8 Résolution numérique d'équations

Le but de ce chapitre est d'exposer quelques méthodes de résolutions d'équations $f(x) = 0$, où $f : I \to \mathbb{R}$ est une fonction définie sur un intervalle I. La fonction f sera supposée continue (et pour certaines méthodes un peu plus que cela même).
L'étude de ces méthodes passera par la description algorithmique de la méthode, permettant l'implémentation dans un langage de programmation, puis par les justifications de convergence et de rapidité. Pour assurer la convergence, il est parfois nécessaire d'ajouter des hypothèses (assurant l'existence d'une solution, ce qui est souvent obtenu par le TVI, ou même parfois, assurant l'existence d'une solution suffisamment proche de la valeur d'initialisation, ce qui nécessite alors une localisation grossière préalable). Par ailleurs, en cas de non unicité des solutions, même en cas de convergence, il n'est pas toujours possible de savoir vers quel zéro on aura convergence, si on n'effectue pas un prétraitement permettant de séparer les racines.

Les méthodes que nous étudions sont tout d'abord la dichotomie (cette méthode est d'ailleurs une preuve possible du TVI assurant l'existence d'une solution), puis la méthode de la sécante, et la version « limite » de cette méthode, qui est la méthode de Newton. La méthode de Newton est en quelque sorte l'aboutissement de la méthode de la sécante et assure une convergence plus rapide, mais elle nécessite la connaissance de la dérivée. Cette dérivée peut être donnée explicitement par l'utilisateur, ou peut être obtenue numériquement (mais cela nécessite que la courbe soit localement suffisamment lisse, car les microvariations ne pourront pas être prises en considération par une méthode numérique). Nous abordons le problème de la dérivation numérique dans la dernière partie de ce chapitre.

La dichotomie et la méthode de Newton sont au programme de l'informatique commune de MPSI/PCSI. La méthode de la sécante, ainsi que la dérivation numérique, sont hors-programme.

I Dichotomie

La dichotomie est une des preuves possibles de la démonstration du théorème des valeurs intermédiaires. Le principe est d'encadrer de plus en plus finement un zéro possible, sa présence étant détectée par un changement de signe de la fonction, supposée continue. Ainsi :

Méthode 8.1.1 (Description de la méthode de dichotomie)

- On initialise l'encadrement par deux valeurs $a_0 < b_0$ telles que f change de signe entre a_0 et b_0, c'est-à-dire $f(a_0)f(b_0) \leqslant 0$.
- On coupe l'intervalle en 2 en son milieu. Puisqu'il y a changement de signe sur l'intervalle, il y a changement de signe sur l'une des deux moitiés. On conserve la moitié correspondante.

