a:=g(a);
ffaire; evalf(sqrt(2));
```

$$1.0,(\mathtt{x}) - > \mathtt{x} - (\mathtt{x}\hat{}\,2 - 2)/10,1.41354549995,1.41421356237$$

Pour notre deuxième problème, on pourrait écrire $g(x) = x$ avec $g = \tan -1/2R$ mais la tangente n'est pas contractante. On écrit donc $g(x) = x$ avec $g(x) = \arctan(x+1/2R)$ et cette fois-ci la fonction est contractante si on reste sur $\mathbb{R}_+$ car $g$ a une dérivée strictement plus petite que 1 en dehors de $-1/2R$.

```
8 a:=1;g(x):=arctan(x+1/(2*6.371*10^6)); pour j de 1 jusque 200000
faire a:=g(a); ffaire;
Temps mis pour l'évaluation : 1.14
```

$$0.00617694643091$$

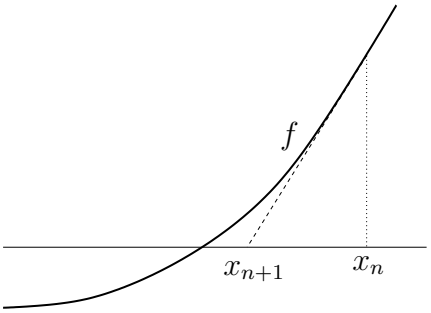
On voit ici que la fonction n'est pas suffisamment contractante car notre solution est proche d'un point où la dérivée de $g$ vaut 1. La convergence n'est pas rapide du tout et il faut beaucoup d'itérations.

◇ Avantages : plus simple à écrire que la dichotomie et pas de tests à faire
◇ Inconvénients : il faut arriver à mettre le problème sous la bonne forme.

## 2.2 Méthode de Newton

La méthode de Newton consiste à l'itération ci-contre : si $x_n$ n'est pas très loin d'un zéro de $f$, alors on tire la tangente en $x_n$ et celle-ci coupe l'axe horizontal en un point $x_{n+1}$ plus proche du zéro de $f$. Le calcul montre que

$$x_{n+1} = \varphi(x_n) := x_n - \frac{f(x_n)}{f'(x_n)} \ .$$




**Proposition 2.6.** *Soit $f$ une fonction de classe $\mathcal{C}^2$ au voisinage d'un point $x^*$ tel que $f(x^*) = 0$ et $f'(x^*) \neq 0$. Alors, pour $x_0$ proche de $x^*$, la suite itérative $x_{n+1} = \varphi(x_n)$ est bien définie et converge quadratiquement vers $x^*$ dans le sens où il existe $C > 0$ tel que*

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^2 \ .$$

***Démonstration :*** Comme $f'(x^*) \neq 0$ et que $f$ est de classe $\mathcal{C}^2$, on a $f'(x) \neq 0$ proche de $x^*$ et donc la division par $f'$ ne sera pas un problème si on reste proche de $x^*$.

On a

$$\varphi'(x) = 1 - \frac{f'(x)^2 - f''(x)f(x)}{f'(x)^2} = \frac{f''(x)f(x)}{f'(x)^2} \ .$$

Comme $f$ est de classe $\mathcal{C}^2$ et s'annule en $x^*$, dans un voisinage de $x^*$, on a

$$\varphi'(x) = \frac{f''(x^*) + o(1))}{f'(x^*)^2 + o(1)}(f'(x^*) + o(1))(x - x^*) = \frac{f''(x^*)}{f'(x^*)}(x - x^*) + o(x - x^*) \ .$$

Donc, dans ce voisinage, $|\varphi'(x)| \leq C|x-x^*|$ et en intégrant $|\varphi(x)-x^*| \leq \frac{C}{2}|x-x^*|^2$ puisque $\varphi(x^*) = x^*$. On obtient donc que si $x$ est à distance $r$ assez petite de $x^*$, alors $\varphi(x)$ est à distance $Cr^2/2$ qui est strictement plus petite que $r$ si $r$ est choisi assez petit. On a donc une suite itérative qui reste dans le même voisinage de $x^*$ et converge quadratiquement vers lui. □

Notons que la convergence quadratique signifie que l'on double le nombre de chiffres significatifs à chaque étape : c'est très rapide. La méthode de Newton demande souvent que 3 ou 4 étapes, pourvu que l'on ait une idée de la solution $x^*$ cherchée et qu'il ne s'agit pas d'un zéro multiple.

Dans certains cas, on peut définir une zone de convergence.

**Proposition 2.7.** *Soit $f$ une fonction de classe $\mathcal{C}^2(\mathbb{R},\mathbb{R})$ et soit $x^*$ une racine de $f$ telle que $f'(x^*) > 0$. On suppose qu'il existe un intervalle $[x^*,a[$ sur lequel $f''(x) \geq 0$, c'est-à-dire que $f$ est convexe. Alors pour tout point de départ $x_0 \in [x^*,a[$, la méthode de Newton converge vers $x^*$.*

*Cette proposition est aussi vraie pour un intervalle à droite avec $f$ concave et localement décroissante et pour un intervalle à gauche avec $f$ convexe et localement décroissante ou concave et localement croissante.*

***Démonstration :*** Notons tout d'abord que $f'(x) \geq f'(x^*) > 0$ sur l'intervalle $[x^*,a[$ et que $f(x) > f(x^*) = 0$. La fonction $\varphi(x) = x - f(x)/f'(x)$ y est donc bien définie et la suite $x_{n+1} = \varphi(x_n)$ est strictement décroissante tant qu'elle reste dans cet intervalle. Le point clef est donc de montrer que si $x \in [x^*,a[$ alors $\varphi(x) > x^*$. En effet, la suite $x_{n+1} = \varphi(x_n)$ serait alors toujours strictement décroissante et minorée par $x^*$ donc aurait une limite $x_\infty$ qui doit vérifier $x_\infty = \varphi(x_\infty)$, ce qui n'est possible que si $x_\infty = x^*$.

On a

$$\begin{aligned}\varphi(x) = x - \frac{f(x)}{f'(x)} &= x - \frac{f(x^*) + \int_{x^*}^{x} f'(\xi)d\xi}{f'(x)} \\ &= x - \int_{x^*}^{x} \frac{f'(\xi)}{f'(x)}d\xi \;>\; x - \int_{x^*}^{x} 1\, d\xi \\ &> x - (x - x^*) = x\end{aligned}$$

ce qu'il fallait démontrer.

Les autres cas de la proposition se déduisent par des symétries horizontales et/ou verticales. □

On peut aussi avoir une estimation de l'erreur commise, qui est utile si on a une estimation de la dérivée, typiquement pour des équations polynomiales.

**Proposition 2.8.** *Soit un intervalle $I$ et soit $f \in \mathcal{C}^1(I,\mathbb{R})$ avec $|f'(x)| \geq \alpha > 0$ sur $I$. Supposons qu'il existe $x^* \in I$ tel que $f(x^*) = 0$. Alors, pour tout $x \in I$, $|x-x^*| \leq |f(x)|/\alpha$.*

***Démonstration :*** On a

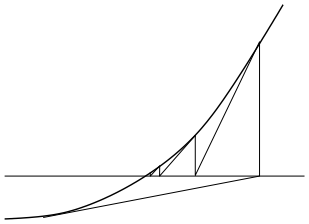
$$f(x) = f(x^*) + \int_{x^*}^{x} f'(\xi)d\xi = \int_{x^*}^{x} f'(\xi)d\xi \ .$$

Comme $f'$ ne s'annule pas sur $I$, on peut supposer $f' > 0$ quitte à changer $f$ en $-f$. Dans ce cas,

$$|f(x)| \geq |x^* - x| \min_{\xi \in [x^*, x]} |f'(\xi)| = \alpha |x^* - x| \ .$$

□

On peut par exemple appliquer cette convergence pour trouver la plus grande racine d'un polynôme scindé à racine simple. Notons que si on est du mauvais côté de la racine mais assez proche, une itération de l'algorithme nous ramène dans la bonne zone.



Une application efficace et classique est le calcul de la racine carrée. On part de $x_0 = a$ et on pose

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{x_n}{2} - \frac{a}{2x_n} \ .$$

On retrouve le fameux et très ancien algorithme de Babylone (ou méthode de Héron).

```
9 phi(x):=x/2+1/x; a:=1.; pour j de 1 jusque 3 faire
a:=phi(a);ffaire; evalf(sqrt(2));

                (x)->x/2+1/x   ,1.0,1.41421568627,1.41421356237
```

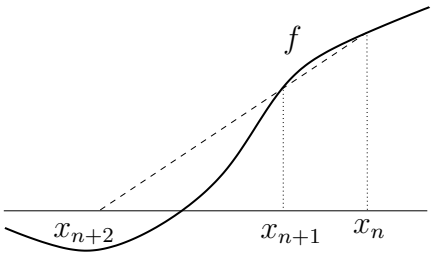
Dans le cas de notre équation modèle, c'est très raisonnable de penser que notre solution est proche de 0 mais nous sommes encore dans un cas problématique car $f'(0) = 0$. La convergence mettra donc du temps à débuter mais sera ensuite bien plus rapide que la dichotomie.

```
10 phi(x):=x-f(x)/(tan(x)^2); a:=1.; pour j de 1 jusque 20 faire
a:=phi(a);ffaire;
                              0.00617483954499
```

◇ Avantages : en général extrêmement rapide et on peut toujours tester si l'algorithme marche même sans aucun a priori.
◇ Inconvénients : le plus souvent, il faut une idée du point de départ ; l'algorithme peut complètement diverger et il converge moins bien dans les cas dégénérés.

## 2.3 Méthode de la sécante

Le problème de la méthode de Newton est qu'on a besoin de connaître la dérivée de la fonction. On pourrait en faire une approximation numérique, mais dans ce cas il y a une méthode plus simple : celle de la sécante. On part de deux points $x_0$ et $x_1$ et on construit ensuite par récurrence la suite $(x_n)$ : le point $x_{n+2}$ est l'intersection de la sécante conduite par $x_n$ et $x_{n+1}$ et de l'axe horizontal, c'est-à-dire



$$x_{n+2} = x_{n+1} - \frac{x_{n+1} - x_n}{f(x_{n+1}) - f(x_n)} f(x_{n+1}) \ .$$

Notons que si $x_{n+1}$ et $x_n$ sont proches, on retrouve une approximation de la méthode de Newton mais c'est alors qu'apparaissent des soustractions de nombres proches sources de grandes erreurs ou bien une division par zéro.

```
11 a:=0.;b:=1.; pour j de 1 jusque 5 faire c:= b -
(a-b)/(a^2-b^2)*(b^2-2); a:=b;b:=c;ffaire;
```

0.0,1.0,1.41421143847

