2.3.2 Variantes de la méthode de Newton

L'avantage majeur de la méthode de Newton par rapport à une méthode de point fixe par exemple est sa vitesse de convergence d'ordre 2. On peut d'ailleurs remarquer que lorsque la méthode ne converge pas, par exemple si l'itéré initial $\boldsymbol{x}^{(0)}$ n'a pas été choisi "suffisamment proche" de $\overline{\boldsymbol{x}}$, alors la méthode diverge très vite...

L'inconvénient majeur de la méthode de Newton est son coût : on doit d'une part calculer la matrice jacobienne $Dg(\boldsymbol{x}^{(k)})$ à chaque itération, et d'autre part la factoriser pour résoudre le système linéaire $Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)})$. (On rappelle que pour résoudre un système linéaire, il ne faut pas calculer l'inverse de la matrice, mais plutôt la factoriser sous la forme LU par exemple, et on calcule ensuite les solutions des systèmes avec matrice triangulaires faciles à inverser, voir Chapitre 1.) Plusieurs variantes ont été proposées pour tenter de réduire ce coût.

"Faux quasi Newton"

Soient $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}$ tels que $g(\bar{\boldsymbol{x}}) = 0$. On cherche à calculer $\bar{\boldsymbol{x}}$. Si on le fait par la méthode de Newton, l'algorithme s'écrit :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n, \\ Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)}), \ n \geq 0. \end{array} \right.$$

La méthode du "Faux quasi-Newton" (parfois appelée quasi-Newton) consiste à remplacer le calcul de la matrice jacobienne $Dg(\boldsymbol{x}^{(k)})$ à chaque itération par un calcul toutes les "quelques" itérations. On se donne une suite $(n_i)_{i\in\mathbb{N}}$, avec $n_0 = 0$ et $n_{i+1} > n_i \ \forall i \in \mathbb{N}$, et on calcule la suite $(\boldsymbol{x}^{(k)})_{n\in\mathbb{N}}$ de la manière suivante :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ Dg(\boldsymbol{x}^{(n_i)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)}) \text{ si } n_i \leq k < n_{i+1}. \end{array} \right. \tag{2.26}$$

Avec cette méthode, on a moins de calculs et de factorisations de la matrice jacobienne $Dg(\boldsymbol{x})$ à effectuer, mais on perd malheureusement la convergence quadratique : cette méthode n'est donc pas très utilisée en pratique.

Newton incomplet

On suppose que g s'écrit sous la forme :

$$g(\boldsymbol{x}) = A\boldsymbol{x} + F_1(\boldsymbol{x}) + F_2(\boldsymbol{x}), \text{ avec } A \in \mathcal{M}_n(\mathbb{R}) \text{ avec } F_1, F_2 \in C^1(\mathbb{R}^n, \mathbb{R}^n).$$

L'algorithme de Newton (2.20) s'écrit alors :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ \big(A + DF_1(\boldsymbol{x}^{(k)}) + DF_2(\boldsymbol{x}^{(k)})\big)\,(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = \\ \qquad\qquad -A\boldsymbol{x}^{(k)} - F_1(\boldsymbol{x}^{(k)}) - F_2(\boldsymbol{x}^{(k)}). \end{array} \right.$$

La méthode de Newton incomplet consiste à ne pas tenir compte de la jacobienne de F_2.

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ (A + DF_1(\boldsymbol{x}^{(k)}))(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -A\boldsymbol{x}^{(k)} - F_1(\boldsymbol{x}^{(k)}) - F_2(\boldsymbol{x}^{(k)}). \end{array} \right. \tag{2.27}$$

On dit qu'on fait du "Newton sur F_1" et du "point fixe sur F_2". Les avantages de cette procédure sont les suivants :

- La méthode ne nécessite pas le calcul de $DF_2(\boldsymbol{x})$, donc on peut l'employer si $F_2 \in C(\mathbb{R}^n, \mathbb{R}^n))$ n'est pas dérivable.
- On peut choisir F_1 et F_2 de manière à ce que la structure de la matrice $A + DF_1(\boldsymbol{x}^{(k)})$ soit "meilleure" que celle de la matrice $A + DF_1(\boldsymbol{x}^{(k)}) + DF_2(\boldsymbol{x}^{(k)})$; si par exemple A est la matrice issue de la discrétisation du Laplacien, c'est une matrice creuse. On peut vouloir conserver cette structure et choisir F_1 et F_2 de manière à ce que la matrice $A + DF_1(\boldsymbol{x}^{(k)})$ ait la même structure que A.