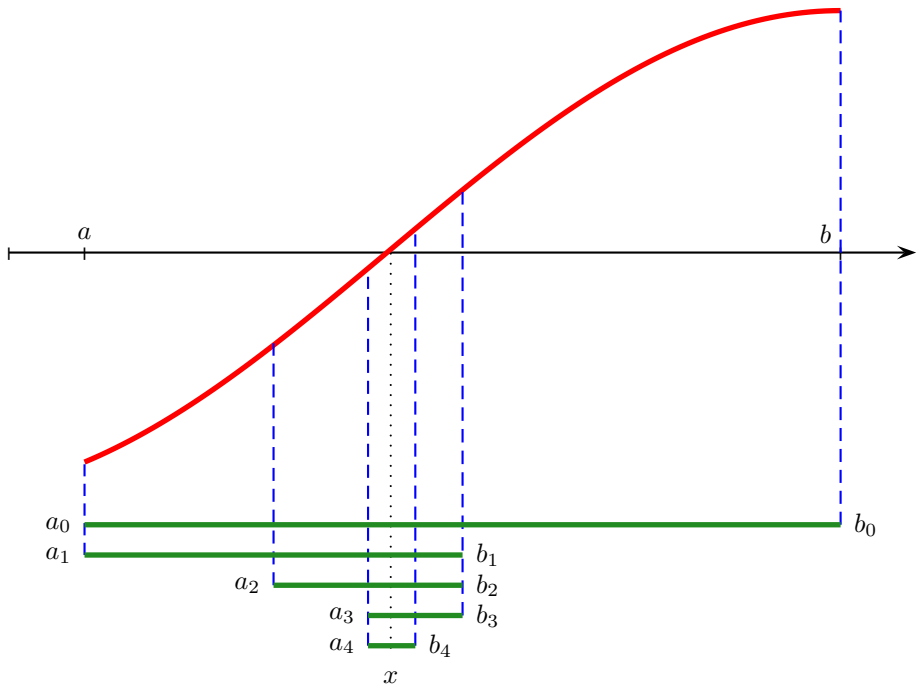

FIGURE 8.1 – Dichotomie

- On continue de la sorte en coupant à chaque étape l'intervalle en deux, en gardant la moitié sur laquelle il y a un changement de signe. On continue jusqu'à ce que l'intervalle obtenu soit de longueur suffisamment petite pour donner une valeur approchée du zéro à la marge d'erreur souhaitée.

La méthode est illustrée par la figure 8.1, et se traduit par l'algorithme suivant, en supposant initialement que $f(a)f(b) < 0$ (il faudrait faire un test et retourner une erreur si ce n'est pas le cas) :

Algorithme 8.1 : Dichotomie

Entrée : f : fonction ;
$a < b$: intervalle initial de recherche, tel que $f(a)f(b) < 0$;
ε : marge d'erreur
Sortie : x : valeur approchée à ε près d'un zéro de f sur $[a, b]$

```
tant que b − a > ε faire
    c ← (a+b)/2 ;
    si f(a)f(c) ⩽ 0 alors
        b ← c
    sinon
        a ← c
    fin si
fin tant que
renvoyer (a+b)/2
```

Remarquez qu'on obtient de la sorte une valeur approchée à $\frac{\varepsilon}{2}$ en fait : on pourrait se contenter de comparer $b - a$ à 2ε.

Proposition 8.1.2 (Validité et rapidité de l'algorithme de dichotomie)
La fonction f étant supposée continue, et en notant x la valeur retournée :

(i) L'algorithme s'arrête (la terminaison est assurée)

(ii) Il existe un zéro de f dans l'intervalle $]x-\varepsilon, x+\varepsilon[$

(iii) Le temps de calcul est en $O(-\ln(\varepsilon))$. Plus précisément, le nombre d'étapes est $\left\lceil \log_2\left(\frac{b-a}{\varepsilon}\right)\right\rceil$

(iv) On peut aussi dire que la convergence est linéaire (en $\Theta(n)$) en le nombre de décimales obtenues (donc en exprimant $\varepsilon = 10^{-n}$)

◁ **Éléments de preuve.**

(i) Les valeurs successives de $b-a$ suivent une progression géométrique de raison $\frac{1}{2}$, donc finissent par être inférieures à ε. Pouvez-vous trouver un variant de boucle ?

(ii) C'est l'inclusion $[a,b] \subset]x-\varepsilon, x+\varepsilon[$ associé au TVI. Quel invariant de boucle considérer pour justifier l'utilisation du TVI ?

(iii) Si a_n et b_n sont les valeurs des variables a et b au début de l'itération n, on a une relation simple entre $b_n - a_n$ et $b-a$, qui permet de faire les calculs explicitement.

(iv) Le nombre de décimales s'exprime avec $\log_{10}$ et $\log_{10}$ et $\log_2$ diffèrent d'un facteur multiplicatif constant.

▷

Remarque 8.1.3

On gagne un facteur 2 en précision à chaque itération. Ainsi, il faut un peu plus de 3 itérations pour gagner une décimale supplémentaire.
Par exemple, en partant d'un intervalle initial de longueur 1, on obtient une précision 10^{-10} au bout de 33 itérations.

Cela reste un bon algorithme, mais si f elle-même est longue à calculer, tout gain de complexité peut être important. Nous verrons un peu plus loin la méthode de Newton, bien plus efficace.

Remarque 8.1.4 (Comparatif de la méthode de dichotomie)

- Points forts :
  - simplicité,
  - convergence assurée,
  - vitesse raisonnable.
- Points faibles :
  - Nécessite un encadrement préalable du zéro ;
  - s'il y a plusieurs zéros dans l'intervalle, on n'en obtient qu'un
  - Moins rapide que la méthode de Newton, ou la méthode de la sécante.

