ECN 6338 Cours 4

Résolution de systèmes d'équations non-linéaires

William McCausland

2022-02-09

Les problèmes univarié et multivarié

- Problème univarié : trouvez $x \in \mathbb{R}$ qui vérifie
$$f(x) = 0,$$
où $f : \mathbb{R} \to \mathbb{R}$.
- Problème multivarié : trouvez $x \in \mathbb{R}^n$ qui vérifie
$$f(x) = 0_n,$$
où $f : \mathbb{R}^n \to \mathbb{R}^n$.
- Problème multivarié, élément par élément : trouvez $(x_1, \ldots, x_n)$ qui vérifie
$$\begin{aligned} f^1(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f^n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

La résolution de systèmes d'équations et l'optimisation

La solution x^* au problème d'optimisation

$$\max_{x \in \mathbb{R}} f(x),$$

où $f : \mathbb{R} \to \mathbb{R}$ et $f \in C^2$, est aussi la solution du système

$$\frac{\partial f(x)}{\partial x^\top} = 0.$$

Cependant, la résolution du système $g(x) = 0$, $g \in C^1$, est plus générale :

- La matrice jacobienne de g n'est pas forcément symmétrique
- La matrice jabobienne de ∇f est la matrice hessienne symmétrique de g.

Systèmes non-linéaires et le nombre de solutions

Dans le cas spécial $f(x) = Ax - b = 0$, où A est une matrice $n \times n$,

- si le rang de A est de n, il y a une solution unique;
- si le rang de A est moins grand, il n'y a aucune solution :

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

  ou il y a un nombre infini de solutions :

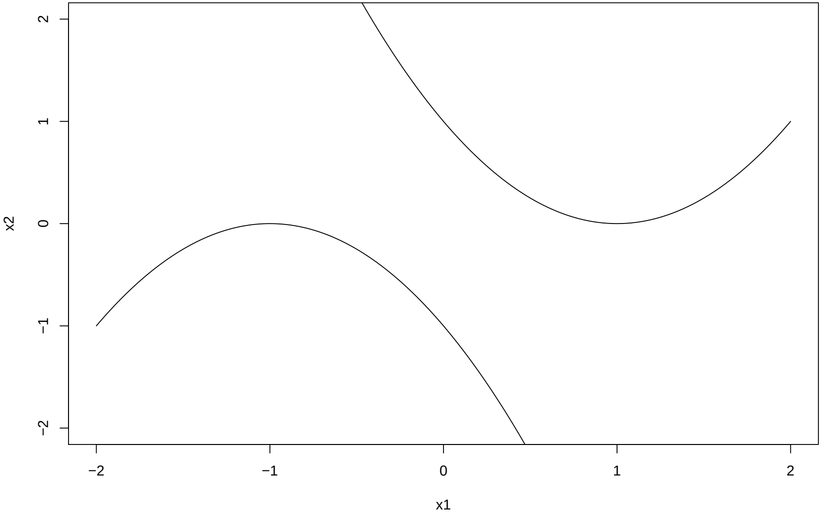
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Dans le cas général, il peut y avoir

- aucune solution, même pour les fonctions f^i très différentes,
- un nombre fini arbitraire de solutions,
- un nombre infini de solutions.