— Dans certains problèmes, on connaît a priori les couplages plus ou moins forts dans les non-linéarités : un couplage est dit fort si la variation d'une variable entraîne une variation forte du terme qui en dépend. Donnons un exemple : Soit f de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par $f(x,y) = (x + \sin(10^{-5}y), \exp(x) + y)$, et considérons le système non linéaire $f(x,y) = (a,b)$ où $(a,b) \in \mathbb{R}^2$ est donné. Il est naturel de penser que pour ce système, le terme de couplage de la première équation en la variable y sera faible, alors que le couplage de deuxième équation en la variable x sera fort.
On a alors intérêt à mettre en oeuvre la méthode de Newton sur la partie "couplage fort" et une méthode de point fixe sur la partie "couplage faible".

L'inconvénient majeur est la perte de la convergence quadratique. La méthode de Newton incomplet est cependant assez souvent employée en pratique en raison des avantages énumérés ci-dessus.

Remarque 2.22. *Si $F_2 = 0$, alors la méthode de Newton incomplet est exactement la méthode de Newton. Si $F_1 = 0$, la méthode de Newton incomplet s'écrit*

$$A(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -A\boldsymbol{x}^{(k)} - F_2(\boldsymbol{x}^{(k)}),$$

En supposant A inversible, on a alors $\boldsymbol{x}^{(k+1)} = -A^{-1}F_2(\boldsymbol{x}^{(k)})$. C'est donc dans ce cas la méthode du point fixe sur la fonction $-A^{-1}F_2$.

Méthode de la sécante

La méthode de la sécante est une variante de la méthode de Newton dans le cas de la dimension 1 d'espace. On suppose ici $n = 1$ et $g \in C^1(\mathbb{R}, \mathbb{R})$. La méthode de Newton pour calculer $\overline{\boldsymbol{x}} \in \mathbb{R}$ tel que $g(\overline{\boldsymbol{x}}) = 0$ s'écrit :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R} \\ g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}), \quad \forall n \geq 0. \end{array} \right.$$

On aimerait simplifier le calcul de $g'(x^{(k)})$, c'est–à–dire remplacer $g'(x^{(k)})$ par une quantité "proche" sans calculer g'. Pour cela, on remplace la dérivée par un quotient différentiel. On obtient la méthode de la sécante :

$$\left\{ \begin{array}{l} x^{(0)}, x^{(1)} \in \mathbb{R} \\ \dfrac{g(x^{(k)}) - g(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}) \quad n \geq 1. \end{array} \right. \tag{2.28}$$

Remarquons que dans la méthode de la sécante, $x^{(k+1)}$ dépend de $x^{(k)}$ et de $x^{(k-1)}$: on a une méthode à deux pas ; on a d'ailleurs besoin de deux itérés initiaux $x^{(0)}$ et $x^{(1)}$. L'avantage de cette méthode est qu'elle ne nécessite pas le calcul de g'. L'inconvénient est qu'on perd la convergence quadratique. On peut toutefois montrer (voir exercice 104 page 182) que si $g(\bar{x}) = 0$ et $g'(\bar{x}) \neq 0$, il existe $\alpha > 0$ tel que si $x^{(0)}, x^{(1)} \in [\bar{x} - \alpha.\bar{x} + \alpha] = I_\alpha$, $x^{(0)} \neq x^{(1)}$, la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par la méthode de la sécante (2.28) est bien définie, que $(x^{(k)})_{n\in\mathbb{N}} \subset I_\alpha$ et que $x^{(k)} \to \bar{x}$ quand $n \to +\infty$. De plus, la convergence est super linéaire, *i.e.* si $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$, alors $\dfrac{x^{(k+1)} - \bar{x}}{x^{(k)} - \bar{x}} \to 0$ quand $n \to +\infty$. On peut même montrer (voir exercice 104 page 182) que la méthode de la sécante est convergente d'ordre d, où d est le nombre d'or.

Méthodes de type "Quasi Newton"

On veut généraliser la méthode de la sécante au cas $n > 1$. Soient donc $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Pour éviter de calculer $Dg(x^{(k)})$ dans la méthode de Newton (2.20), on va remplacer $Dg(x^{(k)})$ par $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ "proche de $Dg(x^{(k)})$". En s'inspirant de la méthode de la sécante en dimension 1, on cherche une matrice $B^{(k)}$ qui, $x^{(k)}$ et $x^{(k-1)}$ étant connus (et différents), vérifie la condition :