II Méthode de la sécante (HP)

La méthode de la sécante consiste à remplacer dans le procédé de dichotomie le choix du point milieu par l'intersection entre l'axe des abscisses et la corde aux extrémités de l'intervalle souhaité, espérant ainsi obtenir une meilleure approximation du zéro recherché. Il faut alors choisir quelle moitié d'intervalle on conserve avant de réitérer l'opération ; Deux choix naturels peuvent se faire, définissant chacun une méthode de résolution :

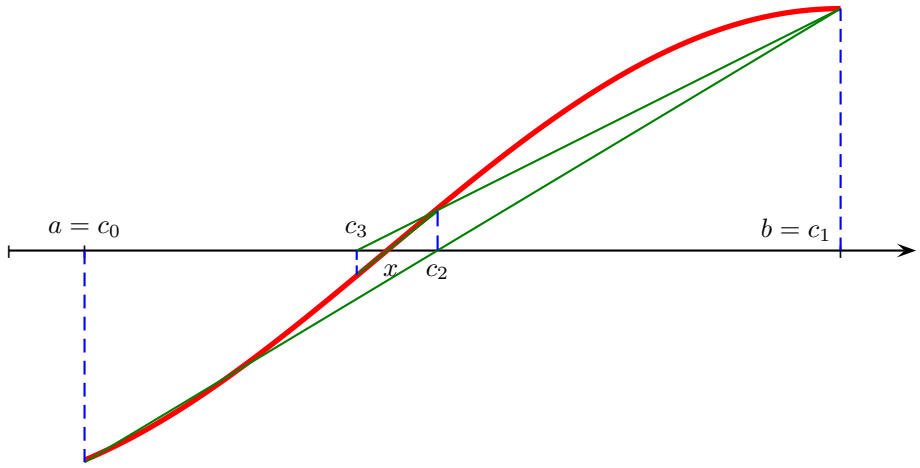

FIGURE 8.2 – Méthode de la sécante

- On peut décider de conserver l'intervalle assurant un changement de signe, et procéder comme dans l'algorithme de dichotomie. On obtient alors un algorithme appelé algorithme de la fausse position, ou *regula falsi*. On peut montrer qu'une des deux bornes de l'intervalle va converger vers la valeur recherchée, l'autre étant stationnaire (en général distincte de la valeur recherchée) : ainsi, contrairement à l'algorithme de la dichotomie, la longueur de l'intervalle considéré ne tend pas vers 0, ce qui pose le problème de la condition d'arrêt. Par ailleurs, même si sous des conditions raisonnables, on contrôle assez bien la complexité (à peu près linéaire en le nombre de décimales souhaitées), certaines situations où la courbe est plate au voisinage du zéro peuvent s'avérer catastrophiques.
- Un deuxième choix possible est de conserver systématiquement les deux dernières valeurs calculées, définissant de la sorte l'algorithme de la sécante. Ainsi, partant d'un intervalle initial $[a, b]$, et en notant $c_0 = a$ et $c_1 = b$, puis c_2 la première intersection, on conserve l'intervalle $[c_1, c_2]$ (éventuellement, les bornes sont dans l'autre sens), on calcule la valeur c_3, puis on conserve $[c_2, c_3]$ etc. Ainsi, l'existence d'un zéro dans l'intervalle conservé n'est pas assuré, et le point suivant peut sortir de l'intervalle précédent, voire le l'intervalle initial, et même du domaine de définition de f. Cela complique un peu l'étude de cet algorithme, en imposant d'étudier des conditions de convergence.

Méthode 8.2.1 (Méthode de la sécante, figure 8.2)

- On part d'un intervalle $[a, b] = [c_0, c_1]$ sur lequel f est continue. On n'impose pas de changement de signe sur cet intervalle.
- On calcule c_2 le point d'intersection de la corde avec l'axe des abscisses.
- On refait pareil sur l'intervalle $[c_1, c_2]$, puis $[c_2, c_3]$ etc, jusqu'à obtenir une approximation suffisante.
- Les intervalles $[c_n, c_{n+1}]$ sont à comprendre par $[c_{n+1}, c_n]$ si $c_{n+1} < c_n$.