Exemple : absence d'une solution

$$f^1(x_1, x_2) = x_2 - (x_1 + 1)^2, \quad f^2(x_1, x_2) = x_2 - (x_1 - 1)^2.$$



Exemple : solutions multiples

$$f^1(x_1, x_2) = x_1^2 + x_2^2 - 1, \quad f^2(x_1, x_2) = 2x_1^2 - x_2 - 1.$$

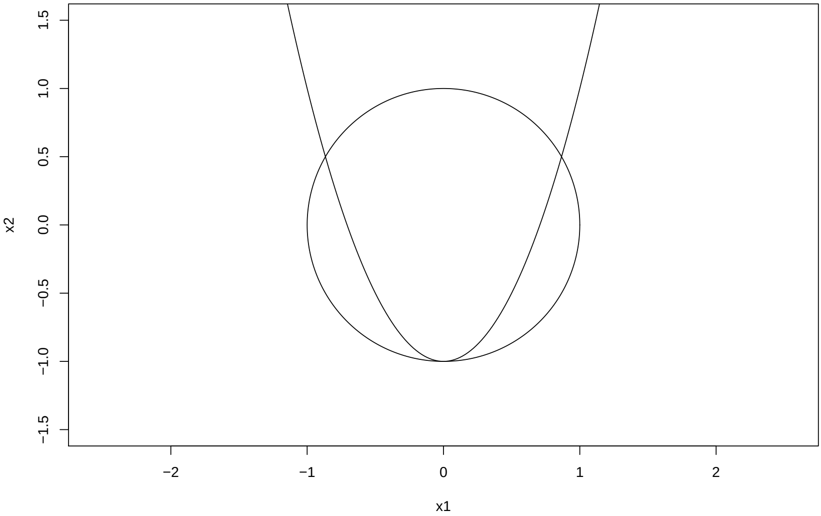


Illustration univariée I : méthode de Newton-Raphson

Considérons la fonction f et sa dérivée, définie sur l'intervalle $[0, 1]$:

$$f(x) = (1 - x)^3 - \log(1 + x), \quad f'(x) = -3(1 - x)^2 - (1 + x)^{-1}.$$

Si on prend le point initial $x_0 = 0$, la droite de tangente est

$$g(x) = f(0) + f'(0)(x - 0) = 1 - 4x.$$

et le point x_1 de l'itération Newton est l'intersection de cette droite et l'axe des abscisses :

$$x_1 = x_0 - \frac{f'(0)}{f(0)} = 0 - \frac{1}{-3 - 1} = \frac{1}{4}.$$

Un pas de Newton de plus donne

$$x_2 = x_1 - \frac{f'(1/4)}{f(1/4)} \approx 0.329892,$$

très près de la racine unique.

Illustration univariée II : échec de la méthode de Newton

- On peut commencer à $x_0 = 1$ où la courbature est plus prononcée.
- On évalue $f(x_0) = -\log 2$, $f'(x_0) = -\frac{1}{2}$ et on calcule

$$x_1 = x_0 - \frac{f'(x_0)}{f(x_0)} \approx -0.3862944,$$

  beaucoup plus loin de la racine et hors de l'intervalle $[0, 1]$.
- À voir aussi : "Pathological Examples", page 153 de Judd.

Illustration univariée III : méthode d'intérpolation linéaire

- Note : fonction $f(x)$ en bleu, droites de tangente en rouge, droite de sécante en vert.
- Pour la première itération, où on calcule x_1, on n'a pas encore deux valeurs précédentes et on utilise la méthode de Newton.
- Une fois qu'on a x_0 et x_1, on peut construire la droite de sécante entre $(x_0, f(x_0))$ et $(x_1, f(x_1))$:

$$h(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

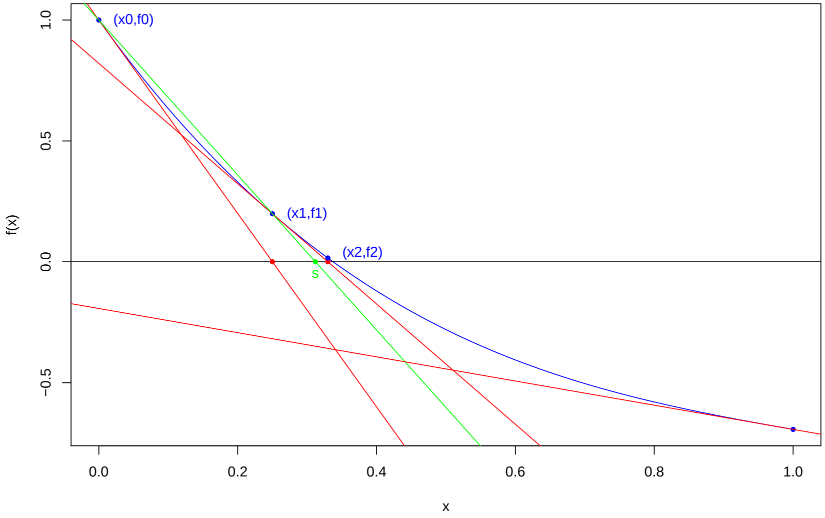
- Le point s de l'itération avec interpolation linéaire est l'intersection de cette droite et l'axe des abscisses :

$$s = x_0 + \frac{(x_1 - x_0)}{f(x_1) - f(x_0)} f(x_0) \approx 0.3120053,$$

un peu plus loin de la racine.