```
12 a:=0.;b:=1.; pour j de 1 jusque 9 faire c:= b -
(a-b)/(a^2-b^2)*(b^2-2); a:=b;b:=c;ffaire;
```

0.0,1.0,*undef*

◇ Avantages : ceux de la méthode de Newton et pas besoin de connaître la dérivée.
◇ Inconvénients : ceux de la méthode de Newton et problème d'imprécisions vers la fin.

# 3 Équations polynomiales

On a vu différentes méthodes pour résoudre des équations non-linéaires. Mais ces méthodes ne sont pas systématiques. Le but de cette partie est de rendre au maximum systématique la résolution dans $\mathbb{R}$ des équations polynomiales

$$x^p + a_{p-1}x^{p-1} + a_{p-2}x^{p-2} + \ldots + a_2x^2 + a_1x + a_0 = 0 \ .$$

## 3.1 Rappels historiques

**Degré 1 :** équations linéaires... c'est facile.

**Degré 2 :** la résolution systématique des équations de degré 2 dans $\mathbb{R}$ nous vient des indiens (Sridhar Acharya, VIIIème siècle). Elle est popularisée par le grand mathématicien arabe Al-Khwârizmî (780-850) et son ouvrage *Abrégé du calcul par la restauration et la comparaison* qui ont donné les mots *algorithme* et *algèbre*.

**Degré 3 :** la méthode a été divulguée dans le *Ars Magna* de Girolamo Cardano (Jérôme Cardan) en 1545. Elle aurait été volée au mathématicien Tartaglia qui en serait le véritable découvreur. Il est difficile d'écrire des formules directes mais le principe de résolution est le suivant.

On part de l'équation $x^3 + ax^2 + bx + c = 0$ et en posant $y = x + a/3$, $p = b - a^2/3$ et $q = c + (2a^3 - 9ab)/27$, on se ramène à la forme réduite

$$y^3 + py + q = 0 \ .$$

En posant $y = z - p/3z$ et en multipliant par $z^3$, on trouve

$$z^6 + qz^3 - \frac{p^3}{27} = 0 \ .$$

Il s'agit d'une équation en $z^3$ de degré deux que l'on peut résoudre. Après avoir obtenu $z$, on revient à $y$ puis $x$.

Il faut noter que le passage par les complexes est nécessaire, même si les racines sont réelles. Ceci explique pourquoi il y a bien trois solutions (alors qu'on pourrait ne penser qu'à deux solutions) car il faut résoudre $z^3 = \lambda$ dans $\mathbb{C}$.

**Degré 4 :** la résolution des équations de degré quatre a été développée par Ferrari, un élève de Cardan, vers 1540. Elle consiste aussi à faire des transformations pour se ramener à une équation de degré inférieur (ici trois) et à appliquer alors la méthode connue.

**Degré $\geq 5$ :** suite aux travaux de Lagrange, Abel et Galois ont montré qu'il n'existe pas de méthode de résolution pour les équations de degré $\geq 5$. On notera que Galois a développé son mémoire à 18 ans avant de mourir en duel à 20 ans. Ce travail sera à l'origine de la théorie des groupes.

## 3.2 Indices et racines : la méthode de Budan-Fourier

On a vu que les équations polynomiales sont en général impossibles à résoudre par des formules, ou simplement la formule est complexe d'utilisation (et demande d'extraire des racines carrées, ce qui n'est pas non plus un calcul informatique standard). On va donc chercher ces racines par un algorithme numérique. Le problème est que la méthode de Newton ou la dichotomie n'assure pas de trouver toutes les racines.

On peut déjà localiser grossièrement les racines.

**Théorème 2.9. Cauchy**
*Soit $P(X) = X^n + a_{n-1}X^{n-1} + \ldots + a_0$ dans $\mathbb{C}[X]$ et soit $\rho = 1 + \max(|a_k|)$. Alors toute racine complexe $z$ de $P$ vérifie $|z| < \rho$.*

***Démonstration :*** Soit $z$ tel que $|z| \geq \rho$. On a

$$\begin{aligned} |P(z)| &\geq |z^n| - |\sum_{k=0}^{n-1} a_k z^k| \;\geq\; |z^n| - \max(|a_k|) \sum_{k=0}^{n-1} |z|^k \;\geq\; |z^n| - (\rho - 1)\frac{1-|z|^n}{1-|z|} \\ &\geq |z^n| - (|z|^n - 1) = 1 \end{aligned}$$

Ce qui montre que $z$ ne peut être racine □

On sait maintenant que nos racines sont entre $-\rho$ et $\rho$. Pour les localiser mieux, on va utiliser un indice. Pour un polynôme $P$ de degré $d$, on introduit la fonction signe comme la fonction

$$\begin{aligned} V_P(x) &= \sum_{k=1}^{d} \frac{1}{2} \left| \text{sign}(P^{(k)}(x)) - \text{sign}(P^{(k-1)}(x)) \right| \\ &= \text{nombre chgt de signe dans la suite } (P(x), P'(x), P''(x), \ldots, P^{(d)}(x)) \end{aligned}$$

On a alors le résultat suivant

**Théorème 2.10. Budan 1807, Fourier 1820**
*Soit $P \in \mathbb{R}[X]$ de degré $d$. On note $m(a,b)$ le nombre des racines de $P$ dans $]a,b]$, comptées avec leur multiplicités. Alors, on a*

$$m(a,b) \leq V_P(a) - V_P(b)$$

*et l'écart est forcément un nombre pair.*

*On a l'égalité $m(a,b) = V_P(a) - V_P(b)$ pour tout $a < b$ dans $\mathbb{R}$ si et seulement si $P$ est de degré $d$ et admet $d$ racines réelles.*

On peut alors utiliser un algorithme de type dichotomie pour localiser les racines. L'avantage est que si $P$ s'annule deux fois sur $]a,b[$, alors les deux racines sont détectées alors que la dichotomie simple ne voit pas de changement de signe.

La démonstration du théorème s'appuie sur plusieurs constats simples :

1. $V_P(a) - V_P(c) = (V_P(a) - V_P(b)) + (V_P(b) - V_P(c))$ donc on peut se ramener à des petits intervalles ne contenant qu'au plus une racine multiple.

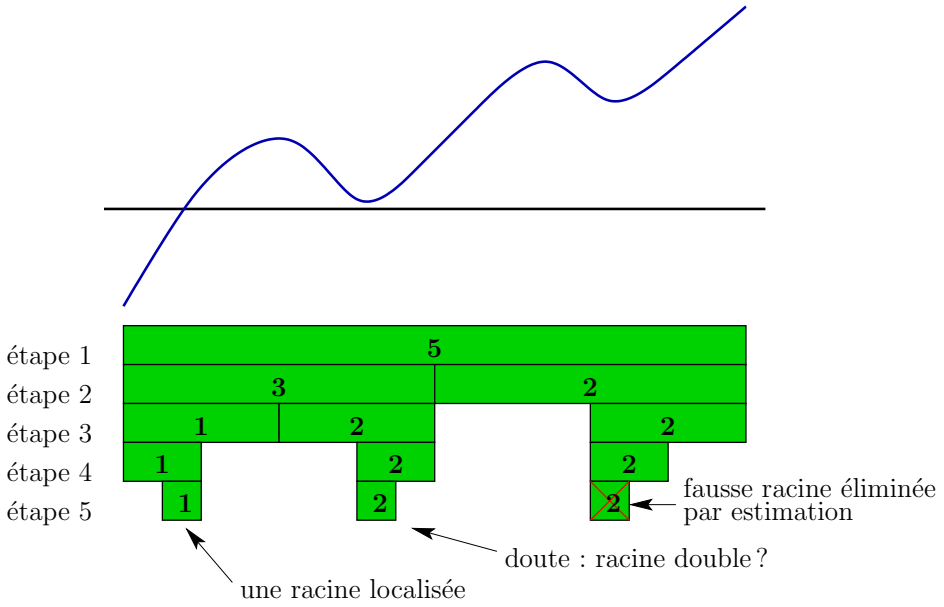
2. Si $P$ ne s'annule pas entre $a$ et $b$, comme $P^{(d)}$ est une constante, alors le changement de $V_p$ entre $a$ et $b$ est pair. Ceci explique que le défaut de l'indice est forcément pair.

3. Le passage de signes $(+,+,+)$ à $(+,-,+)$ entre $a$ et $b$ n'est pas possible car si $P^{(k)}$ est positif, alors $P(k-1)$ est croissant. Idem pour la version symétrique. Donc l'indice $V_p$ ne peut augmenter entre $a$ et $b$ : $V_P(a) - V_P(b)$ est toujours positif et jamais un intervalle ne pourra compter un nombre négatif et cacher une racine voisine.

4. Une racine de multiplicité paire ne provoque pas de changement de signe alors qu'une racine de multiplicité impaire le fait. Pour une racine multiple, la multiplicité de cette racine dans les dérivés alterne entre pair et impair.

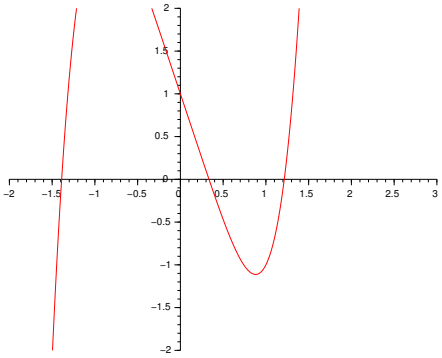
5. Si $P$ a une racine de multiplicité $m$ impaire entre $a$ et $b$, alors le signe de $P$, $P''$,..., $P^{(m-1)}$ change entre $a$ et $b$ et on a $m$ changement des signes dans le calcul de $V_P$. Si $P$ a une racine de multiplicité $m$ paire entre $a$ et $b$, alors le signe de $P'$, $P^{(3)}$,..., $P^{(m-1)}$ change entre $a$ et $b$ et on a $m$ changement des signes dans le calcul de $V_P$.

Il est possible que l'indice de Budan-Fourier sur-évalue le nombre de racines. Notons que dans le cas où $P$ a autant de racines que son degré, on va exactement trouver toutes les racines ainsi. Si $P$ n'a aucune racine dégénérée, c'est-à-dire multiple, alors on aura la position des racines une fois les intervalles assez petits : s'il reste des intervalles $[a,b]$ avec un indice pair, ce sont des artefacts. Si on a aucune information, on peut éventuellement réussir à les éliminer par le calcul de $P(a)$ et une majoration grossière de $P'$ sur $[a,b]$ par les coefficients.




Faisons un exemple avec le polynôme $P = X^5 - 3X + 1$ qui n'a que trois racines réelles.

| $x$ | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| $P = X^5 - 3X + 1$ | - | + | + | - | + | + |
| $P' = 5X^4 - 3$ | + | + | - | + | + | + |
| $P = 20X^3$ | - | - | 0 | + | + | + |
| $P = 60X^2$ | + | + | 0 | + | + | + |
| $P = 120X$ | - | - | 0 | + | + | + |
| $P = 120$ | + | + | + | + | + | + |
| $V_P(x)$ | 5 | 4 | 2 | 1 | 0 | 0 |



Il y a trois racines sûres dans $]-2,-1[$, $]0,1[$ et $]1,2[$ et peut-être deux racines dans $]-1,0[$ (en fait non).

## 3.3 Indices et racines : la méthode de Sturm

Il existe un indice similaire donnant exactement le nombre de racine, dû à Sturm. Il utilise non pas la suite des dérivées mais la suite des polynômes intervenant dans l'algorithme d'Euclide du calcul du PGCD de $P$ et $P'$.

Soit $P$ un polynôme de degré $d$. On définit la suite de polynômes $(A_k)$ par $A_0 = P$, $A_1 = P'$ puis les polynômes suivants se trouvent en prenant l'opposé du reste de la division euclidienne des deux précédents :

$$A_k = A_{k+1}Q_{k+2} - A_{k+2} \quad \text{avec } d(A_{k+2}) < d(Q_{k+2}) \ .$$

Soit $A_p$, le dernier reste non nul : c'est le $PGCD$ de $P$ et $P'$ (algorithme d'Euclide). Cest un polynôme constant si $P$ n'a pas de racine multiple. Sinon, il faut diviser tous les $A_k$

par $A_p$. On définit maintenant les changements de signes de la suite comme plus haut

$$\begin{aligned}\tilde{V}_P(x) &= \sum_{k=1}^{p} \frac{1}{2} \left|\text{sign}(A_k(x)) - \text{sign}(A_{k-1}(x))\right| \\ &= \text{nombre chgt de signe dans la suite } (A_0(x), A_1(x), \ldots, A_p(x))\end{aligned}$$

**Théorème 2.11. Sturm (1829)**
*Le nombre de racines réelles de $P$ dans l'intervalle $]a,b]$ est égal à $\tilde{V}_P(a) - \tilde{V}_P(b)$.*

Faisons un exemple avec le polynôme $P = X^5 - 3X + 1$ qui n'a que trois racines réelles alors que l'indice de Budan-Fourier en avait détecté cinq. Pour cela, il faut effectuer les divisions euclidiennes. Pour calculer $A_2$, on fait la division de $P$ par $P'$.

$$\begin{array}{rl|l} X^5 - 3X & +1 & 5X^4 - 3 \\ -\frac{12}{5}X & +1 & \hline \frac{1}{5}X \end{array} \tag{2.1}$$

Donc $A_2 = \frac{12}{5}X - 1$, puis on divise $P'$ par $A_2$ et on trouve $A_3 = \frac{59083}{20736}$ qui est le dernier non nul.

| $x$ | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| $A_0 = P = X^5 - 3X + 1$ | - | + | + | - | + | + |
| $A_1 = P' = 5X^4 - 3$ | + | + | - | + | + | + |
| $A_2 = \frac{12}{5}X - 1$ | - | - | - | + | + | + |
| $A_3 = \frac{59083}{20736}$ | + | + | + | + | + | + |
| $\tilde{V}_P(x)$ | 3 | 2 | 2 | 1 | 0 | 0 |

On obtient bien cette fois exactement la localisation des trois zéros.

## 3.4 Recherche des racines rationnelles

Soit

$$P = a_d X^d + a_{d-1} X^{d-1} + \ldots + a_1 X + a_0 \ \in \mathbb{Z}[X]$$

un polynôme avec des coefficients entiers (ce qui est le cas de la majorité des polynômes que l'on écrit en général). On peut être intéressé par les racines rationnelles, en particulier pour écrire une factorisation exacte (ce qu'on ne peut pas faire avec des valeurs approchées des racines). Cela peut aussi servir à trouver suffisamment de racines pour faire descendre le degré jusqu'à 4, à partir duquel on a des méthodes exactes. On peut se ramener à un nombre fini de tests par le constat suivant. Si $p/q$ est une racine sous forme de fraction irréductible, alors

$$P\left(\frac{p}{q}\right) = a_d \left(\frac{p}{q}\right)^d + \ldots + a_1 \frac{p}{q} + a_0 = 0$$

et donc

$$a_d p^d + a_{d-1} p^{d-1} q + a_{d-2} p^{d-2} q^2 + \ldots + a_1 p q^{d-1} + a_0 q^d = 0 \ .$$

On en déduit que $p$ divise $a_0 q^d$ puis, comme $p/q$ est irréductible, que $p$ divise $a_0$. De même $q$ divise $a_d$. On trouve donc un nombre fini de possibilités pour $p$ et $q$, qu'il reste à toutes

tester. Notons qu'il peut y avoir beaucoup de tests à faire et que factoriser des grands nombres est un problème difficile. Il existe donc des améliorations de ce principe pour les applications concrètes.

**Exemple :** on considère le polynôme

$$P = 4X^3 - 15X^2 + 25X - 12 \ .$$

Si $p/q$ est racine, alors $p$ divise 12 et $q$ divise 4. On a donc comme possibilités

$$1 \ , \ 2 \ , \ 3 \ , \ 4 \ , \ 6 \ , \ 12 \ , \ \frac{1}{2} \ , \ \frac{3}{2} \ , \ \frac{1}{4} \ , \ \frac{3}{4} \quad \text{et leur opposés.}$$

Ces 20 possibilités testées, on trouve que $3/4$ est racine et on obtient la factorisation

$$P = 4\left(X - \frac{3}{4}\right)(X^2 - 3X + 4)$$

qui montre qu'il n'y a pas d'autres racines réelles.

# Chapitre 3 : Algèbre linéaire

## 1 Remarques sur les erreurs et les coûts

On munit $\mathbb{R}^n$ de la norme euclidienne

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \ .$$

Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice, on appelle *norme triple de A* le nombre

$$|||A||| = \max_{\|x\|=1} \|Ax\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \ .$$

Par définition, on a $\|Ax\| \leq |||A|||.\|x\|$ pour tout vecteur $x$.

**Exemples :**

— la matrice $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ a pour norme triple $|||A||| = 2$ puisque $Ax = 2x$.

— la matrice $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ a pour norme triple $|||A||| = 2$ aussi.

Si $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, alors

$$\langle x|y \rangle = x_1 y_1 + \ldots + x_n y_n$$

demande $2n - 1$ opérations. L'erreur absolue est multipliée par $n$ et par le max des $|x_k|$ et $|y_k|$. L'erreur relative peut devenir très mauvaise si le produit est quasiment nul.

De même, si $A = (a_{ij})$ est une matrice, le calcul de $Ax$ demande $(2n-1) \times n$ opérations et l'erreur relative peut être mauvaise en cas de grand coefficients de $A$ ou si $Ax$ est quasiment nul. Plus précisément :

**Proposition 3.1.** *Soit $x \in \mathbb{R}^n$ connu avec une erreur $\delta x$ et soit $y = Ax$ et $(y + \delta y) = A(x + \delta x)$ avec $A \in \mathcal{M}_n(\mathbb{R})$ inversible. Alors*

$$\frac{\|\delta y\|}{\|y\|} \leq |||A|||.|||A^{-1}||| \ \frac{\|\delta x\|}{\|x\|} \ .$$

*Le nombre $cond(A) = |||A|||.|||A^{-1}|||$ est appelé le conditionnement de A.*

***Démonstration :*** On a $A\delta x = \delta y$ par linéarité. Donc $\|\delta y\| \leq |||A|||.\|\delta x\|$. D'autre part, $x = A^{-1}y$ donc $\|x\| \leq |||A^{-1}|||.\|y\|$, ce qui conclut. □

Le calcul du déterminant par la formule

$$\det(A) = \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma) a_{1,\sigma(1)} \dots a_{n,\sigma(n)}$$

demande $n.n!$ opérations. Pour $n = 100$ et les ordinateurs actuels, il faut plus que l'âge de l'univers en temps de calcul ! Le calcul par le développement sur les lignes ou colonnes conduit au même problème (coût au rang $n$ est $n$ fois le coût au rang $n-1$). S'en suit qu'inverser $A$ par

$$A^{-1} = \frac{1}{\det(A)} {}^t\mathrm{com}(A)$$

est une mauvaise idée.

Pour calculer les valeurs propres d'une matrice, calculer le polynôme caractéristique de façon naïve n'est pas forcément une bonne idée : si on n'a pas de logiciel de calcul formel, cela ne peut se faire directement et si on passe par Newton ou quelque chose du genre, il faut recalculer un déterminant à chaque itération.

# 2 Pivot de Gauss et applications

On ne va pas rappeler en quoi consiste le pivot de Gauss. Nous allons juste considérer qu'il s'agit de faire des opérations sur les lignes d'une matrice $A$ parmi :

i) permuter deux lignes,
ii) ajouter $\mu$ fois une ligne à une ligne plus bas,
iii) multiplier une ligne par $\lambda \neq 0$,

pour arriver à une matrice triangulaire supérieure, c'est-à-dire de la forme

$$U = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & * \end{pmatrix}.$$

**Avertissement numérique :** la théorie nous dit que le pivot doit être un coefficient non nul. La pratique veut que ce coefficient ne doit pas non plus être petit. Prendre le plus grand (en valeur absolue) est le meilleur choix.

En voici une illustration. Soit $\varepsilon > 0$ tel que $1 + \varepsilon$ soit indifférentiable de 1 par la précision machine. On veut résoudre le système

$$\begin{cases} \varepsilon x + y &= 1 \\ x + 2y &= 3 \end{cases}$$

qui a pour solution $(x,y) = \left(\frac{1}{1-2\varepsilon}, \frac{1-3\varepsilon}{1-2\varepsilon}\right) \simeq (1,1)$. Si on utilise $\varepsilon x$ comme pivot, en éliminant $x$ dans la deuxième ligne, on obtient

$$\begin{cases} x &= \frac{1}{\varepsilon}(1-y) \\ (2-\frac{1}{\varepsilon})y &= 3 - \frac{1}{\varepsilon} \end{cases}$$

ce qui conduit à $y = 1$ puis $x = 0$, ce qui est très mauvais. Si maintenant on prend comme pivot le terme $x$ de la deuxième ligne, on aurait

$$\begin{cases} x &= 3 - 2y \\ (1-2\varepsilon)y &= 1 - 3\varepsilon \end{cases}$$

ce qui conduit à $y \simeq 1$ puis $x \simeq 1$, ce qui est la réponse attendu. En pratique, même si $1+\varepsilon \neq 1$ dans l'ordinateur, on perdra beaucoup en précision en prenant le mauvais pivot.

**Methode LU :** on a vu que le pivot permet de transformer $A$ en une matrice $U$ triangulaire supérieure. Remarquons maintenant que les opérations sur les lignes peuvent se traduire matriciellement. Ainsi, multiplier la ligne $i$ par $\lambda \neq 0$ revient à multiplier la matrice à gauche par

$$L_i(\lambda) = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \lambda & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} -- i\text{ème ligne}$$

De même, ajouter $\mu$ fois la ligne $j$ à la ligne $i < j$ revient à multiplier la matrice à gauche par

$$M_{i,j}(\mu) = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \mu & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} -- i\text{-ème ligne, } j\text{-ème colonne}$$

Donc, si on peut appliquer le pivot de Gauss à une matrice $A$ pour la transformer en une matrice triangulaire supérieure $U$, cela veut dire que $K_1 K_2 \dots K_p A = U$ avec $K_m$ un $L_i(\lambda)$ ou un $M_{i,j}(\mu)$. Or pour $\lambda \neq 0$, $L_i(\lambda)$ est inversible d'inverse $L_i(1/\lambda)$ et $M_{i,j}(\mu)$ est inversible d'inverse $M_{i,j}(-\mu)$ (il suffit de raisonner en opérations inverses sur les lignes). On en conclut que $A = N_1 N_2 \dots N_p U$ avec $N_m$ un $L_i(1/\lambda)$ ou un $M_{i,j}(-\mu)$. Comme le produit de matrices triangulaires inférieures est une matrice triangulaire inférieure, on obtient la décomposition $A = LU$ avec $L$ triangulaire inférieure et $U$ triangulaire supérieure.

Si on doit changer d'ordre les lignes pour prendre un meilleur pivot, il suffit de les permuter avant le pivot, c'est-à-dire de multiplier $A$ pour une matrice de permutation $P_\sigma = (p_{ij})$ avec $p_{ij} = 1$ si $i = \sigma(j)$ et 0 sinon. On obtient alors une décomposition $A = PLU$ avec $L$ triangulaire inférieure, $U$ triangulaire supérieure et $P = P_{\sigma^{-1}}$ inverse de $P_\sigma$.

Notons qu'il faut de l'ordre de $n(n-1)/2$ opération sur les lignes donc $\mathcal{O}(n^3)$ calculs pour une matrice de taille $n$.

**Applications :** comme $L$ et $U$ sont triangulaires, alors leur déterminant est le produit des coefficients diagonaux. Du coup, $\det(A) = \det(L)\det(U)$ devient facile à calculer ($\mathcal{O}(n^4)$ opérations par cette méthode, éventuellement améliorable, à la place de $n!$ par le

développement sur les colonnes). De même, l'inverse de $U$ est facile à faire (on continue le pivot en remontant) et idem pour l'inverse de $L$. On peut donc calculer $A^{-1}$ polynomialement. Notons qu'il vaut mieux de ne pas calculer $L$ et plutôt garder la liste des $L_i$ et $M_{ij}$ car leurs inverses et leur déterminant sont encore plus rapides à faire. On peut bien sûr utiliser cette forme pour résoudre directement des systèmes $Ax = b$ car les $L_i$ et $M_{ij}$ s'inversent rapidement et comme $U$ est triangulaire supérieure, on trouve $x_n$ puis $x_{n-1}\dots$ en remontant.

# 3 Méthode de la puissance

Nous voulons calculer les valeurs propres et vecteurs propres d'une matrice $A$. Si on utiliser la méthode du polynôme caractéristique, il y a plusieurs soucis, le premier étant que les matrices de la vraie vie ont parfois des dimensions bien plus grandes que $100 \times 100$ et on obtiendrait des équations polynomiales trop difficiles à résoudre (sans parler de la propagation des erreurs dans ce cas).

Dans de nombreux cas, nous sommes intéressés par la première valeur propre (la plus grande en module) et il existe une méthode pour l'obtenir de façon stable.

**Théorème 3.2.** *Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice telle qu'il existe une valeur propre simple $\lambda \in \mathbb{C}$ vérifiant que tout autre valeur propre $\mu$ de $A$ est telle que $|\mu| \leq |\lambda| - \varepsilon$ avec $\varepsilon > 0$.*

*Alors, pour presque tout $x_0 \in \mathbb{R}^d$, la suite définie par récurrence*

$$x_{n+1} = \frac{1}{\|Ax_n\|} Ax_n$$

*converge vers un vecteur propre $x_\infty$ de $A$ pour la valeur propre $\lambda$.*

***Démonstration :*** Soit $(e_1,e_2,\dots,e_d)$ une base dans laquelle $A$ est une matrice de type Jordan. Quitte à changer les notations, supposons que $e_1$ est un vecteur propre de la valeur propre simple $\lambda$. Si $e_j$ est un vecteur propre pour la valeur propre $\mu_j$, on a $A^n e_j = \mu_j^n e_j$. Si $e_{j+1},\dots,e_{j+p}$ est un bloc de Jordan, alors

$$A^n e_{j+p} = \mu_j^n e_{j+p} + \mu_j^{n-1} e_{j+p-1} + \dots + \mu_j^{n-p} e_j \ .$$

Donc, si $x = \sum c_k e_k$, on a

$$A^n x = \lambda^n c_1 e_1 + \mathcal{O}(|\lambda| - \varepsilon)^n \ .$$

Comme $x_n$ n'est qu'une renormalisation à chaque étape de $A^n x$, on obtient que $x_n = e_1 + \mathcal{O}\left(\left(1 - \frac{\varepsilon}{\lambda}\right)^n\right)$... enfin sauf si $c_1 = 0$. Mais le cas $c_1 = 0$ correspond à un vecteur $x_0$ pris dans un hyperplan de $\mathbb{R}^d$. Si on a tiré au sort $x_0$, la probabilité de tomber dans cet hyperplan est nul. C'est pour cela que l'algorithme fonctionne pour « presque tout » $x_0$. □

**Proposition 3.3.** *La méthode de la puissance est stable dans le sens où, si une erreur de taille $\eta$ est commise à chaque étape, alors l'erreur sur $x_\infty$ est au plus de $\lambda\eta/\varepsilon$.*

***Démonstration :*** Imaginons qu'il y ait une petite erreur à chaque étape, c'est-à-dire $x_{n+1} = \frac{1}{\|Ax_n\|}Ax_n + \eta_n$. La partie de $\eta_n$ selon $e_1$ ne change pas grand chose car il est dans la direction qui nous intéresse et sera gommé lors de la renormalisation. Le problème vient de la partie transverse qui rajoute des termes petits à chaque étape. Même si ces termes sont petits, ils deviennent importants face à la partie $\mathcal{O}(|\lambda| - \varepsilon)^n$ quand $n$ est grand. Le point clef ici est de constater qu'aux étapes suivantes, ces erreurs sont multipliées à chaque fois par $(|\lambda| - \varepsilon)$. L'erreur totale à l'étape $n$ est donc du type

$$\sum_{m=0}^{n} \eta_{n-m}(|\lambda| - \varepsilon)^m \leq \frac{\lambda}{\varepsilon} \max(|\eta_n|) \ .$$

L'accumulation des erreurs n'explosent donc pas quand on itère le processus : on dit que le processus est stable [1]. □

Si en outre $A$ est une matrice symétrique, c'est-à-dire que ${}^tA = A$, alors les sous-espaces propres sont orthogonaux entre eux.

**Proposition 3.4.** *Soit $A = (a_{ij})$ une matrice symétrique, c'est-à-dire que ${}^tA = A$ ou encore que $a_{ij} = a_{ji}$. Soient $x$ et $y$ deux vecteurs propres de $A$ pour deux valeurs propres $\lambda \neq \mu$. Alors $x \perp y$.*

***Démonstration :*** On a

$$\lambda\langle x|y\rangle = \langle Ax|y\rangle = \sum_i (\sum_j a_{ij}x_j)y_i = \sum_{ij} a_{ij}x_jy_i = \sum_{ij} a_{ji}x_jy_i = \langle x|Ay\rangle = \mu\langle x|y\rangle$$

et donc $\langle x|y\rangle = 0$ car $\lambda \neq \mu$. □

Si on prend $y_0 \in \mathbb{R}^d$ qui est perpendiculaire au vecteur propre $x_\infty$ trouvé plus haut, alors la méthode de la puissance $y_{n+1} = Ay_n/\|Ay_n\|$ doit donner une suite qui reste orthogonale à $x_\infty$ et converge vers un deuxième vecteur propre pour la deuxième plus grande valeur propre en module.

1. Le lecteur attentif pourra trouver cette démonstration peu rigoureuse car on passe vite sur le procédé de normalisation qui est non-linéaire. Une justification plus rigoureuse serait de présenter le problème comme une itération du type « point fixe » et d'utiliser les fonctions implicites pour quantifier comment le point fixe limite est influencé par l'erreur commise sur la fonction itérée.