Ainsi, on n'a cette fois qu'une suite à calculer, déterminée par la relation de récurrence exprimant l'intersection de la corde et de l'axe des abscisses :

$$c_{n+2} = c_{n+1} - f(c_{n+1}) \frac{c_{n+1} - c_n}{f(c_{n+1}) - f(c_n)}.$$

Remarque 8.2.2 (À propos de la condition d'arrêt)

- On n'a pas une condition d'arrêt aussi fiable que dans la méthode de dichotomie, du fait qu'on n'est pas assuré de la présence d'un zéro dans l'intervalle $[a, b]$. On peut faire un test de chan-

Algorithme 8.2 : Méthode de la sécante

Entrée : f : fonction ;
$a < b$: intervalle initial ;
ε : marge d'erreur
Sortie : x : zéro de f

répéter
  $c \leftarrow a - \frac{(b-a)f(a)}{f(b)-f(a)}$;
  $a, b \leftarrow b, c$
jusqu'à ce que *(f change de signe sur $[c-\varepsilon, c+\varepsilon]$ ou) $b-a$ petit*;
renvoyer (c)

gement de signe entre $c-\varepsilon$ et $c+\varepsilon$ (en considérant le signe en ces 2 valeurs), mais ce n'est pas satisfaisant (par exemple si la fonction est positive, l'algorithme ne s'arrête pas).

- On montrera plus loin que si la suite c_n converge, alors elle converge très vite : si $|c_n - c_{n-1}| < \varepsilon$, alors la somme des erreurs accumulées aux étapes suivantes reste petite, de l'ordre de ε. C'est pourquoi on préfère souvent donner une condition d'arrêt sous la forme $|b-a|$ suffisamment petit (on quantifiera mieux le « suffisamment » plus loin).

Avertissement 8.2.3 (La méthode n'est pas toujours bien définie)

- Comme on n'impose plus de changement de signe de f sur $[c_n, c_{n+1}]$, la valeur de c_{n+2} peut sortir de cet intervalle, et même de l'intervalle $[a,b]$ initial : la corde peut être presque parallèle à l'axe des abscisses et c_{n+2} est dans ce cas éloigné de c_n et c_{n+1}. Dans ce cas, c_{n+2} peut sortir du domaine de définition.
- Il n'est pas exclu d'ailleurs que la corde soit parallèle à l'axe des abscisses. Dans ce cas, c_{n+2} n'est pas défini du tout.
- La suite (c_n), mais ne pas converger.

Pour cette raison, il est nécessaire d'améliorer un peu le test d'arrêt, en limitant le nombre d'itérations. Par exemple, si on n'a pas atteint la condition d'arrêt assurant la convergence au bout de 100 itérations, on peut estimer que la méthode ne converge pas.
Que faire pour les problèmes de domaine de définition ?

On suppose par la suite que $f(c_n)$ n'est jamais nul (faute de quoi on peut arrêter l'algorithme !)

Lemme 8.2.4 (Condition suffisante locale de convergence de (c_n))

1. *Supposons f de classe $\mathcal{C}^2$ sur $[a,b]$, telle que $f(a)f(b) < 0$. Supposons que f' ne s'annule pas sur $[a,b]$. Alors f admet un unique zéro x dans $]a,b[$.*
2. *Soit alors $m_1 = \min_{[a,b]}(|f'|) > 0$, et $M_2 = \max_{[a,b]}(|f''|)$, et $\delta > 0$ suffisamment petit pour que :*

$$2\delta \times \frac{M_2}{m_1} < \frac{1}{2} \qquad \text{et} \qquad B(x,\delta) \subset [a,b].$$

Si pour une valeur $n_0 \in \mathbb{N}^*$, deux valeurs consécutives c_{n_0} et c_{n_0+1} sont dans $B(x,\delta)$, alors la suite $(c_n)_{n\in\mathbb{N}}$ est bien définie, et pour tout $n \geqslant n_0 - 1$, il existe deux réels d_n et d'_n l'un dans $[x, c_{n+1}]$, l'autre dans $[c_n, c_{n+1}]$, tels que

$$|c_{n+2} - x| \leqslant \frac{M_2}{m_1} \cdot |c_{n+1} - x| \cdot |d'_n - d_n|.$$

La suite $(|c_n - x|)_{n\in\mathbb{N}}$ est alors décroissante à partir du rang n_0 et converge vers 0.