Illustration (Newton et interpolation linéaire)

```
source('root_uni.R')
```



Méthode de dichotomie

Intrants à l'itération $k+1$: points a_k, b_k, valeurs $f(a_k)$, $f(b_k)$ tels que

1. $a_k < b_k$,
2. $f(a_k)f(b_k) < 0$.

À l'itération $k+1$:

1. Calculer $m = \frac{1}{2}(a_k + b_k)$.
2. Évaluer $f(m)$, si $f(m) = 0$, terminer avec m.
3. Si $f(m)$ a la même signe que a_k,

$$a_{k+1} = m, \quad b_{k+1} = b_k.$$

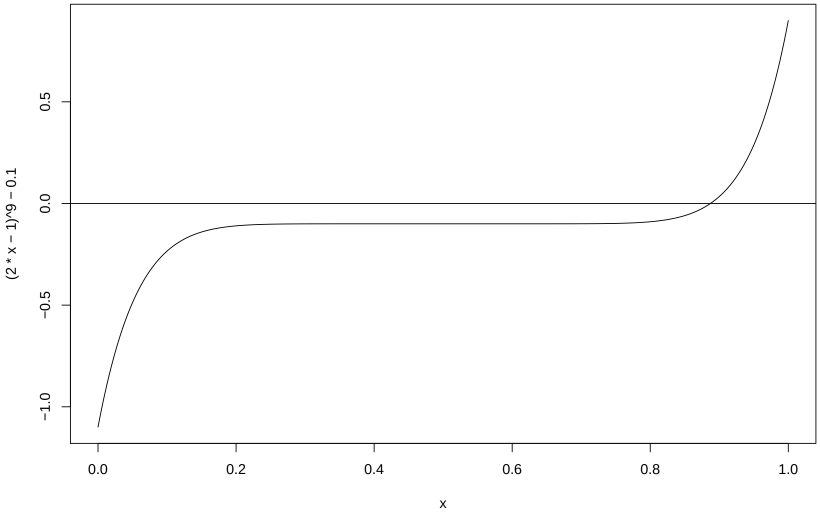
   sinon

$$a_{k+1} = a_k \quad b_{k+1} = m.$$

   Si $b_{k+1} - a_{k+1} < \delta$, terminer avec $\frac{1}{2}(a_{k+1} + b_{k+1})$.

Quand la méthode de dichotomie marche relativement bien

```
x = seq(0, 1, by=0.0001)
plot(x, (2*x-1)^9 - 0.1, type='l')
abline(h=0)
```



Discussion, méthode de dichotomie

- Pour la méthode de dichotomie,
  - on gagne 1 bit de précision à chaque itération (pas beaucoup, mais sûr),
  - on sait à l'avance combien d'itérations il faut pour atteindre ces deux conditions : $b_k - a_k < \delta$, $[a_k, b_k]$ contient une racine.
- Considérez des jeux zéro-somme entre
  - joueur 1 qui choisit une fonction continue $f(x)$ avec $f(a_0)f(b_0) < 0$, et veut maximiser le nombre d'itérations pour trouver un intervalle $[a, a+\delta]$ qui contient une racine.
  - joueur 2 qui choisit une algorithme pour trouver un intervalle $[a, a+\delta]$ contenant une racine.
- Conjecture : Si joueur 2 joue en premier, la méthode de dichotomie est optimale (minmax).
- Mais la méthode de dichotomie est très sous-optimale pour les fonctions habituelles.
- On veut accélérer la convergence et en même temps garantir un intervalle court en un nombre borné d'itérations.

Méthodes du type Dekker-Brent

Intrants à l'itération $k+1$: points a_k, b_k, b_{k-1} ($b{-}1 = a_0$) et valeurs $f(a_k)$, $f(b_k)$ et $f(b_{k-1})$ tels que

1. $|f(a_k)| \leq |f(b_k)|$ (point b_k, contrepoint a_k)
2. $f(a_k)f(b_k) < 0$.