```
1 A:=[[2,0,1],[0,-1,3],[1,3,2]]
```

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & -1 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

```
2 V:=[1.,0.,0.]; pour j de 1 jusque 50 faire V:=A*V; V:=V/norm(V);
ffaire; print(V)
V :[0.36758125538,0.465377349647,0.805175721895]

3 (A*V)./V
```

$$[4.19047002564, 4.19046998639, 4.19047005065]$$

```
4 W:=[1.,0.,0.]; W:=W-(V*W)*V; pour j de 1 jusque 50 faire W:=A*W;
W:=W/norm(W); ffaire; print(W)
W :[1.04720405515e-07,0.671845953128,-0.740690903998]

5 (A*W)./W
```

$$[-7073031.19116, -4.30741400116, -0.721159330057]$$

On trouve bien un vecteur propre $V$ pour la valeur propre $\lambda \simeq 4{,}1904\ldots$. Mais en prenant ensuite un point de départ orthogonal, la convergence ne marche pas bien. En effet, le soucis est que la moindre petite erreur dans le calcul de $V$ et le procédé d'orthogonalisation induira un terme selon la première direction propre qui va grossir exponentiellement vite par rapport à la deuxième valeur propre. Il faut donc rendre bien $W$ orthogonal à $V$ à chaque étape pour stabiliser ce terme d'erreur et le garder petit.

```
6 W:=[1.,0.,0.]; W:=W-(V*W)*V; pour j de 1 jusque 50 faire
W:=A*W;W:=W-(V*W)*V; W:=W/norm(W); ffaire; print(W)
W :[0.108761330581,0.838335537288,-0.534195188943]

7 (A*W)./W
```

$$[-2.91162792959, -2.91162785728, -2.91162780339]$$

On trouve comme deuxième valeur propre $\mu \simeq -2{,}9116\ldots$ et on peut trouve de même la troisième.

```
8 X:=[1.,0.,0.]; X:=X-(V*X)*V-(W*X)*W; pour j de 1 jusque 50 faire
X:=A*X; X:=X-(V*X)*V-(W*X)*W; X:=X/norm(X); ffaire; print(X)
X :[0.923609762651,-0.283932121035,-0.257541369456]

9 (A*X)./X
```

$$[1.72115780942, 1.72115780896, 1.72115780972]$$

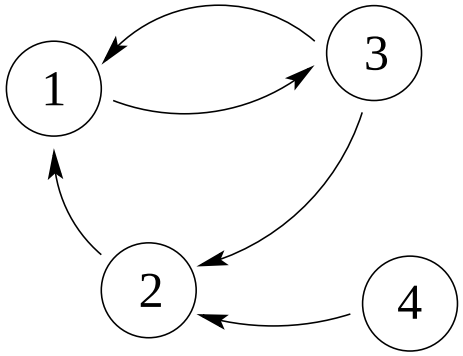
Vérifions notre calcul grâce au déterminant et à la trace de $A$ :

```
10 det(A); (A*V)./V.*(A*W)./W.*(A*X)./X; (A*V)./V+(A*W)./W+(A*X)./X;
```

$$-21,[-21.0000005878, -20.999999864, -20.9999998066]$$
$$[2.99999990547,2.99999993806,3.00000005698]$$

# 4 L'algorithme PageRank de Google

Le réseau internet est composé de milliers de milliards de pages reliées entre elles. On peut le modéliser par un graphe orienté où deux pages sont reliées par une flèche si l'une renvoie sur l'autre. Ci-contre un exemple de graphe avec seulement 5 pages. La question est de savoir quelle est la page la plus pertinente dans ce réseau. Pour cela, on va attribuer une note $x_i \geq 0$ à la page $i$ et la plus pertinente sera celle avec la meilleure note.




**Première idée : comptage des liens**
L'idée la plus simple est de définir $x_i$ comme le nombres de pages envoyant sur la page $i$. Dans notre exemple, les pages 1, 2 et 3 sont les gagnantes. La méthode est simple mais posent plusieurs problèmes :

- Un lien venant d'une page très réputée a le même poids que le lien venant d'une page inintéressante.
- Le lien venant d'une page avec très peu de liens choisis a le même poids qu'un lien venant d'une page citant tout le monde sans distinction.
- L'algorithme est facilement détournable par la création de pages vides pointant sur une page donnée.

**Deuxième idée : mesure stationnaire**
On va donc essayer de trouver un équilibre pour les notes tel que le lien venant d'une page $j$ pointant sur la page $i$ amène une note proportionnelle à la note de la page $j$ et inversement proportionnelle au nombre de liens $n_j$ de la page $j$. On veut donc que

$$\forall i \ , \ \ x_i = \sum_{j \text{ pointant vers } i} \frac{1}{n_j} x_j \ .$$

Cela peut se représenter matriciellement sous la forme suivante. Soit $A$ la matrice telle que $a_{ij} = 0$ si $j$ ne pointe pas sur $i$ et $a_{ij} = 1/n_j$ si $j$ pointe sur $i$. Dans notre exemple, on a ainsi

$$A = \begin{pmatrix} 0 & 1 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \ .$$

On cherche $x = (x_i)$ tel que $Ax = x$, c'est-à-dire que $x$ est vecteur propre pour la valeur propre 1 pour $A$.

**Théorème 3.5. Perron-Frobenius**
*Soit $A$ une matrice avec $a_{ij} \geq 0$ et telle que la somme sur chaque colonne vérifie $\sum_i a_{ij} = 1$. Alors $A$ a au moins un vecteur propre $x$ pour la valeur propre 1 qui a toutes ses coordonnées positives.*

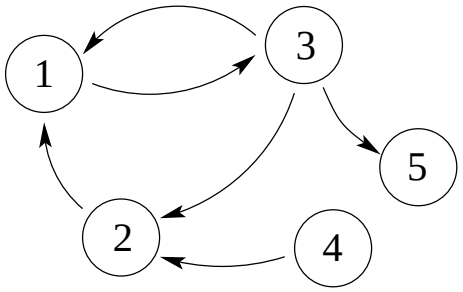
Nous prouverons ce théorème plus bas. Nous pouvons l'appliquer au réseau exemple à l'aide d'une méthode itérative.