◁ **Éléments de preuve.**

1. Le TVI et la stricte monotonie de f nous assure que f admet un unique zéro dans $[a, b]$.
2. • Commencer par montrer que si c_n et c_{n+1} sont dans $B(x, \delta)$, l'inégalité est correcte. Pour cela, exprimer $c_{n+2} - x$ à l'aide de la relation de récurrence, et forcer la factorisation par $c_{n+1} - x$ (qu'on peut supposer non nul). Remarquer que $f(c_{n+1}) = f(c_{n+1}) - f(x)$, ce qui permet d'utiliser le TAF pour définir d_n. Utiliser aussi le TAF pour obtenir d'_n, puis l'IAF sur f' cette fois.
   • Grâce à cette inégalité, montrer par récurrence double que pour tout $n \geqslant n_0$, $c_n \in B(0, \delta)$ (et donc que l'inégalité est vraie à partir du rang n_0. On pourra majorer $|d'_n - d_n|$ par 2δ (ces deux réels sont dans $B(0, \delta)$).
   • $(|c_n - x|)$ est donc ultimement sous-géométrique de raison $\frac{1}{2}$, et décroissante par définition de δ.

▷

Corollaire 8.2.5

II Sous les mêmes hypothèses que le lemme 8.2.4, pour tout $n \geqslant n_0$,

$$|c_{n+2} - x| \leqslant \frac{1}{b-a}|c_{n+1} - x| \cdot |c_n - x|.$$

◁ **Éléments de preuve.**

Par décroissance de $|c_n - x|$, d_n et d'_n sont dans $B(x, |c_n - x|)$. L'inégalité découle alors du lemme 8.2.4 ▷

Pour exploiter cette inégalité, on utilise les lemmes suivants :

Lemme 8.2.6

Soit $(y_n)_{n\in\mathbb{N}}$ une suite positive telle que pour tout $n \in \mathbb{N}$, $y_{n+2} \leqslant y_{n+1}y_n$. Alors

$$\forall n \in \mathbb{N}, \quad y_n \leqslant \alpha^{\varphi^n},$$

où $\varphi = \frac{1+\sqrt{5}}{2}$ est le nombre d'or, solution positive de l'équation $\varphi^2 = \varphi + 1$, et $\alpha = \max(y_0, y_1^{1/\varphi})$.

◁ **Éléments de preuve.**

Récurrence d'ordre 2 immédiate. ▷

Lemme 8.2.7

Soit (y_n) une suite positive telle qu'il existe $\alpha \in]0, 1[$, $M > 0$ et $\rho > 0$ tels que pour tout $n \in \mathbb{N}$, $y_n \leqslant Ma^{\rho^n}$. Alors il existe $N > 0$ et $\beta \in]0, 1[$ tels que pour tout $n \geqslant N$, $y_n \leqslant \beta^{\rho^n}$.

◁ **Éléments de preuve.**

Prendre $b \in]a, 1[$ et comparer l'ordre de grandeur asymptotique des 2 majorants ▷

Lemme 8.2.8

Soit (y_n) une suite strictement positive telle qu'il existe $K \in \mathbb{R}$ et n_0 tel que pour tout $n \geqslant n_0$, $y_{n+2} \leqslant Ky_{n+1}y_n$. Alors il existe $N > 0$ et $\beta \in]0, 1[$ tels que pour tout $n \geqslant N$, $y_n \leqslant \beta^{\rho^n}$.

◁ **Éléments de preuve.**

- Se ramener au cas où $n_0 = 0$ en grossissant éventuellement K, en considérant le maximum des $\frac{y_{n+2}}{y_{n+1}y_n}$, pour $n < n_0$.
- Utiliser alors le premier lemme avec la suite (Ky_n), puis le deuxième lemme pour se débarrasser de la constante.

▷

Théorème 8.2.9 (Convergence locale de la méthode de la sécante)

Sous les mêmes hypothèses que dans le lemme 8.2.4, il existe $\beta \in]0,1[$, et $N \in \mathbb{N}$, tel que pour tout $n \geqslant N$:

$$|c_n - x| \leqslant \beta^{\varphi^n}.$$

◁ **Éléments de preuve.**

Conséquence immédiate du dernier lemme ci-dessus appliqué à l'inégalité du corollaire . ▷

On dit que la méthode est *d'ordre φ*.