À l'itération $k+1$:

1. Calculer $m = \frac{1}{2}(a_k + b_k)$.
2. Calculer s comme fonction de a_k, b_k, $f(a_k)$, $f(b_k)$, b_{k-1}, $f(b_{k-1})$. (détails à venir)
3. Choisir entre $b_{k+1} = s$ et $b_{k+1} = m$. (détails à venir)
4. Évaluer $f(b_{k+1})$, si $f(b_{k+1}) = 0$, terminer avec b_{k+1}.
5. Choisir entre $a_{k+1} = a_k$ et $a_{k+1} = b_k$ tel que $f(a_{k+1})f(b_{k+1}) < 0$. (Condition 2.)
6. Si $|f(a_{k+1})| < f(b_{k+1})|$, échanger a_{k+1} et b_{k+1}. (Condition 1.)
7. Si $|a_{k+1} - b_{k+1}| < \delta$, terminer avec b_{k+1}.

Calculer s (étape 2) par interpolation linéaire (droite sécante)

$$s = b_k - \frac{b_k - b_{k-1}}{f(b_k) - f(b_{k-1})} f(b_k)$$

Notes :

1. s n'est pas une fonction de a_k.
2. Si on choisit s par interpolation linéaire, une condtion nécessaire pour choisir $b_{k+1} = s$ (étape 3) est que s se trouve entre m et b_k.

Calculer s (étape 2) par interpolation inverse quadratique

- Supposez que $f(a_k)$, $f(b_k)$ et $f(b_{k-1})$ sont distinctes
- Voici une fonction quadratique $g(y)$ qui passe par les points $(f(a_k), a_k)$, $(f(b_k), b_k)$ et $(f(b_{k-1}), b_{k-1})$:

$$\begin{aligned} g(y) = & \frac{(y - f(a_k))(y - f(b_k))}{(f(b_{k-1}) - f(a_k))(f(b_{k-1} - f(b_k))} b_{k-1} \\ & + \frac{(y - f(a_k))(y - f(b_{k-1}))}{(f(b_k) - f(a_k))(f(b_k) - f(b_{k-1}))} b_k \\ & + \frac{(y - f(b_{k-1}))(y - f(b_k))}{(f(a_k) - f(b_{k-1}))(f(a_k) - f(b_k))} a_k \end{aligned}$$

- Notez que la fonction inverse $f^{-1}(y)$ passe par les mêmes points.
- Défine $s = g(0)$, un zéro de la fonction $g^{-1}(x)$

Calculer s par interpolation inverse quadratique (cont.)

$$
\begin{aligned}
s = {} & \frac{f(a_k)f(b_k)}{(f(b_{k-1}) - f(a_k))(f(b_{k-1} - f(b_k))} b_{k-1} \\
& + \frac{f(a_k)f(b_{k-1})}{(f(b_k) - f(a_k))(f(b_k) - f(b_{k-1}))} b_k \\
& + \frac{f(b_{k-1})f(b_k)}{(f(a_k) - f(b_{k-1}))(f(a_k) - f(b_k))} a_k
\end{aligned}
$$

Notes :

1. Habituellement, c'est une amélioration, mais on peut toujours utiliser l'interpolation linéaire quand $k = 1$ où quand deux valeurs sont très près l'une à l'autre.
2. Si on choisit s par interpolation inverse quadratique, une condition nécessaire pour choisir $b_{k+1} = s$ (étape 3) est que s se trouve entre $\frac{3}{4}b_k + \frac{1}{4}a_k$ et b_k.

Choisir entre s et m (étape 3)

- $b_{k+1} = m$ est plus sécure que $b_{k+1} = s$, mais le deuxième est habituellement meilleur.
- On ajoute aux conditions nécessaires déjà mentionnées pour choisir s d'autres conditions :
  - Après un pas de bisection (pour b_k), on ajoute les conditions $|b_k - b_{k-1}| > \delta$ et $\frac{1}{2}|b_k - b_{k-1}| > |s - b_k|$.
  - Après un pas d'interpolation, on ajoute les conditions $|b_{k-1} - b_{k-2}| > \delta$ et $\frac{1}{2}|b_{k-1} - b_{k-2}| > |s - b_k|$.

Méthode de Newton

- L'expansion linéaire de Taylor autour du point actuel x^k est
$$g(x) = f(x^k) + J(x^k)(x - x^k).$$
- Si la matrice jacobienne est inversible, il y a un zéro de g à
$$x^* = x^k - J(x^k)^{-1} f(x^k).$$