```
11 A:=[[0,1,1/2,0],[0,0,1/2,1],[1,0,0,0],[0,0,0,0]];
12 X:=[[1.],[1.],[1.],[1.]]; pour j de 1 jusque 50 faire X:=A*X;
X:=X/norm(X);ffaire;
```

$$\begin{pmatrix} 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 0.666666674945 \\ 0.333333325055 \\ 0.666666662527 \\ 0.0 \end{pmatrix}$$

On trouve donc que $x = (2/3{,}1/3{,}2/3{,}0)$ est une notation qui convient et donne les pages 1 et 3 comme les meilleures. On note juste que cela semble un peu sévère pour la page 4 qui n'a aucun poids.

Mais en fait, il y a plusieurs problèmes. D'abord, il peut exister plusieurs vecteurs propres. Pire, on a un problème dans le cas d'une page qui n'a aucun lien. Par exemple dans le cas ci-dessous.




$$B = \begin{pmatrix} 0 & 1 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \end{pmatrix}.$$

Le théorème ne peut s'appliquer à cause de la dernière colonne et on peut même voir qu'il ne peut y avoir de mesure stationnaire, c'est-à-dire de vecteur $x$ positif tel que $Ax = x$. En effet, dans ce cas, on doit avoir $x_5 = 0$ car sinon le poids de $x_5$ est perdu dans $Ax$. Mais alors, il faut que $x_3 = 0\ldots$ et donc que $x \equiv 0$.

**Interprétation stochastique : le « surfeur aléatoire »**
On peut interpréter les matrices ci-dessus en terme de « chaînes de Markov » et de « processus stochastiques ». Partons d'un vecteur $x$ qui vaut 1 sur une case $i$ et 0 ailleurs. Si cette case $i$ pointe sur deux cases $j$ et $k$, alors $Ax$ est un vecteur qui vaut $1/2$ sur ces cases et 0 ailleurs. Ainsi $Ax$ mesure la probabilité d'être sur une case si on part de la case $i$ et on suit un lien de façon aléatoire depuis cette case. Quand on itère, $A^n x$ mesure la probabilité de se trouver sur les différentes cases après avoir suivi $n$ liens. On cherche une mesure stationnaire c'est-à-dire telle que $Ax = x$ avec l'idée qu'on a une bonne probabilité d'être sur une page si celle-ci est indiquée par beaucoup de pages de façon plus ou moins directe.

Nous pouvons maintenant imaginer d'effectuer l'expérience plutôt que de mesurer une probabilité. Posons un petit robot sur une page et laissons le robot suivre des liens : à

chaque page, il tire au sort un lien de la page et le suit. En répétant de nombreuses fois l'expérience, on peut espérer que la moyenne du temps passer par le robot sur une page mesure bien son importance. On peut montrer mathématiquement que, sous de bonnes hypothèses sur le réseau (ou sur $A$), alors la moyenne du temps passé par le robot sur les pages converge bien vers $x$ tel que $Ax = x$.

Sous cette interprétation, on voit bien le problème du deuxième réseau : le robot finira bien par aller dans la page 5 et y restera indéfiniment. Cette page « trou noir » fait échec au processus. De manière générale, le robot ne pourra pas voir des parties du réseau qui ne sont pas connectées. Dans tous ces cas, les conditions des théorèmes mathématiques ne sont pas remplis.

**Troisième idée : une téléportation aléatoire**
Pour éviter des cas comme le « trou noir » de la page 5 ci-dessus, on ajoute que le robot a toujours une probabilité $\theta \in ]0,1[$ de quitter la page actuelle et de se téléporter sur une page tirée au hasard. Ce saut devient automatique pour une page « trou noir » ce qui revient à rajouter des lien de cette page vers toutes les autres.

Soit $n$ le nombre total de pages et $C$ la matrice remplie par des termes $1/n$. Si $\sum x_i = 1$, alors $Cx = y := (1/n,1/n,\ldots)$. Notre problème revient donc à trouver $x$ tel que

$$x_i \geq 0 \quad , \quad \sum_i x_i = 1 \quad \text{et} \quad (1-\theta)Ax + \theta y = x \ .$$

**Théorème 3.6.** *Soit $A$ une matrice avec $a_{ij} \geq 0$ et telle que la somme sur chaque colonne vérifie $\sum_i a_{ij} = 1$, soit $\theta \in ]0,1[$ et soit $y = (1/n,1/n,\ldots)$. Si*

$$\mathcal{X} = \{x \in \mathbb{R}^n \ , \ x_i \in [0,1] \ \textit{et} \ \sum_i x_i = 1\} \ .$$

*alors*

$$\Phi \ : \ x \in \mathcal{X} \ \longmapsto \ (1-\theta)Ax + \theta y \in \mathcal{X}$$

*est bien définie et admet un unique point fixe $x^*$ dans $\mathcal{X}$, qui a toutes ses coordonnées dans $]0,1[$. En outre, la suite $(x^k)$ définie par $x^0 \in \mathcal{X}$ et $x^{k+1} = \Phi(x^k)$ converge vers $x^*$.*

***Démonstration :*** Par définition de $A$ et $y$, il est facile de voir que si $x \in \mathcal{X}$, alors les coordonnées de $(1-\theta)Ax + \theta y$ sont positives. En outre

$$\begin{aligned}\sum_i (\Phi(x))_i &= \sum_{i=1}^n \left( (1-\theta) \sum_j a_{ij} x_j + \frac{\theta}{n} \right) \\ &= (1-\theta) \sum_{i,j} a_{ij} x_j + \theta = (1-\theta) \sum_j x_j (\sum_i a_{ij}) + \theta \\ &= (1-\theta) \sum_j x_j + \theta = 1\end{aligned}$$

Donc $\Phi$ est bien définie de $\mathcal{X}$ dans lui-même. Il nous reste simplement à montrer qu'il

s'agit d'une fonction contractante. On munit $\mathcal{X}$ de la norme $\|x\| = \sum |x_i|$. On a alors

$$\begin{aligned}\|\Phi(x) - \Phi(x')\| &= (1-\theta)\|Ax - Ax'\| = (1-\theta)\|A(x-x')\| \\ &= (1-\theta)\sum_i |\sum_j a_{ij}(x_j - x'_j)| \leq (1-\theta)\sum_j \sum_i a_{ij}|x_j - x'_j| \\ &\leq (1-\theta)\sum_j |x_j - x'_j| = (1-\theta)\|x - x'\| \ .\end{aligned}$$

On trouve donc que $\Phi$ est $(1-\theta)-$lipschitzienne pour cette norme et donc contractante. Elle a un unique point fixe $x^*$ dans $\mathcal{X}$ qui peut s'approcher par toute suite itérative. En outre, comme $y$ n'a que des coordonnés strictement positive et $Ax^*$ que des coordonnées positives ou nulles, $x^* = \Phi(x^*)$ n'a que des coordonnés strictement positives. □

Avant de tester ce théorème, on peut en déduire le théorème précédent.
***Démonstration du théorème 3.5 :*** Pour chaque $k \in \mathbb{N}$, la fonction $\Phi_k(x) = (1 - 1/k)Ax + y/k$ a un unique point fixe $x^k$ dans $\mathcal{X}$. Comme $\mathcal{X}$ est un fermé borné de $\mathbb{R}^n$, il est compact. On peut extraire une sous-suite convergente $x^{\varphi(k)}$ vers un $x^*$ dans $\mathcal{X}$. En passant à la limite dans $x^{\varphi(k)} = (1 - 1/k)Ax^{\varphi(k)} + y/k$, on trouve que $x^* = Ax^*$. Notons que cette existence d'une limite pour une sous-suite ne garantie par l'unicité d'un tel $x^*$. □

Faisons marcher ce système de notation des pages webs pour nos deux exemples.

13 `X:=[[1.],[0.],[0.],[0.]] ; Y:=[[0.25],[0.25],[0.25],[0.25]] ; pour j de 1 jusque 50 faire X:=0.9*A*X+0.1*Y; ffaire;`

$$\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \begin{pmatrix} 0.386659436038 \\ 0.215347071678 \\ 0.372993492284 \\ 0.025 \end{pmatrix}$$

14 `B:=[[0,1,1/3,0,1/5],[0,0,1/3,1,1/5],[1,0,0,0,1/5], [0,0,0,0,1/5],[0,0,1/3,0,1/5]] ;`

15 `X:=[[1.],[0.],[0.],[0.],[0.]] ; Y:=[[0.2],[0.2],[0.2],[0.2],[0.2]] ; pour j de 1 jusque 50 faire X:=0.9*B*X+0.1*Y; ffaire;`

$$\begin{pmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}, \begin{pmatrix} 0.307067227127 \\ 0.183210537646 \\ 0.321952498201 \\ 0.0455919937843 \\ 0.142177743241 \end{pmatrix}$$

Dans les deux cas, la page 4 est la moins bien notée et les pages 1 et 3 sont les meilleures. Par contre, leur ordre change quand on ajoute la page 5. On peut interpréter cela en disant que la page 3 met moins en valeur la page 1 puisqu'elle possède plus de liens.

**En pratique**
La définition du poids des pages donnée ci-dessus est bien celle qui apparaît dans *The Anatomy of a Large-Scale Hypertextual Web Search Engine* de Sergey Brin et Lawrence Page publié en 1998. Le brevet était à l'origine à l'université de Stanford, qui employait les auteurs mais il a été rapidement donné à la Start-Up Google. Il est aujourd'hui dans le domaine publique.

Dans l'article originel, on peut lire

> *We assume page A has pages* $T1 \dots Tn$ *which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also* $C(A)$ *is defined as the number of links going out of page A. The PageRank of a page A is given as follows :*
>
> $$PR(A) = (1 - d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$
>
> *Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages PageRanks will be one.*

Donc, au moins au départ, on avait $\theta = 0{,}15$. Il y a très peu de détails mathématiques, mais plutôt une grande discussion sur les moyens pratiques et matériels d'obtenir ces notes.

L'algorithme PageRank est bien sûr un peu trop naïf et il a été perfectionné. Le poids des pages dépend aussi de nombreux facteurs, dont la plupart sont gardés secrets pour ne pas faciliter les tentatives de manipulation des poids.

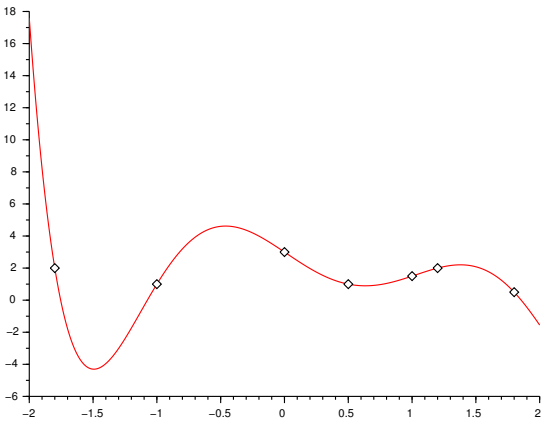
# Chapitre 4 : Approximation polynomiale

Nous avons vu que les seuls fonctions vraiment calculables sont les polynômes, et même il est mieux de se limiter au bas degrés. Si on a une fonction $f : [0,1] \to \mathbb{R}$, on aimerait l'approcher par des polynômes de façon globale sur [0,1] (et non pas locale comme le fait le développement de Taylor). On verra qu'une des applications est le calcul de son intégrale une fois la fonction approchée. L'utilisation de polynôme approchant ou interpolant est pratique pour stocker un tracer en quelques coefficients et se retrouve à la base des dessins sur les ordinateurs : splines pour la CAO ou courbes de Bézier pour les polices de caractères.

## 1 Polynômes de Lagrange

### 1.1 Interpolation de Lagrange

On considère $n+1$ points $(x_k,y_k)_{k=0..n} \subset \mathbb{R}^2$ avec $x_k \neq x_{k'}$. Notre but est de trouver un polynôme de degré $n$ qui passe par tous ces points. La courbe de ce polynôme fera donc une interpolation entre ces différents points de façon relativement lisse. Le degré $n$ est le degré minimum pour pouvoir réaliser cette interpolation en général à cause du théorème suivant.




**Théorème 4.1.** *Soit $P \in \mathbb{R}[X]$ un polynôme non nul de degré $n$. Alors $P$ a au plus $n$ racines réelles.*

***Démonstration :*** Nous avons vu l'indice de Budan-Fourier et montré qu'il majore le nombre de racines de $P = c_n X^n + \ldots + c_0$ dans l'intervalle $[a,b]$. Cet indice considère le nombre de changements de signe dans la suite $(P(x),P'(x),P''(x),\ldots,P^{(n)}(x))$. Or, quand $x$ tend vers $+\infty$, seul le terme dominant compte et tous les polynômes seront du signe du coefficients $c_n$ : il n'y aura pas de changement de signes. A l'inverse, quand $x$ tend vers $-\infty$, les signes des termes dominants alterneront selon la parité de la puissance et on aura $n$ changements de signes. Le théorème de Budan-Fourier montre donc que l'on a au plus $n$ racines. □

Rappel : le théorème de d'Alembert-Gauss indique qu'il y a en fait exactement $n$ racines si on compte les racines complexes et les multiplicités.

Le résultat précédent montre qu'il faut au moins aller au degré $n$ car si $Q$ est de degré $n-1$ et vérifie $Q(k) = k^n$ pour $k = 0..n$, alors $X^n - Q$ est de degré $n$, est non nul et possède $n+1$ racines. Le degré $n$ est en fait suffisant.

**Théorème 4.2.** *Soient $n+1$ points $(x_k,y_k)_{k=0..n} \subset \mathbb{R}^2$ avec $x_k \neq x_{k'}$. Alors il existe un unique polynôme $P \in \mathbb{R}_n[X]$ de degré au plus $n$ tel que $P(x_k) = y_k$ pour tout $k = 0..n$. En outre, $P$ est donné par*

$$P = \sum_{k=0}^{n} y_k \prod_{j \neq k} \left( \frac{X - x_j}{x_k - x_j} \right) \ .$$

***Démonstration :*** Pour l'existence, il suffit de tester la formule. Elle donne bien un polynôme de degré $n$ et quand on l'évalue en $x_{k^*}$, tous les termes de la somme sont nuls sauf celui pour $k = k^*$ pour lequel on a le produit qui fait exactement 1. Donc $P(x_{k^*}) = y_{k^*}$.

Pour l'unicité, on reprend l'argument ci-dessus. Soient $P$ et $Q$ deux polynômes de degré au plus $n$ réalisant l'interpolation. Alors $P - Q$ est un polynôme de degré au plus $n$ ayant $n+1$ racines. Il ne peut donc s'agir que du polynôme nul, c'est-à-dire que $P = Q$. □

En fait, cette méthode permet d'obtenir une base agréable pour les polynômes.

**Théorème 4.3.** *Soient $n+1$ points $(x_k)_{k=0..n} \subset \mathbb{R}$ avec $x_k \neq x_{k'}$. Les polynômes*

$$L_k = \prod_{j \neq k} \left( \frac{X - x_j}{x_k - x_j} \right) \ ,$$

*appelés polynômes de Lagrange, forment une base de $\mathbb{R}_n[X]$. En outre, si $P \in \mathbb{R}_n[X]$ est un polynôme de degré au plus $n$, alors*

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

***Démonstration :*** Le théorème précédent nous dit que, puisque $P$ et $\sum_{k=0}^{n} P(x_k) L_k$ interpolent tous les deux les points $(x_k,P(x_k))$, alors ils sont égaux. Cela montre que les polynômes de Lagrange forment une famille génératrice et donc une base car la famille est du bon cardinal. □

Nous venons de trouver une première manière d'interpoler les fonctions : l'interpolation de Lagrange. Soit $n+1$ points $(x_k)_{k=0..n} \subset \mathbb{R}$ avec $x_k \neq x_{k'}$ et soit $f$ une fonction. L'interpolation de Lagrange de $f$ aux points $(x_k)$ est le polynôme de degré au plus $n$ donné par

$$L = \sum_{k=0}^{n} f(x_k) L_k \ .$$

Notons que trouver ce polynôme interpolateur en cherchant juste les coefficients tels que $P(x_k) = f(x_k)$ demande à résoudre un système $(n+1) \times (n+1)$ plein. Alors que si on passe bien par la base des polynômes de Lagrange, le calcul est immédiat car diagonalisé.

Cette interpolation est exacte sur les polynômes de degré au plus $n$. Quelle erreur commet-on pour une fonction quelconque ?

**Proposition 4.4.** *Soit $f \in \mathcal{C}^{n+1}([a,b],\mathbb{R})$ une fonction $(n+1)$ fois dérivable sur un intervalle $[a,b]$. Soient $a \leq x_0 < x_1 < \ldots < x_n \leq b$ des points dans $[a,b]$ et soit $P$ le polynôme interpolateur de Lagrange aux $(n+1)$ points $(x_k)$. Alors pour tout $x \in [a,b]$, il existe $\xi_x \in ]\min(x,x_0), \max(x,x_n)[$ tel que*

$$f(x) - P(x) = \frac{1}{(n+1)!} \left( \prod_{k=0}^{n} (x - x_k) \right) f^{(n+1)}(\xi_x) \ .$$

***Démonstration :*** Si $x$ est un des $(x_k)$, le résultat est trivial donc on va maintenant supposer $x \neq x_k$. On introduit $Q$ le polynôme interpolateur de $f$ aux $(n+2)$ points distincts $(x_k) \cup x$. Le polynôme $Q - P$ est de degré $n+1$ et s'annule sur les $n+1$ points $x_k$ donc il existe $c \in \mathbb{R}$ tel que

$$Q - P = c \prod_{k=0}^{n} (X - x_k) \ .$$

On considère la fonction $g(\xi) = f(\xi) - Q(\xi) = f(\xi) - P(\xi) - c\prod_{k=0}^{n}(\xi - x_k)$ qui est $(n+1)$ fois dérivable et s'annule aux $(n+2)$ points $(x_k) \cup x$. Par le théorème de Rolle, sa dérivée s'annule sur $(n+1)$ points intercalés, donc sa dérivée seconde sur $n$ points... et par récurrence on obtient que $g^{(n+1)}$ s'annule sur un point $\xi_x$. Mais $P^{(n+1)} = 0$, donc $g^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - c(n+1)!$ et donc $c = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x)$. Il ne reste plus qu'à voir que

$$\begin{aligned} f(x) - P(x) &= (f(x) - Q(x)) + c \prod_{k=0}^{n} (x - x_k) = 0 + c \prod_{k=0}^{n} (x - x_k) \\ &= \frac{1}{(n+1)!} \left( \prod_{k=0}^{n} (x - x_k) \right) f^{(n+1)}(\xi_x) \ . \end{aligned}$$

□

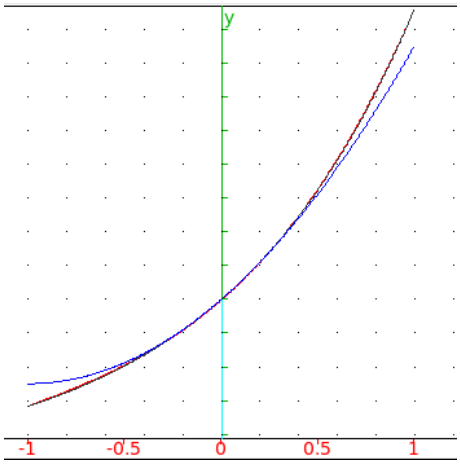
**Exemple d'application :** on veut calculer le polynôme caractéristique d'une matrice $A$ de taille $n \times n$. Pour cela, il suffit de calculer $(n+1)$ déterminants $\det(A - \lambda Id)$ avec $\lambda = 0, \ldots, n$. On applique ensuite l'interpolation de Lagrange.

## 1.2 Problème de la convergence

Fixons nous un intervalle $[a,b]$ et une fonction $f \in \mathcal{C}^{\infty}([a,b],\mathbb{R})$. On veut approcher au mieux la fonction par des polynômes et donc on peut penser à l'interpoler par les polynômes de Lagrange aux points équidistants $x_k = a + kh$ pour $k = 0..n$ avec $h = (b-a)/n$. Ensuite, en faisant tendre $n$ vers l'infini, on espère converger vers la fonction $f$. Faisons

le test.
Dans le premier essai, nous faisons l'interpolation de Lagrange de la fonction exponentielle sur $[-1,1]$ avec 3 points. Nous en profitons pour afficher aussi le polynôme de Taylor de degré 2 en zéro pour comparer.

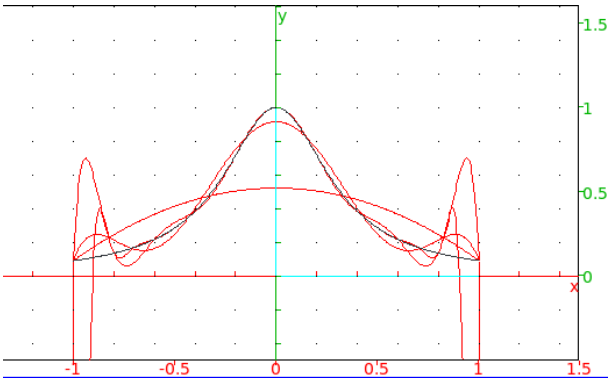
```
1 n:=3.; points:=[(-1+k*2./n)$(k=0..n)];
P:=lagrange(points,exp); plot([P(x),exp(x),
1+x+x^2/2.],x=-1..1, couleur=[rouge,noir,
bleu]);
```




Nous voyons un résultat quasiment parfait avec seulement une parabole (l'exponentielle est en noir et son interpolation en rouge). Le développement de Taylor (en bleu) est bon près de zéro mais rapidement mauvais plus loin.

Faisons maintenant le test avec la fonction $x \mapsto 1/(1+10x^2)$. Au début, l'interpolation de Lagrange ne donne pas une très bonne approximation. Nous allons donc tester avec 10 points.