Remarque 8.2.10

Plus généralement, une méthode est d'ordre ω si, en notant (x_n) l'approximation de rang n et x la valeur recherchée, il existe $\alpha \in]0,1[$ tel que pour tout n à partir d'un certain rang, on ait :

$$|c_n - x| \leqslant \alpha^{\omega^n}.$$

Dire qu'une méthode est d'ordre ω signifie grossièrement que le nombre de décimales correctes est multiplié par ω à chaque étape.

En effet,

$$-\log_{10}(\alpha^{\omega^n}) = -\omega^n \log_{10}(\omega).$$

Or, cette expression donne à peu près le rang du premier chiffre non nul après la virgule de α^{ω^n}, donc une minoration du rang du premier chiffre de $|c_n - x|$, donc une minoration du nombre des premières décimales communes à c_n et x.
Ainsi, le nombre de décimales correctes est multiplié par environ 1.618 par la méthode de la sécante. En supposant initialement $b - a = 1$, et vérifiant les hypothèses du théorème, en 5 itérations, on a déjà 11 décimales correctes, et en 10 itérations, on a plus de 100 décimales ! La convergence est donc très rapide, beaucoup plus que la dichotomie.

Remarque 8.2.11

Si $f'(x) \neq 0$, le résultat précédent assure la convergence de la méthode de la sécante dès lors que l'initialisation se fait suffisamment proche du zéro x. En effet, on peut contrôler localement f' et f et restreindre l'intervalle initial suffisamment pour récupérer les hypothèses idoines.

Remarque 8.2.12 (Comparatif de la méthode de la sécante)

- Points forts :
  - Convergence très rapide (méthode d'ordre φ)
  - Facilité d'expression (ne nécessite pas la dérivée contrairement à la méthode de Newton)
- Points faibles :
  - Instabilité : on n'est pas assuré de la convergence ; il faut se placer suffisamment près de la racine pour avoir la convergence. la distance initiale dépend d'un minorant de $|f'|$ et d'un

majorant de $|f''|$. Il nous faut même deux valeurs consécutives près de la racine (donc une bonne appromination initiale)

∗ Condition d'arrêt mal assurée.

Proposition 8.2.13 (Retour sur la condition d'arrêt)

Sous les hypothèses du lemme 8.2.4, $(|c_n - x|)$ est sous-géométrique de raison $\frac{1}{2}$ à partir d'un certain rang N. Si pour $n \geqslant N$, on a $|c_n - c_{n+1}| \leqslant \varepsilon$, alors $|c_{n+1} - x| \leqslant \varepsilon$. Une condition d'arrêt acceptable est donc $|b - a| \leqslant \varepsilon$.

◁ **Éléments de preuve.**

Par IT, $2|c_{n+1} - x| \leqslant |c_n - x| \leqslant |c_{n+1} - c_n| + |c_{n+1} - x|$. ▷

III Méthode de Newton

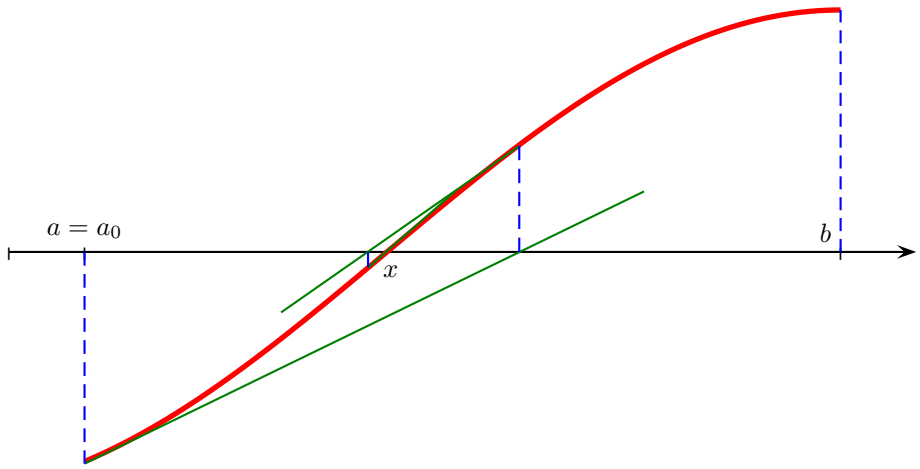

FIGURE 8.3 – Méthode de Newton

La méthode de Newton est l'aboutissement. Lorsque la méthode de la sécante converge, les valeurs des c_i sont de plus en plus proches, et la corde est alors une bonne approximation de la tangente. La méthode de Newton consiste à remplacer dans l'algorithme de la sécante la corde par la tangente. On n'a alors plus besoin des deux bornes de l'intervalle (la seconde ne servait qu'à calculer la corde).