$$B^{(k)}(x^{(k)} - x^{(k-1)}) = g(x^{(k)}) - g(x^{(k-1)}) \tag{2.29}$$

Dans le cas où $n = 1$, cette condition détermine entièrement $B^{(k)}$;car on peut écrire : $B^{(k)} = \dfrac{g(x^{(k)}) - g(x^{(k-1)}}{x^{(k)} - x^{(k-1)}}$.
Si $n > 1$, la condition (2.29) ne permet pas de déterminer complètement $B^{(k)}$. Il y a plusieurs façons possibles

de choisir $B^{(k)}$, nous en verrons en particulier dans le cadre des méthodes d'optimisation (voir chapitre 4, dans ce cas la fonction g est un gradient), nous donnons ici la méthode de Broyden [6]. Celle-ci consiste à choisir $B^{(k)}$ de la manière suivante : à $x^{(k)}$ et $x^{(k-1)}$ connus, on pose $\delta^{(k)} = x^{(k)} - x^{(k-1)}$ et $y^{(k)} = g(x^{(k)}) - g(x^{(k-1)})$; on suppose $B^{(k-1)} \in \mathcal{M}_n(\mathbb{R})$ connue (et $\delta^{(k)} \neq 0$), et on cherche $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ telle que

$$B^{(k)}\delta^{(k)} = y^{(k)} \tag{2.30}$$

(c'est la condition (2.29), qui ne suffit pas à déterminer $B^{(k)}$ de manière unique) et qui vérifie également :

$$B^{(k)}\xi = B^{(k-1)}\xi, \quad \forall \xi \in \mathbb{R}^n \text{ tel que } \xi \perp \delta^{(k)}. \tag{2.31}$$

Proposition 2.23 (Existence et unicité de la matrice de Broyden)**.**
Soient $y^{(k)} \in \mathbb{R}^n$, $\delta^{(k)} \in \mathbb{R}^n$, $\delta^{(k)} \neq 0$, et $B^{(k-1)} \in \mathcal{M}_n(\mathbb{R})$. Il existe une unique matrice $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ vérifiant (2.30) *et* (2.31)*; la matrice $B^{(k)}$ s'exprime en fonction de $y^{(k)}$, $\delta^{(k)}$ et $B^{(k-1)}$ de la manière suivante :*

$$B^{(k)} = B^{(k-1)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t. \tag{2.32}$$

DÉMONSTRATION – L'espace des vecteurs orthogonaux à $\delta^{(k)}$ est de dimension $n-1$. Soit $(\gamma_1, \ldots, \gamma_{n-1})$ une base de cet espace, alors $(\gamma_1, \ldots, \gamma_{n-1}, \delta^{(k)})$ est une base de $\mathbb{R}^n$ et si $B^{(k)}$ vérifie (2.30) et (2.31), les valeurs prises par l'application linéaire associée à $B^{(k)}$ sur chaque vecteur de base sont connues, ce qui détermine l'application linéaire et donc la matrice $B^{(k)}$ de manière unique. Soit $B^{(k)}$ définie par (2.32), on a :

$$B^{(k)}\delta^{(k)} = B^{(k-1)}\delta^{(k)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t\delta^{(k)} = y^{(k)},$$

et donc $B^{(k)}$ vérifie (2.30). Soit $\xi \in \mathbb{R}^n$ tel que $\xi \perp \delta^{(k)}$, alors $\xi \cdot \delta^{(k)} = (\delta^{(k)})^t\xi = 0$ et donc

$$B^{(k)}\xi = B^{(k-1)}\xi + \frac{(y^{(k)} - B^{(k-1)}\delta^{(k)})}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t\xi = B^{(k-1)}\xi, \ \forall \xi \perp \delta^{(k)}.$$

■

L'algorithme de Broyden s'écrit donc :

$$\begin{cases} \text{Initialisation : } x^{(0)},\ x^{(1)} \in \mathbb{R}^n,\ x^{(0)}, \neq x^{(1)},\ B_0 \in \mathcal{M}_n(\mathbb{R}) \\ \text{Itération k : } x^{(k)},\ x^{(k-1)} \text{ et } B^{(k-1)} \text{ connus, on pose} \\ \qquad \delta^{(k)} = x^{(k)} - x^{(k-1)} \text{ et } y^{(k)} = g(x^{(k)}) - g(x^{(k-1)}); \\ \text{Calcul de } B^{(k)} = B^{(k-1)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t, \\ \text{résolution de } B^{(k)}(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}). \end{cases}$$

Une fois de plus, l'avantage de cette méthode est de ne pas nécessiter le calcul de $Dg(x)$, mais l'inconvénient est la perte du caractère quadratique de la convergence .

6. C. G. Broyden, "A Class of Methods for Solving Nonlinear Simultaneous Equations." *Math. Comput.* 19, 577-593, 1965.