```
2 n:=10.; points:=[(-1+k*2./n)$
(k=0..n)]; f(x):=1/(1+10*x^2);
P:=lagrange(points,f);plot([P(x),f(x)],
x=-1..1,couleur=[rouge,noir]);
```




Puis même 30 points (ci-contre plusieurs tests avec de plus en plus de points). Étrangement, c'est de pire en pire !

Le but de ce paragraphe est de comprendre ce dernier phénomène qui est appelé *phénomène de Runge*.

**Proposition 4.5.** *Soit* $[a,b]$, $n \in \mathbb{N}$ *et* $h = (b-a)/n$ *et soit* $x_k = a + kh$ *pour* $k = 0..n$ *les* $(n+1)$ *points équirépartis dans* $[a,b]$. *Alors*

$$\max_{x\in[a,b]} \left| \prod_{k=0}^{n} (x - x_k) \right| \leq h^{n+1} n! = \mathcal{O}\left(\frac{b-a}{e}\right)^{n+1} .$$

***Démonstration :*** On a

$$\left| \prod_{k=0}^{n} (x - x_k) \right| = h^{n+1} \left| \prod_{k=0}^{n} (s - k) \right| := h^{n+1} \phi(s)$$

avec $s = (x-a)/h$. On voit que $\phi(s)$ est symétrique par rapport à $n/2$, c'est-à-dire que $\psi(n-s) = \psi(s)$, donc on peut se contenter de regarder $s \in [0,n/2]$. Par ailleurs, on a dans cet intervalle $\phi(s-1)/\phi(s) = (n+1-s)/s > 1$ et donc $\phi$ est maximum sur $[0,1]$. On trouve donc

$$\phi(s) \leq \max_{s\in[0,1]} \left| \prod_{k=0}^{n} (s - k) \right| \leq n!$$

et on peut remarquer que les gros problèmes viennent du bord de l'intervalle, ce que l'on a vu dans nos tests. Pour conclure, on a que

$$h^{n+1}n! \;=\; \frac{(b-a)^{n+1}}{n^{n+1}}n! \;\sim\; \frac{(b-a)^{n+1}}{n^{n+1}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \;=\; \mathcal{O}\left(\frac{b-a}{e}\right)^{n+1}$$

par l'équivalent de Stirling (nous sommes surtout intéressé par la partie puissance, ce qui explique qu'on ne cherche pas à gérer mieux le $\sqrt{n}$). □

Maintenant, il nous reste à évaluer la partie $f^{(n+1)}(\xi_x)$. Pour cela, on va supposer que notre fonction est développable en série entière.

**Définition 4.6.** *On dit que $f$ est développable en série entière en $x_0 \in \mathbb{R}$ avec un rayon de convergence $R > 0$ si, pour tout $x \in ]x_0 - R, x_0 + R[$ on a*

$$f(x) = \sum c_n (x - x_0)^n$$

*avec les coefficients $c_n = f^{(n)}(x_0)/n!$.*

On peut alors montrer le résultat de convergence suivant.

**Théorème 4.7.** *Soit $f : [a,b] \to \mathbb{R}$ développable en série entière en $(a+b)/2$ avec un rayon de convergence $R > 0$. Soit $n \in \mathbb{N}$, $h = (b-a)/n$ et $x_k = a + kh$ pour $k = 0..n$ les $(n+1)$ points équirépartis dans $[a,b]$. Soit $P_n$ le polynôme d'interpolation de Lagrange correspondant à $f$ et ces points. Alors si $R > \left(\frac{1}{e} + \frac{1}{2}\right)(b-a)$, il existe $\lambda \in [0,1[$ tel que*

$$\max_{x\in[a,b]} |f(x) - P_n(x)| = \|f - P_n\|_\infty \;=\; \mathcal{O}(\lambda^n) \;\xrightarrow[n\longrightarrow\infty]{}\; 0\ .$$

***Démonstration :*** Il reste à estimer $f^{(n+1)}(\xi_x)$. Pour cela remarque que pour tout $r < R$, on a que la série $\sum c_n r^n$ converge et donc qu'il existe $C(r)$ tel que $|c_n| \leq C(r) r^{-n}$. On a ainsi pour $\xi = x_0 + x$ avec $|x| < r$

$$\begin{aligned}|f^{(n+1)}(\xi)| &\leq C(r)\sum_{k\geq 0}\frac{1}{r^k}\frac{d^{n+1}}{dx^{n+1}}(x^k) = C(r)\frac{d^{n+1}}{dx^{n+1}}\left(\sum_{k\geq 0}\left(\frac{x}{r}\right)^k\right) = C(r)\frac{d^{n+1}}{dx^{n+1}}\left(\frac{r}{r-x}\right)\\ &\leq \frac{(n+1)!\, rC(r)}{(r-x)^{n+1}}\end{aligned}$$

Soit $R > \left(\frac{1}{e} + \frac{1}{2}\right)(b-a)$, on a alors un $r < R$ avec encore $r - x > (b-a)/e$ pour tout $x < (b-a)/2$. On a donc

$$\begin{aligned}\frac{1}{(n+1)!}|f^{(n+1)}(\xi)| \max_{s\in[a,b]}\left|\prod_{k=0}^{n}(s - x_k)\right| &\leq rC(r)K\left(\frac{b-a}{e(r-x)}\right)^{n+1}\\ &\xrightarrow[n\longrightarrow\infty]{} 0\ .\end{aligned}$$

□

On comprend donc pourquoi notre interpolation est bonne pour l'exponentielle (rayon de convergence $R = \infty$) mais pas pour $f(x) = 1/(1+10x^2)$ (rayon de convergence $R = 1/\sqrt{10}$). En fait, on peut démontrer dans ce dernier cas que $\max_{x\in[-1,1]} |f(x) - P_n(x)| \to +\infty$ quand le nombre de points tend vers $+\infty$.

# 2 Polynômes orthogonaux

## 2.1 Rappels sur les espaces préhilbertiens

**Définition 4.8.** *Un espace préhilbertien est un espace vectoriel réel $E$ muni d'un produit scalaire $\langle\cdot|\cdot\rangle$ et de sa norme associée $\|\cdot\|$.*

La différence avec les espaces euclidiens réside donc seulement dans le fait qu'un espace préhilbertien peut être (mais pas obligatoirement) de dimension infinie. On pourra aussi penser à un espace de dimension finie mais très grande.

Dans ce chapitre, notre exemple favori va être l'espace des polynômes $\mathbb{R}[X]$ muni du produit scalaire $\langle P|Q\rangle = \int_{-1}^{1} P(x)Q(x)dx$. On peut aussi imaginer d'autres produits scalaire comme $\langle P|Q\rangle = \int_0^\infty e^{-x}P(x)Q(x)dx$.

Le concept principal de cette partie sera l'approximation par projection orthogonale sur un sous-espace de dimension finie. L'intérêt est de remplacer au mieux une description ayant besoin d'un nombre très grand, voire infini, de dimensions, par une description très correcte en quelques nombres.

**Théorème 4.9. Projection orthogonale**
*Soit $E$ un espace préhilbertien et soit $F$ un sous-espace de dimension finie de $E$, dont $(e_1,\ldots,e_d)$ est une base orthonormale.*

*Alors, pour tout $x \in E$, les trois définitions suivantes de $\Pi x$ sont équivalentes.*

*i) $\Pi x$ est l'unique vecteur de $F$ tel que $x - \Pi x$ est orthogonal à $F$.*
*ii) $\Pi x$ est l'unique vecteur de $F$ tel que $\|x - \Pi x\| = \min_{y\in F} \|x - y\|$.*
*iii) $\Pi x = \langle x|e_1\rangle e_1 + \ldots + \langle x|e_d\rangle e_d$.*

***Démonstration :*** Il est clair que la dernière définition correspond à un unique vecteur $\Pi x$ dans $F$. Ensuite, le $\Pi x$ de iii) vérifie $\langle x - \Pi x|e_i\rangle = \langle x|e_i\rangle - \langle x|e_i\rangle\langle e_i|e_i\rangle = 0$ et donc vérifie la description i). Pour conclure que i) et iii) sont équivalentes, il suffit donc de montrer qu'il n'existe pas plus d'un vecteur $y$ de $F$ tel que $x - y$ est orthogonal à $F$. Imaginons qu'il existe un autre vecteur $y'$ avec cette propriété. Alors $y - y'$ est dans $F$ et est égal à $(x - y') - (x - y)$ donc est orthogonal à $F$. Ceci n'est possible que si $y - y' = 0$, c'est-à-dire $y = y'$.

Supposons que $\Pi x$ vérifie $\|x - \Pi x\| = \min_{y\in F} \|x - y\|$. Alors, pour tout $y \in F$ et tout $\lambda \in \mathbb{R}$, on a $\|x - (\Pi x + \lambda y)\|^2 = \|x - \Pi x\|^2 + 2\lambda\langle x - \Pi x|y\rangle + |\lambda|^2\|y\|^2$. Quand $\lambda$ tend vers 0, on trouve que l'on doit avoir $\lambda\langle x - \Pi x|y\rangle \geq 0$ pour que $\Pi x$ réalise bien le minimum $\min_{y\in F} \|x - y\|$. Comme cela est vrai pour $\lambda$ positif comme négatif, cela implique que $\langle x - \Pi x|y\rangle = 0$ pour tout $y \in F$ et donc que $x - \Pi x$ est orthogonal à $F$. Supposons maintenant que $\Pi x \in F$ vérifie que $x - \Pi x$ est orthogonal à $F$. Soit $y \in F$, on a $\|x - y\|^2 = \|x - \Pi x + \Pi x - y\|^2 = \|x - \Pi x\|^2 + \|\Pi x - y\|^2$ par le théorème de Pythagore et donc $\|x - y\| \geq \|x - \Pi x\|$ ce qui montre ii). Comme l'on sait déjà que $\Pi x$ est unique dans la définition i), l'équivalence ci-dessus justifie l'unicité dans la définition ii). □

L'inérêt du théorème 4.9 est de donner une meilleure approximation qui soit calculable rapidement par des produits scalaires. C'est souvent pour cela que les approximations « par moindres carrés », qui sont liées à des produits scalaires, sont plus utilisées.

**Exemple d'application : la régression linéaire.** On mesure une variable $x(t)$ sur une suite de temps $t_1$, …, $t_n$, avec $n$ assez grand, et on trouve des valeurs $x_1$, …, $x_n$ . On sait que, en théorie, $x(t)$ est une droite $at + b$ et on aimerait retrouver $a$ et $b$ à partir des mesures. Comme il y a des erreurs et imprécisions sur les mesures, les points $(t_n,x_n)$ ne sont pas vraiment sous la forme d'une droite. On aimerait trouver « la meilleure droite possible approchant les points $(t_n,x_n)$ », mais il faudrait savoir ce qu'on entend par là : la droite $at + b$ minimisant le plus grand écart entre $at_i + b$ et $x_i$ ? Celle minimisant l'écart moyen entre les $at_i + b$ et $x_i$ ? Ces deux critères correspondent respectivement à minimiser la distance au sens des normes $\ell^\infty$ et $\ell^1$. Cela pourrait se faire, mais il est bien plus pratique de chercher à minimiser la distance $\ell^2$ car elle provient d'un produit scalaire et le théorème de projection orthogonal est à notre disposition. C'est *une méthode des moindres carrées*, puisqu'on cherche à minimiser la somme des carrés des écarts.

Prenons le formalisme suivant : $E = \mathbb{R}^n$ est muni du produit scalaire canonique et on pose $x = (x_n)$. On considère le sous-espace $F$ de $E$ engendré par les vecteurs $\mathbb{1} = (1,1,\ldots,1)$ et $T = (t_1,\ldots,t_n)$. On souhaite trouver $a$ et $b$ tel que $\|x - aT + b\mathbb{1}\|$ soit minimum. Il suffit d'orthonormaliser la base $(\mathbb{1},T)$ en $(\tilde{e}_1,\tilde{e}_2)$ puis de calculer la projection orthogonale de $x$ sur $F$ à l'aide de cette base. En repassant à la base $(\mathbb{1},T)$, on obtient les coefficients $a$ et $b$. Notons que tous ces calculs, y compris le procédé d'orthonormalisation de Gram-Schmidt sont parfaitement algorithmiques et programmables. En fait, on peut montrer les formules

$$a = \frac{\sum_i (t_i - \bar{t})(x_i - \overline{x})}{\sum_i (t_i - \bar{t})^2} = \frac{\frac{1}{n}\sum_i t_i x_i - \bar{t}.\overline{x}}{\frac{1}{n}\sum_i t_i^2 - \bar{t}^2} = \frac{\text{covariance}(x,t)}{\text{variance}(t)} \qquad b = \overline{x} - a\bar{t}$$

où $\bar{t} = \frac{1}{n}\sum_i t_i$ et $\overline{x} = \frac{1}{n}\sum_i x_i$ sont les moyennes des $(t_i)$ et des $(x_i)$.

## 2.2 Polynômes orthogonaux de Legendre

On souhaite approcher une fonction $f \in \mathcal{C}^0([-1,1],\mathbb{R})$ de la meilleure façon possible par un polynôme de degré donné. Comme on l'a déjà discuté, il y a plusieurs notions d'approximation. Celle des moindres carrés a l'avantage que le théorème 4.9 nous donne un moyen simple de la calculer.

**Exemple de la méthode :** on veut approcher la fonction exponentielle par un polynôme de degré 4 sur $[-1,1]$. On part de la base $(1,X,X^2,X^3,X^4)$ et on lui applique la méthode de Gram-Schmidt. On peut faire les calculs à la main, mais un logiciel de calcul formel est aussi parfait pour cela.