Méthode 8.3.1 (Méthode de Newton)

On suppose f dérivable.

- On part d'une valeur initiale x_0.
- On construit x_1 comme l'intersection de la tangente en x_0 et de l'axe des abscisses.
- On itère cette construction.

Proposition 8.3.2 (Récurrence associée à la méthode de Newton)

On a alors, pour tout $n \in \mathbb{N}$, si la suite (x_n) est définie :

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

◁ **Éléments de preuve.**

Exprimer l'équation de la tangente en x_n et résoudre une petite équation linéaire... ▷

On obtient donc l'algorithme 8.3.

Algorithme 8.3 : Méthode de Newton

Entrée : f : fonction ; f' : dérivée de f ;
a : valeur initiale ;
ε : marge d'erreur
Sortie : x : zéro de f

répéter
  $a' \leftarrow a$;
  $a \leftarrow a - \frac{f(a)}{f'(a)}$
jusqu'à ce que *(f change de signe sur $[a-\varepsilon, a+\varepsilon]$) ou* $|a'-a| \leqslant \varepsilon$;
renvoyer (c)

Exemple 8.3.3 (Méthode de Heron)

La méthode de Heron pour le calcul de $\sqrt{a}$ n'est rien d'autre que la méthode de Newton appliquée à la fonction $x \mapsto x^2 - a$. La relation de récurrence est alors :

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Avertissement 8.3.4

Comme pour la méthode de la sécante, la méthode de Newton peut être mal définie (impossible à itérer).

Pour assurer la convergence, il faut donc se placer, comme pour la méthode de la sécante, suffisamment près du zéro recherché.

Théorème 8.3.5 (Convergence locale de la méthode de Newton)

Soit f de classe $\mathcal{C}^2$ s'annulant en x. Supposons donné un réel δ tel que f soit définie sur $B(x,\delta)$, et vérifie $|f'| \geqslant m_1$ et $|f''| \leqslant M_2$ sur $B(x,\delta)$, tel que $2\delta\frac{M_2}{m_1} < 1$ (ce qu'on peut obtenir en diminuant δ). Alors, si $x_0 \in B(x,\delta)$,

(i) pour tout $n \in \mathbb{N}$, $|x_{n+1} - x| \leqslant |x_n - x|^2 \frac{M_2}{m_1}$

(ii) il existe $a \in]0,1[$ tel que pour tout n assez grand, $|x_n - x| \leqslant a^{2^n}$.

En particulier, la méthode de Newton est d'ordre 2 : on double le nombre de décimales correctes à chaque itération.

◁ **Éléments de preuve.**

(i) Même démarche que pour la méthode de la sécante, en plus simple car la relation est d'ordre 1 : considérer $x_{n+1} - x$, mettre $(x_n - x)$ en facteur, utiliser le TAF pour f entre x_n et x, puis l'IAF pour f'.

(ii) Poser K convenable tel que $y_n = K|x_n - x|$ vérifie $y_{n+1} \leqslant y_n^2$. Pour quelle valeur de b a-t-on alors $y_n \leqslant b^{2^n}$? Comment se débarrasser de K ?

▷

La méthode de Newton assure donc une convergence très rapide, meilleure que la méthode de la sécante.

Remarque 8.3.6

- Le résultat reste vrai à partir du rang n_0 si on a seulement $x_0 \in B(x, \delta)$: le point fixe x est attractif, il suffit de passer suffisamment près pour avoir convergence.
- Comme pour la méthode de la sécante, la convergence est très rapide (encore plus rapide), et sous-géométrique de raison $\frac{1}{2}$. On peut donc utiliser la même condition d'arrêt.

Nous donnons un critère simple, fréquemment vérifié (au moins après restriction), permettant d'assurer la convergence globale de la méthode de Newton.

Théorème 8.3.7 (Cas simple de convergence globale de la méthode de Newton)

Soit f une fonction de classe $\mathcal{C}^2$ sur un intervalle $[a, b]$, strictement convexe et croissante, telle que $f(a) < 0 < f(b)$. Alors la méthode de Newton initialisée par le point b est convergente, et les valeurs approchées (x_n) successives calculées forment une suite décroissante à partir du rang 1.
Cet énoncé s'adapte aux cas de décroissance ou concavité (s'aider d'un dessin pour savoir sur quel bord initialiser)

◁ **Éléments de preuve.**

Par stricte convexité, si $x_n > x$, alors $x_{n+1} > x$ (la courbe est au-dessus de sa tangente). Ainsi, cette inégalité est vérifiée pour tout n, et $f(x_n) > 0$. La positivité de la pente de la tangente en x_n assure alors la décroissance. On passe alors à la limite dans la relation de récurrence, en utilisant la continuité de f et f'. ▷

Remarque 8.3.8 (Comparatif de la méthode de Newton)

- Points forts :
  - Convergence extrêmement rapide
- Points faibles :
  - Mêmes problèmes d'instabilité que la méthode de la sécante : il faut initialiser près du zéro pour être assuré de la convergence. Mais une seule valeur suffit.
  - Utilisation de la dérivée de f, nécessitant d'être fournie par l'utilisateur, ou calculée numériquement (mais cela n'est satisfaisant que pour une fonction n'ayant pas trop de microvariations).

IV Le problème de la dérivation numérique (HP)

La méthode de Newton nécessitant l'utilisation d'une dérivation numérique, on étudie maintenant une méthode de dérivation numérique, en essayant d'optimiser l'erreur.

Méthode 8.4.1 (Dérivation numérique)

On approche $f'(x)$ par $\frac{f(x+h)-f(x-h)}{2h}$, où h est suffisamment petit.

Proposition 8.4.2

L'erreur d'approximation théorique est de l'ordre de h^2, et l'erreur d'arrondi de l'ordre de $\frac{|f(x)|}{|h|}r$, où r est l'erreur élémentaire.

◁ **Éléments de preuve.**

En effet, on peut montrer à l'aide de la formule de Taylor-Young, appliquée entre x et $x+h$, puis entre x et $x-h$, que $\frac{f(x+h)-f(x-h)}{2h}$ diffère de $f'(x)$ d'un terme de l'ordre de grandeur de αh^2. De plus, les grandeurs qui interviennent dans ces calculs sont de l'ordre de grandeur de $\frac{f(x)}{h}$, donc de $\frac{1}{h}$. L'erreur d'arrondi sur ces termes sera donc de l'ordre de $\frac{r}{h}$. ▷

Corollaire 8.4.3 (Valeur optimale de h)

Sous l'hypothèse que f et des premières dérivées sont de l'ordre de grandeur de 1 *(ni trop petit, ni trop gros) la valeur optimale de h minimisant l'erreur entre $f'(x)$ et la valeur calculée $\frac{f(x+h)-f(x-h)}{2h}$ est de l'ordre de $\sqrt[3]{r}$, où r est l'erreur élémentaire.*

◁ **Éléments de preuve.**

En sommant l'erreur d'arrondi et l'erreur théorique, on a une erreur de l'ordre de $\alpha h^2 + \frac{r}{h}$. En minimisant cette fonction en h (par une étude de fonction par exemple), on se rend compte qu'elle admet un minimum en une quantité $\beta\sqrt[3]{r}$. ▷

Dans cette estimation, on a négligé les erreurs d'arrondi sur le calcul de $x+h$ et $x-h$. Ces erreurs sont à peu près réduites à 0 si h est une puissance de 2 (cela revient juste à changer une successions de bits $01\ldots1$ en $10\ldots0$, ce qui se fait de façon exacte, sauf si cela rajoute un chiffre dans l'écriture binaire, ce qui fait sauter le chiffre de poids minimal. L'erreur est donc très petite, négligeable devant les autres termes. Ainsi, on a tout intérêt à considérer pour h une puissance de 2. En norme IEEE 754, la mantisse étant constituée de 52 bits, l'erreur élémentaire est de l'ordre de 2^{-52}, donc la valeur de h optimale est de l'ordre de 2^{-17} (ce qui correspond à 10^{-5} environ, mais comme on l'a expliqué, il est préférable de rester en puissances de 2).