```
1 P0(x):=1/sqrt(2);

        (x)->1/(sqrt(2))
```

```
2 P1:=unapply(x -int(t*P0(t),t=-1..1)*P0(x),x);
P1:=unapply(P1(x)/sqrt(int(P1(t)^2,t=-1..1)),x);
```

$$(x) -> x,(x) -> x * 1/(sqrt(6)) * 3$$

```
3 P2:=unapply(simplify(x^2 -int(t^2*P0(t),
t=-1..1)*P0(x)-int(t^2*P1(t),t=-1..1)*P1(x)),x);
P2:=unapply(simplify(P2(x)/sqrt(int(P2(t)^2,t=-1..1))),x);
```

$$(x) -> (3 * x^2 - 1)/3,(x) -> (3 * x^2 * sqrt(10) - (sqrt(10)))/4$$

```
4 P3:=unapply(simplify(x^3 -int(t^3*P0(t),
t=-1..1)*P0(x)-int(t^3*P1(t),t=-1..1)*P1(x) -int(t^
3*P2(t),t=-1..1)*P2(x)),x); P3:=unapply(simplify(P3(x)
/sqrt(int(P3(t)^2,t=-1..1))),x);
```

$$(x) -> (5 * x^3 - 3 * x)/5,(x) -> (5 * x^3 * sqrt(14) - 3 * x * sqrt(14))/4$$

```
5 P4:=unapply(simplify(x^4- int(t^4*P0(t),
t=-1..1)*P0(x)-int(t^4*P1(t),t=-1..1)*P1(x) -int(t^
4*P2(t),t=-1..1)*P2(x) -int(t^4*P3(t),t=-1..1)*P3(x)),x);
P4:=unapply(simplify(P4(x)/sqrt(int(P4(t)^2,t=-1..1))),x);
```

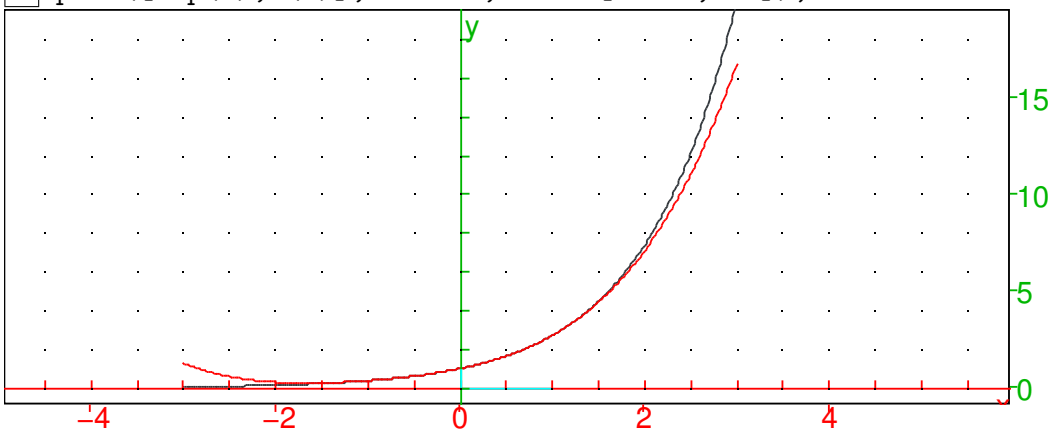
$$(x) -> (35 * x^4 - 30 * x^2 + 3)/35$$

$$(x) -> (105 * x^4 * sqrt(2) - 90 * x^2 * sqrt(2) + 9 * sqrt(2))/16$$

```
6 P:=unapply(simplify(int(exp(t)*P0(t),t=-1..1)*P0(x)+int(exp(t)*P1(t),
t=-1..1)*P1(x)+int(exp(t)*P2(t),t=-1..1)*P2(x)+int(exp(t)*P3(t),
t=-1..1)*P3(x)+int(exp(t)*P4(t),t=-1..1)*P4(x)),x);
```

$$(x) -> (5670 * x^4 * exp(1)^2 - 41895 * x^4 - 350 * x^3 * exp(1)^2$$
$$+2590 * x^3 - 4830 * x^2 * exp(1)^2 + 35700 * x^2 + 210 * x * exp(1)^2$$
$$-1530 * x + 480 * exp(1)^2 - 3525) * 1/8/exp(1)$$

```
7 plot([exp(x),P(x)],x=-3..3,color=[black,red]);
```




On note qu'on obtient bien une approximation parfaite sur $[-1,1]$ mais pas sur $[-3,3]$.

Les polynômes que nous avons calculés ne sont pas ceux que l'Histoire a conservée. La normalisation la plus standard est de prendre $P_i(1) = 1$. On appelle alors ces polynômes orthogonaux (mais pas orthonormaux) les *polynômes de Legendre*.

**Théorème 4.10.** *On muni $\mathbb{R}[X]$ du produit scalaire $\langle P|Q\rangle = \int_{-1}^{1} P(x)Q(x)dx$. Soit $(P_n)$ la base orthonormale obtenue par le procédé d'orthogonalisation de Gram-Schmidt à partir de $(1,X,X^2,\ldots)$ avec la normalisation $P_n(1) = 1$. Alors on a*

$$P_n(X) = \frac{1}{2^n\, n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left[(x^2-1)^n\right] \ .$$

***Démonstration :*** On note que la formule définit bien un polynôme de degré $n$ qui commence par le terme $X^n$. On note que les $n$ dérivées tapant sur $(x^2-1)^n$ peuvent taper soit sur un $(x^2-1)$ soit sur les termes sortant à cause des dérivées de $x^2$. Le seul terme ne s'annulant pas en $x = \pm 1$ est celui où toutes les dérivées tapent sur un $(x^2-1)$, ce qui donne $(2x)^n n!$ et donc $P_n(1) = 1$. Enfin, si on prend $m > n$, alors le calcul de $\int_{-1}^{1} P_m(x)P_n(x)dx$ peut se faire avec $m$ intégrations par parties. Les termes de bord sont nuls par un argument comme ci-dessus et $P_n$ n'est que de degré $n < m$ donc on obtient 0. □

On peut aussi montrer que $\int_{-1}^{1} |P_n(x)|^2 \,\mathrm{d}x = \frac{2}{2n+1}$, ce qui permet de faire la normalisation rapidement. On calcule facilement les polynômes de Legendre grâce à la formule du théorème.

```
8 pour n de 0 jusque 6 faire P(x):=(x^2-1)^n; pour j de 1
jusque n faire P:=unapply(simplify(diff(P(x),x)),x); ffaire;
P:=simplify(P/(2^n*n!));print(P(x))ffaire;
1
x
(3*x^2-1)/2
(5*x^3-3*x)/2
(35*x^4-30*x^2+3)/8
(63*x^5-70*x^3+15*x)/8
(231*x^6-315*x^4+105*x^2-5)/16
```

Il existe plein d'autres polynômes orthogonaux en fonction du produit scalaire utilisé :

| | |
|---|---|
| polynômes de Tchebychev | $\langle P\|Q\rangle = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} P(x)Q(x)\,\mathrm{d}x$ |
| polynômes de Hermite | $\langle P\|Q\rangle = \int_{-\infty}^{\infty} e^{-x^2} P(x)Q(x)\,\mathrm{d}x$ |
| polynômes de Laguerre | $\langle P\|Q\rangle = \int_{0}^{\infty} e^{-x} P(x)Q(x)\,\mathrm{d}x$ |

On peut montrer de nombreux résultats sur les polynômes orthogonaux. En voici un qui nous sera utile au chapitre suivant.

**Proposition 4.11.** *Soit $(P_n)$ la suite des polynômes orthogonaux obtenues par le procédé de Gram-Schmidt pour le produit scalaire $\langle P|Q\rangle = \int_a^b w(x)P(x)Q(x)\,\mathrm{d}x$ avec $w > 0$. Alors $P_n$, qui est de degré $n$ change $n$ fois de signe dans $]a,b[$ et en particulier a exactement $n$ zéros dans $]a,b[$.*

***Démonstration :*** Supposons que $P_n$ ne change que $k$ fois de signe dans $]a,b[$ avec $k < n$ et notons $a < x_1 < \ldots < x_k < b$ les points où le signe change. Soit $Q = (x - x_1)(x - x_2)\ldots(x - x_k)$. On a $Q$ de degré $k < n$ donc $Q$ doit être orthogonal à $P_n$. Mais $\langle P_n|Q\rangle = \int_a^b w(x)P_n(x)Q(x)\,\mathrm{d}x$ et $P_nQ$ est de signe constant sur $[a,b]$ et non nul, ce qui est absurde. □

# Chapitre 5 : Intégration numérique

Notre but est de calculer une intégrale comme $\int_a^b f(x)\,\mathrm{d}x$. Pour la majorité des fonctions, on ne connaît pas une primitive de $f$. La valeur de l'intégrale n'est en général accessible que numériquement.

## 1 Premières méthodes

### 1.1 Méthodes des rectangles

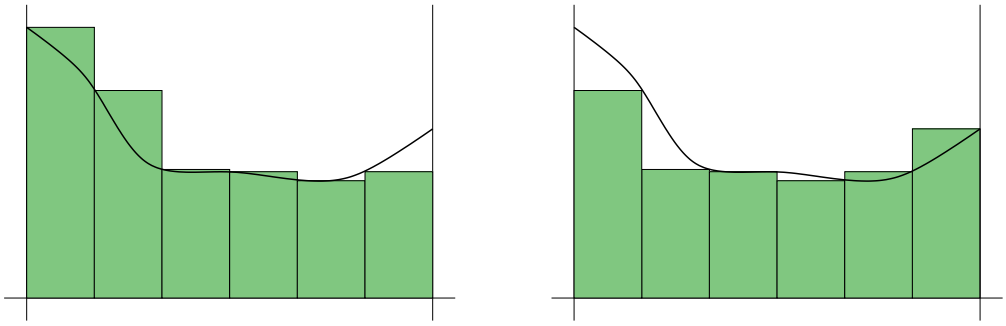
La méthode des rectangles à gauche consiste à dire que

$$\int_a^b f(x)\,\mathrm{d}x \;\simeq\; \frac{(b-a)}{n}\sum_{k=0}^{n-1} f(x_{k,n}) \qquad \text{avec } x_{k,n} = a + \frac{k}{n}(b-a)$$

et celle des rectangles à droite que

$$\int_a^b f(x)\,\mathrm{d}x \;\simeq\; \frac{(b-a)}{n}\sum_{k=1}^{n} f(x_{k,n}) \qquad \text{avec } x_{k,n} = a + \frac{k}{n}(b-a)\ .$$

On pose $h = (b-a)/n$, qui est le *pas*, c'est-à-dire l'espace entre deux points. Un dessin permet de bien comprendre cette approximation de façon géométrique.



*Les méthodes des rectangles à gauche (à gauche) et à droite (à droite) approchent l'aire sous la courbe par l'aire des rectangles verts.*

On a le résultat de convergence suivant.

**Théorème 5.1.** *Soit $[a,b] \subset \mathbb{R}$ et soit $x_{k,n} = a + \frac{k}{n}(b-a)$. Si $f$ est une fonction réglée sur $[a,b]$ (ou simplement de classe $\mathcal{C}^0([a,b],\mathbb{R})$), alors la méthode des rectangles à gauche converge dans le sens où*

$$\frac{(b-a)}{n}\sum_{k=0}^{n-1} f(x_{k,n}) \;\xrightarrow[n\longrightarrow+\infty]{}\; \int_a^b f(x)\,\mathrm{d}x$$

*et de même pour la méthode des rectangles à droite.*

*Si $f$ est dérivable sur [a,b] alors*

$$\left| \int_a^b f(x)\,\mathrm{d}x \;-\; \frac{(b-a)}{n} \sum_{k=0}^{n-1} f(x_{k,n}) \right| \;\leq\; h(b-a) \max_{x\in[a,b]} |f'(x)|$$

*et de même pour la méthode des rectangles à droite.*

***Démonstration :*** La première partie découle de la définition même de l'intégrale de Riemann. Pour obtenir la deuxième estimation, on note que sur chaque $[x_{k,n},x_{k+1,n}]$, on a $f(x) - f(x_{k,n}) = f'(\xi_x)(x - x_{k,n})$ et donc $|f(x) - f(x_{k,n})| \leq h \max_{x\in[x_{k,n},x]} |f'(x)|$. On obtient donc que

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x \;-\; \frac{b-a}{n} \sum_{k=0}^{n-1} f(x_{k,n}) \right| &= \left| \int_a^b f(x)\,\mathrm{d}x \;-\; \sum_{k=0}^{n-1} \int_{x_{k,n}}^{x_{k+1,n}} f(x_{k,n})\,\mathrm{d}x \right| \\
&= \left| \sum_{k=0}^{n-1} \int_{x_{k,n}}^{x_{k+1,n}} (f(x) - f(x_{k,n}))\,\mathrm{d}x \right| \\
&\leq \sum_{k=0}^{n-1} h^2 \max_{x\in[a,b]} |f'(x)| \\
&\leq h(b-a) \max_{x\in[a,b]} |f'(x)|
\end{aligned}$$

□

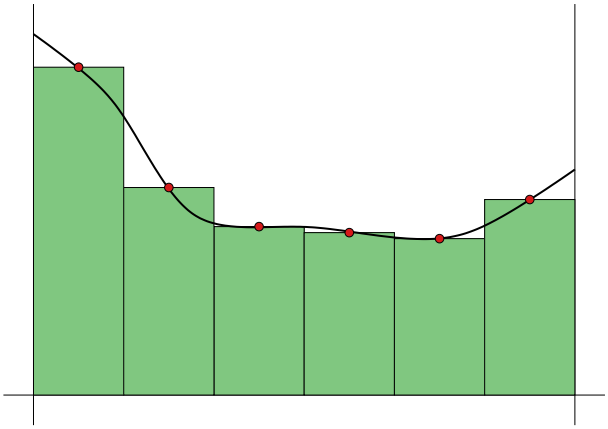
Testons la méthode des rectangles à gauche sur Xcas.

```
1 f(x):=cos(3*x); Iex:=1/3*sin(3);
2 n:=10; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f(j*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.14618629716
-0.0991462944733

3 n:=100; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f(j*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.0569864371164
-0.00994643442982

4 n:=1000; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f(j*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.0480349636543
-0.000994960967711
```

On remarque bien que la méthode converge avec une erreur proportionnelle au pas de discrétisation.

## 1.2 Méthode du point milieu

La méthode du point milieu consiste à prendre sur $[x_{k,n},x_{k+1,n}]$ la valeur de $f$ au point milieu plutôt que sur les bords.



On peut montrer les mêmes résultats de convergence que pour les méthodes des rectangles, mais on peut même gagner en précision.

**Théorème 5.2.** *Soit $[a,b] \subset \mathbb{R}$ et soit $x_{k,n} = a + \frac{k}{n}(b-a)$. Si $f$ est de classe $\mathcal{C}^2([a,b],\mathbb{R})$ alors*

$$\left| \int_a^b f(x)\,\mathrm{d}x \;-\; \frac{(b-a)}{n} \sum_{k=0}^{n-1} f\left(\frac{x_{k,n}+x_{k+1,n}}{2}\right) \right| \;\leq\; h^2 \frac{(b-a)}{24} \max_{x\in[a,b]} |f''(x)| \;.$$

***Démonstration :*** On utilise le développement à l'ordre deux : sur $[x_{k,n},x_{k+1,n}]$, on a

$$f(x) = f(m_{k,n}) + f'(m_{k,n})(x-m_{k,n}) + f''(\xi_x)\frac{(x-m_{k,n})^2}{2}$$

avec $m_{k,n} = \frac{x_{k,n}+x_{k+1,n}}{2}$. On obtient donc que

$$\begin{aligned}
\left| \int_{x_{k,n}}^{x_{k+1,n}} f(x)\,\mathrm{d}x \;-\; \frac{b-a}{n} f(m_{k,n}) \right| &= \left| \int_{x_{k,n}}^{x_{k+1,n}} (f(x) - f(m_{k,n}))\,\mathrm{d}x \right| \\
&= \left| \int_{x_{k,n}}^{x_{k+1,n}} f'(m_{k,n})(x-m_{k,n}) + f''(\xi_x)\frac{(x-m_{k,n})^2}{2}\,\mathrm{d}x \right| \\
&\leq \frac{1}{3}\left(\frac{h}{2}\right)^3 \max_{x\in[x_{k,n},x^n_{n+1}]} |f''(x)| \\
&\leq \frac{h^3}{24} \max_{x\in[x_{k,n},x^n_{n+1}]} |f''(x)|
\end{aligned}$$

où on a utilisé que $\int_{x_{k,n}}^{x_{k+1,n}} (x-m_{k,n}) = 0$ par imparité. Il ne reste plus qu'à appliquer cette estimation sur chacun des $n$ intervalles. □

Le test sous Xcas nous confirme bien que cette méthode est plus rapide que les rectangles car converge comme le carré du pas.

```
5 n:=10; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f((j+1/2)*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.0472168668478
-0.000176864161198

6 n:=100; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f((j+1/2)*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.047041766733
-1.76404635388e-06

7 n:=1000; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*f((j+1/2)*h)); ffaire; print(I); print(evalf(Iex-I));
I :0.047040020326
-1.76394221452e-08
```

# 2 Les méthodes composées

On veut approcher une intégrale d'une fonction $f$ sur un segment $[a,b]$. Les méthodes composées consistent au procédé suivant. On introduit d'abord une *méthode élémentaire* qui consiste à approcher une intégrale sur $[-1,1]$ par une combinaison de valeurs :

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \;\simeq\; 2\sum_{k=1}^{K} \omega_k f(x_k)\ .$$

Puis on découpe $[a,b]$ en $N$ segments $[\alpha_n,\alpha_{n+1}]$ avec $\alpha_n = a + hn$ où $h = (b-a)/N$ est le *pas de discrétisation*. Dans chaque segment, on applique la méthode élémentaire après une translation et homothétie :

$$\int_{\alpha_n}^{\alpha_{n+1}} f(x)\,\mathrm{d}x \;\simeq\; h\sum_{k=1}^{K} \omega_k f(x_{k,n}) \quad \text{avec } x_{k,n} = \frac{\alpha_n + \alpha_{n+1}}{2} + \frac{h}{2}x_k\ .$$

On obtient donc la méthode

$$\int_{a}^{b} f(x)\,\mathrm{d}x \;\simeq\; h\sum_{n=0}^{N-1}\sum_{k=1}^{K} \omega_k f(x_{k,n})\ .$$

**Définition 5.3.** *Une méthode est dite d'ordre $p$ si elle est exacte sur les polynômes de degré $p$ mais pas sur les polynômes d'ordre plus élevé.*

**Exemples :**

- Les méthodes des rectangles sont des méthodes composées avec un point $x_1 = \pm 1$ et $\omega_1 = 1$. Elles sont d'ordre 0 car elles sont exactes seulement sur les fonctions $f$ constantes.
- La méthode du point milieu est une méthode composée avec un point $x_1 = 0$ et $\omega_1 = 1$. Elle est d'ordre 1 car elle est exacte aussi sur les polynômes de degré 1 (vu que $\int_{-1}^{1} x\,\mathrm{d}x = 0$).

**Théorème 5.4.** *On considère une méthode composée*

$$\int_a^b f(x)\,\mathrm{d}x \;\simeq\; h\sum_{n=0}^{N-1}\sum_{k=1}^{K}\omega_k f(x_{k,n})$$

*que l'on suppose d'ordre $p$. Alors, si $f$ est de classe $\mathcal{C}^{p+1}([a,b],\mathbb{R})$, on a*

$$\left|\int_a^b f(x)\,\mathrm{d}x \;-\; h\sum_{n=0}^{N-1}\sum_{k=1}^{K}\omega_k f(x_{k,n})\right| \;\leq\; h^{p+1}\frac{(b-a)}{2^{p+1}(p+1)!}\left(1+\sum_{k=1}^{K}|\omega_k|\right)\max_{x\in[a,b]}|f^{(p+1)}(x)|\,.$$

*On remarque en outre que si les coefficients $\omega_k$ sont positifs, alors $\sum|\omega_k|=1$.*

***Démonstration :*** Sur le segment $[\alpha_n,\alpha_{n+1}]$, on note $m_n=(\alpha_n+\alpha_{n+1})$ le milieu. On a

$$f(x)=f(m_n)+f'(m_n)(x-m_n)+\ldots+\frac{f^{(p)}(m_n)}{p!}(x-m_n)^p+\int_{m_n}^{x}\frac{(x-t)^p}{p!}f^{(p+1)}(t)\,\mathrm{d}t\,.$$

Comme la dérivée $(p+1)-$ième de $f$ est bornée sur $[a,b]$, le reste intégral se borne par

$$\left|f(x)-\sum_{j=0}^{p}\frac{f^{(j)}(m_n)}{j!}(x-m_n)^j\right|\leq\frac{h^{p+1}}{2^{p+1}(p+1)!}\max_{x\in[a,b]}|f^{(p+1)}(x)|\,.$$

Ce qui approche $f$ est un polynôme de degré $p$, donc la méthode d'intégration est exacte dessus. En remplaçant $f$ par cette approximation, on trouve

$$\begin{aligned}\left|\int_{\alpha_n}^{\alpha_{n+1}} f(x)\,\mathrm{d}x \;-\; h\sum_{k=1}^{K}\omega_k f(x_{k,n})\right| &\leq \int_{\alpha_n}^{\alpha_{n+1}}\frac{h^{p+1}}{2^{p+1}(p+1)!}\max|f^{(p+1)}|\,\mathrm{d}x\\ &\quad+h\sum_{k=1}^{K}|\omega_k|\frac{h^{p+1}}{2^{p+1}(p+1)!}\max|f^{(p+1)}|\\ &\leq\left(1+\sum_{k=1}^{K}|\omega_k|\right)\frac{h^{p+2}}{2^{p+1}(p+1)!}\max|f^{(p+1)}|\end{aligned}$$

En faisant la somme des $N=(b-a)/h$ segments, on obtient une erreur de taille $\mathcal{O}(h^{p+1})$.

Pour finir, remarquons que si $\omega_k\geq 0$, alors $\sum_{k=1}^{K}|\omega_k|=\sum_{k=1}^{K}\omega_k=1$ car la méthode est exacte pour le polynôme constant égal à 1. □

On note que l'on retrouve bien les estimations des méthodes des rectangles et point milieu.

# 3 Méthodes de Newton-Cotes

Les méthodes de Newton-Cotes sont des méthodes composées basées sur l'interpolation de Lagrange. D'après le paragraphe précédent, il suffit de nous concentrer sur la méthode élémentaire sur $[-1,1]$. Pour ce faire, on dispose $K$ points équirépartis dans $[-1,1]$ (avec

$K \geq 2$) : $x_k = -1 + 2(k-1)/K$. Puis on remplace $f$ par son interpolation de Lagrange en ces points

$$f(x) \simeq \sum_{k=1}^{K} f(x_k) \prod_{j \neq k} \frac{x - x_j}{x_k - x_j} .$$

Enfin, on obtient

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \simeq 2\sum_{k=1}^{K} \omega_k f(x_k) \quad \text{avec } \omega_k = \frac{1}{2}\int_{-1}^{1} \prod_{j \neq k} \frac{x - x_j}{x_k - x_j}\,\mathrm{d}x .$$

On obtient immédiatement que la méthode est exacte pour les polynômes de degré plus petit que $K-1$. On peut en fait dire un peu mieux.

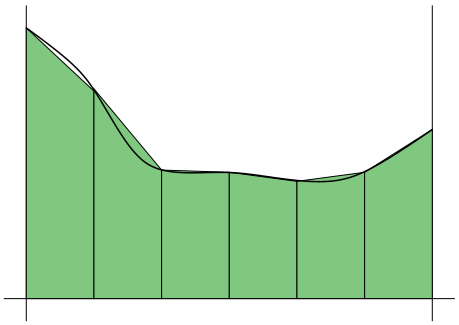
**Proposition 5.5.** *Si $K$ est pair, la méthode est d'ordre $K-1$. Si $K$ est impair, alors la méthode est d'ordre $K$.*

***Démonstration :*** On a déjà vu que la méthode est par définition exacte sur les polynômes de degré plus petit que $K-1$ car l'interpolation de Lagrange est exacte. Si $K$ est impair, alors les points $x_k$ sont symétriques par rapport à 0 et les $\omega_k$ sont égaux en les points symétriques (petits calculs de symétrie). Donc non seulement $\int_{-1}^{1} x^K\,\mathrm{d}x = 0$, mais aussi $\sum \omega_k x_k^K = 0$ par symétrie. Donc le calcul est exact encore pour les polynômes de degré $K$. Il faut encore bricoler un polynôme de degré plus grand pour lequel la méthode ne marche pas, mais nous allons admettre ceci. □

Pour $K = 2$, on obtient la méthode des trapèzes :

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \simeq 2\frac{f(-1)+f(1)}{2}$$

qui ressemble, autant pour la formule que pour l'ordre, à la méthode du point milieu.



Pour $K \geq 3$, la proposition ci-dessus fait que l'on n'utilise que des méthodes avec $K$ impair. La méthode pour $K = 3$ est appelée la *méthode de Simpson* d'ordre 3 :

$$\begin{array}{lll} x_1 = -1 & x_2 = 0 & x_3 = 1 \\ \omega_1 = 1/6 & \omega_2 = 2/3 & \omega_3 = 1/6 \end{array}$$

Puis vient la méthode de Boole-Villarceau :

$$\begin{array}{lllll} x_1 = -1 & x_2 = -1/2 & x_3 = 0 & x_2 = 1/2 & x_3 = 1 \\ \omega_1 = 7/90 & \omega_2 = 16/45 & \omega_3 = 2/15 & \omega_2 = 16/45 & \omega_3 = 7/90 \end{array}$$

et la méthode de Weddle-Hardy d'ordre 7. A partir de $K = 8$, on obtient des poids $\omega_k$ négatifs, ce qui devient problématique pour les erreurs. On n'utilise donc pas ces méthodes. Pour gagner en précision, il faut jouer sur le pas $h$ et non pas le nombre de points de la méthode élémentaire.

Faisons un essai de la méthode de Simpson sous Xcas pour vérifier que cette méthode est bien d'ordre 3, c'est-à-dire converge comme $h^{-4}$.

```
8 n:=10; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*(f(j*h)/6+2/3*f((j+1/2)*h)+f((j+1)*h)/6)); ffaire;
afficher(I); afficher(evalf(Iex-I));
I :0.0470401353418
-1.32655219032e-07

9 n:=100; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*(f(j*h)/6+2/3*f((j+1/2)*h)+f((j+1)*h)/6)); ffaire;
afficher(I); afficher(evalf(Iex-I));
I :0.0470400026998
-1.3175682767e-11
```

# 4 Méthodes de Gauss-Legendre

On a obtenu des méthodes d'ordre $K$ avec $K$ points. Mais peut-on faire mieux ? On peut déjà voir que l'on sera limité.

**Proposition 5.6.** *Soit $\int_{-1}^{1} f(x)\,\mathrm{d}x \simeq 2\sum_{k=1}^{K} \omega_k f(x_k)$ une méthode à $K$ points. Alors elle ne peut être d'ordre $p \geq 2K$.*

***Démonstration :*** On pose $P = (x-x_1)^2 \dots (x-x_K)^2$ qui est un polynôme de degré $2K$, qui est positif non nul et $\sum_{k=1}^{K} \omega_k P(x_k) = 0$. Donc la méthode n'est pas exacte sur $P$. □

Les méthodes de Gauss-Legendre consiste à optimiser au mieux selon cette limitation.

**Théorème 5.7.** *Soit $K \geq 1$, il existe une unique méthode élémentaire $\int_{-1}^{1} f(x)\,\mathrm{d}x \simeq 2\sum_{k=1}^{K} \omega_k f(x_k)$ qui est exacte sur tous les polynômes de degré $p \leq 2K-1$.*
*En outre, les points $x_k$ sont les zéros du polynôme de Legendre de degré $K$ et*

$$\omega_k = \frac{1}{2}\int_{-1}^{1} L_k(x) \qquad \text{avec } L_k = \prod_{j\neq k} \frac{x-x_j}{x_k-x_j} \text{ le polynôme interpolateur de Lagrange.}$$

***Démonstration :*** On va raisonner par analyse-synthèse. Imaginons que notre méthode optimale existe. Soit $P_K$ le polynôme de Legendre de degré $K$, qui est orthogonal à tous les polynômes de degré plus petit. Donc

$$\int_{-1}^{1} Q(x)P_K(x)\,\mathrm{d}x \;=\; 2\sum_{k=1}^{K} \omega_k Q(x_k)P(x_k) \;=\; 0$$

pour tout polynôme $Q \in \mathbb{R}_{K-1}[X]$. Comme on peut, par interpolation de Lagrange, obtenir toutes les valeurs possibles de $(Q(x_k))_{k=1..K}$, c'est que $P(x_k) = 0$ pour tout $k$. Puis la définition des $\omega_k$ est imposée par

$$\int_{-1}^{1} L_{k_0}(x)\,\mathrm{d}x \;=\; 2\sum_{k=1}^{K} \omega_k L_{k_0}(x_k) = 2L_{k_0}(x_{k_0})$$

avec $L_{k_0}$ le polynôme de Lagrange qui vaut 0 sur les $x_j$ et 1 sur $x_{k_0}$.

Faisons maintenant la synthèse. On sait (cf chapitre précédent) que $P_K$ a $K$ zéros distincts dans $]-1{,}1[$ et on peut donc choisir $x_k$ comme ces zéros puis $\omega_k$ comme ci-dessus. Prenons maintenant $P$ de degré au plus $2K-1$. Par division euclidienne, il s'écrit

$$P = QP_K + R = QP_K + \sum_{k=1}^{K} R(x_k)L_k$$

avec $R$ de degré au plus $K-1$. Mais $R(x_k) = P(x_k)$ puisque $P_K(x_k) = 0$ et de plus

$$\begin{aligned}\int_{-1}^{1} P(x)\,\mathrm{d}x &= \int_{-1}^{1} Q(x)P_K(x) + \sum_{i=k}^{K} R(x_k)L_k(x)\,\mathrm{d}x \\ &= \langle Q|P_K\rangle + \sum_{k=1}^{K} R(x_k)\int_{-1}^{1} L_k(x)\,\mathrm{d}x = 0 + \sum_{k=1}^{K} P(x_k)2\omega_k\ .\end{aligned}$$

Donc la méthode est bien d'ordre $2K-1$. □

Avec un seul point, l'optimisation est donc la méthode du point milieu. Avec deux points, on peut déjà aller à l'ordre 3

$$\begin{array}{ll} x_1 = -1/\sqrt{3} & x_2 = 1/\sqrt{3} \\ \omega_1 = 1/2 & \omega_2 = 1/2 \end{array}$$

et avec 3 points, on obtient de l'ordre 5.

$$\begin{array}{lll} x_1 = -\sqrt{3/5} & x_2 = 0 & x_3 = \sqrt{3/5} \\ \omega_1 = 5/18 & \omega_2 = 4/9 & \omega_3 = 5/18 \end{array}$$

Testons cette dernière méthode qui permet d'avoir une erreur pour la méthode composée en $\mathcal{O}(h^6)$.

```
10 n:=1; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*(5/18*f((j+(1/2-sqrt(3/20)))*h) +4/9*f((j+1/2)*h)
+5/18*f((j+(1/2+sqr t(3/20)))*h))); ffaire; afficher(I);
afficher(evalf(Iex-I));
I :0.0470638758661
-2.38731794358e-05

11 n:=10; I:=0; h:=1/n; pour j de 0 jusque (n-1) faire
I:=I+evalf(h*(5/18*f((j+(1/2-sqrt(3/20)))*h) +4/9*f((j+1/2)*h)
+5/18*f((j+(1/2+sqr t(3/20)))*h))); ffaire; afficher(I);
afficher(evalf(Iex-I));
I :0.0470400027037
-1.70652381115e-11
```

On note aussi que la méthode se généralise à des situations avec des poids différents, par exemple on peut ainsi calculer des intégrales généralisée du type $\int_{\mathbb{R}} f(x)e^{-x^2}\,\mathrm{d}x$.

# Chapitre 6 : Discrétisation des Équations Différentielles Ordinaires

## 1 Principe général

On veut approcher les solutions d'une équation différentielle autonome dans $\mathbb{R}^d$

$$x(0) = x_0 \quad \text{et} \quad \forall t \in [0,T[ \;\; \dot{x}(t) = f(x(t)) \tag{6.1}$$

avec $T > 0$ un temps jusqu'auquel la solution $x(t)$ va exister. On supposera que $f \in \mathcal{C}^1(\mathbb{R}^d,\mathbb{R}^d)$ et en particulier est lipschitzienne en $x$ sur les bornés. D'après le théorème de Cauchy-Lipschitz, il existe une solution $x(t)$ locale sur un intervalle $[0,T[$.

Le but est d'approcher cette solution en discrétisant le problème. On choisit un pas de temps $h > 0$ et on pose $t_n = hn$. On part de $x_0$ et on calcule par récurrence

$$x_{n+1} = x_n + h\Phi(x_n,h) \tag{6.2}$$

où $\Phi \in \mathcal{C}^0(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+)$ est une méthode à définir. On va voir que ces méthodes sont très liées aux méthodes d'intégration puisque $x(t)$ est solution de (6.1) si et seulement si

$$x(t+h) = x(t) + \int_0^h f(x(t+\tau))\,\mathrm{d}\tau = x(t) + h\int_0^1 f(x(t+hs))\,\mathrm{d}s \ . \tag{6.3}$$

Ce chapitre n'est qu'une introduction. Nous allons admettre les résultats suivants et passer les définitions plus précises. Nous nous concentrons aussi sur les équations autonomes pour simplifier une partie des notations.

**Définition 6.1.** *On dit que la méthode associée à $\Phi$ est d'ordre au moins $p$ si*

$$\forall 0 \le k \le p-1 \ , \ \ \frac{\partial^k}{\partial h^k}\Phi(x,0) \ = \ \frac{1}{k+1}\Theta^k f(x) \qquad \textit{avec } \Theta g = (Dg).g \ .$$

On note que si $x(t)$ est une solution de (6.1), alors

$$\dot{x}(t) = f(x(t)) = (\Theta^0 f)(x(t)) \ , \ \ \ddot{x}(t) = Df(x(t)).\dot{x}(t) = Df(x(t)).f(x(t)) = (\Theta^1 f)(x(t)) \ \ldots$$

et ceci explique l'apparition de cet opérateur $\Theta$ par le développement au moins formel

$$\begin{aligned} x(t+h) &= x(t) + \sum_{k\ge1} (\Theta^{k-1}f)(x(t))\frac{h^k}{k!} = x(t) + h\sum_{k\ge0}\frac{\partial^k}{\partial h^k}\Phi(x,0)\frac{h^k}{k!} \\ &= x(t) + \sum_{k\ge1}\frac{\partial^{k-1}}{\partial h^k}\Phi(x,0)\frac{h^k}{(k-1)!} \end{aligned}$$

On admettra le résultat suivant.

**Théorème 6.2.** *Soit $x(t)$ la solution exacte de* (6.1) *sur un intervalle* $[0,T[$. *On suppose que la méthode associée à $\Phi$ est d'ordre au moins $p$ et que $\Phi$ est lipschitzienne sur un voisinage de $\{x(t)\}$. Alors il existe une constante $C$ telle que la suite définie par* (6.2) *vérifie*

$$\forall t_n \leq T \ , \ \ |x(t_n) - x_n| \leq Ch^p \ .$$

# 2 Quelques méthodes

## 2.1 Méthode d'Euler explicite

L'idée la plus simple pour avoir une méthode est de prendre $\Phi(x_n,h) = f(x_n)$. On a alors

$$x_{n+1} = x_n + hf(x_n) \ .$$

On a donc remplacé $\int_0^h f(x(t+\tau))\mathrm{d}\tau$ par $hf(x(t))$, c'est-à-dire qu'il s'agit d'une méthode des rectangles à gauche. On peut aussi dire qu'on a écrit que $f(x_n) \simeq \dot{x}(t_n) \simeq (x(t_{n+1}) - x(t_n))/h$.

Il s'agit d'une méthode d'ordre 1 car

$$\Phi(x,0) = f(x) \quad \text{et} \quad \frac{\partial}{\partial h}\Phi(x,0) = 0 \neq \frac{1}{2}Df(x).f(x) \ .$$

## 2.2 Méthodes avec une pente

On peut approcher aussi $\int_0^h f(x(t+\tau))\mathrm{d}\tau$ par $hf(x(t+\tau))$ pour n'importe quel $t+\tau$. Pour $\tau = h$ on a une méthode type rectangles à droite, pour $\tau = h/2$, il s'agit d'un point milieu. Le problème est d'obtenir ces valeurs. Pour cela, on commence par dire que $x(t+\tau) \simeq x(t) + \tau f(x(t))$ et on calcule la pente $p = f(x(t) + \tau f(x(t))$. Puis on aura $x(t+h) \simeq x(t) + hp$. On obtient donc, pour chaque $\alpha \in [0,1]$, une méthode décrite par

$$\Phi(x,h) = f(x + \alpha h f(x)) \ .$$

On a $\Phi(x,0) = f(x)$ donc la méthode est d'ordre au moins 1. On calcule ensuite

$$\frac{\partial}{\partial h}\Phi(x,0) = \alpha Df(x).f(x) = \alpha(\Theta f)(x) \ .$$

Donc la méthode est d'ordre au moins 2 seulement pour $\alpha = 1/2$, c'est-à-dire pour le point milieu. Cette méthode n'est pas d'ordre 3 car alors

$$\frac{\partial^2}{\partial h^2}\Phi(x,0) = \alpha^2 D^2 f(x)(f(x),f(x)) \neq \frac{1}{3}D(Df.f).f = \frac{1}{3}D^2 f(f,f) + Df.Df.f \ .$$

## 2.3 Méthodes de Heun et de Runge-Kutta

On peut s'inspirer de la méthode d'intégration des trapèzes en utilisant deux pentes en $x(t)$ et $x(t+h)$. Pour approcher la pente en $x(t+h)$, on va utiliser la pente en $x(t)$. On obtient

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

On calcule de même que la méthode est d'ordre 2 comme celle du point milieu.

En fait les méthodes que l'on a vues font partie d'une série de méthodes baptisées *méthodes de Runge-Kutta*. Elles consistent à enchaîner les calculs de différentes pentes, chacune utilisant les pentes précédentes, puis de combiner le tout sous la forme d'une méthode proche d'une méthode d'intégration. La méthode de Runge-Kutta la plus classique est la suivante :

$$\begin{aligned} p_1 &= f(x) \\ p_2 &= f(x + p_1 h/2) \quad \text{calcul au point milieu} \\ p_3 &= f(x + p_2 h/2) \quad \text{calcul au point milieu avec la nouvelle pente} \\ p_4 &= f(x + p_3 h) \quad \text{calcul au point à droite} \\ \Phi(x,h) &= \frac{1}{6}p_1 + \frac{2}{6}p_2 + \frac{2}{6}p_3 + \frac{1}{6}p_4 \quad \text{combinaison du type Simpson.} \end{aligned}$$

On peut montrer qu'il s'agit d'une méthode d'ordre 4. C'est la méthode la plus utilisée quand on veut faire de l'ordre élevé car elle reste assez simple et est stable (coefficients positifs).

## 2.4 Méthode d'Euler implicite

On a parlé de méthode d'Euler explicite, c'est donc qu'il y en a une implicite. Elle consiste à définir non pas $x_{n+1} = x_n + hf(x_n)$ mais

$$x_{n+1} = x_n + hf(x_{n+1}) \ .$$

Le problème est donc que la définition de $x_{n+1}$ se fait par une équation non linéaire. On peut par exemple la résoudre par une méthode de Newton. On peut aussi voir que si $f$ est lipschitzienne, alors pour $h$ assez petit $x \mapsto x_n + hf(x)$ est contractante et $x_{n+1}$ est donc son unique point fixe que l'on obtient en itérant la fonction en partant de $x_n$.

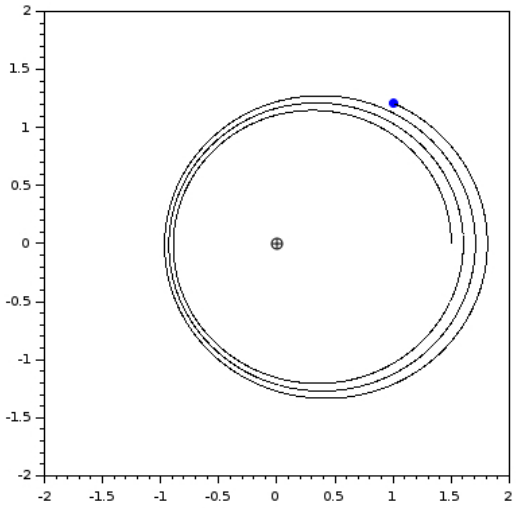
# 3 Quelques illustrations

Dans un premier temps, nous allons nous intéresser au mouvement d'un planète autour du Soleil, placé au point (0,0). Les équations de Newton nous disent que la trajectoire $x(t)$ dans $\mathbb{R}^2$ est telle que

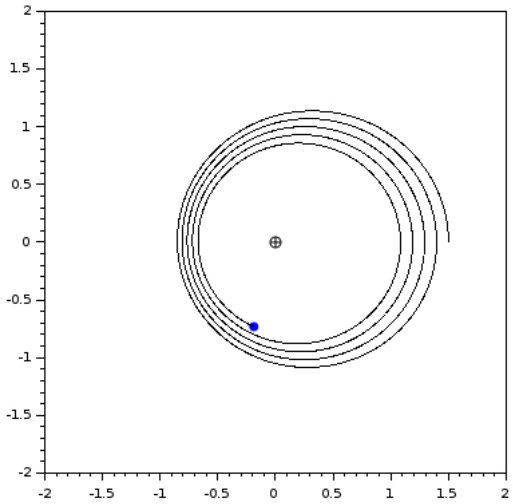
$$\frac{d}{dt}\begin{pmatrix} x(t) \\ \dot{x}(t) \end{pmatrix} = \begin{pmatrix} \dot{x}(t) \\ -\mathcal{G}\frac{x(t)}{\|x(t)\|^3} \end{pmatrix}$$

Il s'agit donc bien d'une équation différentielle qui rentre dans notre cadre, du moment que $x(t) \neq (0,0)$ c'est-à-dire que la planète reste loin du Soleil. On note aussi que la vraie variable est le vecteur d'état $(x,\dot{x})$ car l'équation est d'ordre 2.

Commençons avec les méthodes d'Euler

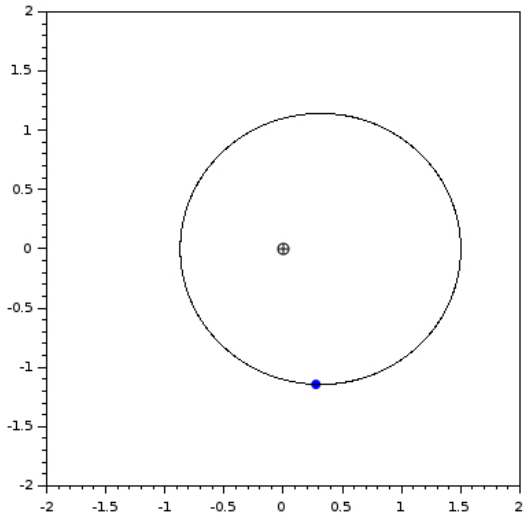



*Méthode d'Euler explicite*



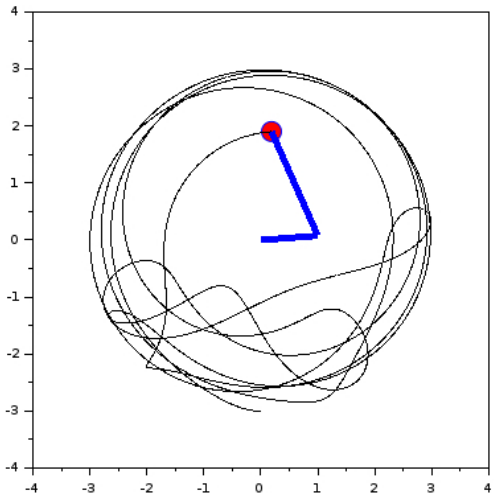

*Méthode d'Euler implicite*

On remarque que la méthode d'Euler explicite produit une planète qui s'éloigne du Soleil. En effet, la méthode remplace la trajectoire entre $t$ et $t + h$ par la tangente en $t$ qui est toujours trop sortante par rapport à la trajectoire. L'erreur est donc toujours dans le même sens. A l'inverse, la méthode implicite utilise la tangente en $t + h$ qui est toujours rentrante et la planète se rapproche du Soleil. En toute logique, le point milieu est un bon compromis et on peut même montrer qu'il conserve exactement l'énergie physique.




*Méthode du point milieu*

La méthode de Runge-Kutta est parfaitement adaptée quand on souhaite avoir une grande précision comme dans les cas contenant des phénomènes chaotique comme le pendule double.




*Méthode de Runge-Kutta-4 pour un pendule double chaotique*