Petit complément sur la méthode de Newton

www.h-k.fr/publications/objectif-agregation

1. Présentation

Cette note complète la sous-section 1.3.3 de [BPM] à propos de la méthode de Newton. L'exemple 1.53 de cette sous-section justifie l'introduction de la méthode de Newton pour calculer le minimum d'une fonction f, et présente brièvement ses qualités et ses défauts. Dans ce complément, nous nous focalisons sur sa propriété majeure : nous montrons que, sous de « bonnes » hypothèses, la méthode converge vers un minimum de la fonction, et que cette convergence est rapide.

2. Résultat

Nous reprenons les notations de la sous-section 1.3.3 de [BPM]. Soit $f\colon \mathbb{R}^n \to \mathbb{R}$ une fonction deux fois différentiable sur $\mathbb{R}^n$. Son gradient en un point x est noté $\nabla f(x)$ et son hessien $\mathrm{H}f(x)$. La suite construite par la méthode de Newton est alors définie par

$$x_0 \in \mathbb{R}^n \qquad \text{et} \qquad x_{k+1} = x_k - \big(\mathrm{H}f(x_k)\big)^{-1}\nabla f(x_k),$$

sous réserve que les matrices hessiennes soient inversibles. Le théorème suivant établit qu'un « bon » minimum $x^\star$ de f attire la suite de Newton.

Théorème 1 — Convergence quadratique de Newton. Soit f une fonction de classe C^3. Supposons que $x^\star$ est un minimum local de f vérifiant les conditions suffisantes de minimalité c'est-à-dire

$$\nabla f(x^\star) = 0 \qquad \text{et} \qquad \mathrm{H}f(x^\star) \text{ définie positive.}$$

Alors pour x_0 suffisamment proche de $x^\star$, la suite de Newton est bien définie, elle converge vers $x^\star$, et cette convergence est quadratique, c'est-à-dire

$$\|x_{k+1} - x^\star\| \leqslant \mathrm{O}\big(\|x_k - x^\star\|^2\big).$$

Preuve. La démonstration repose essentiellement sur l'utilisation d'un développement de Taylor pour aboutir à l'inégalité $(*)$ ci-dessous, à partir de laquelle tout est déduit.

L'application $x \mapsto \det(\mathrm{H}f(x))$ étant continue, $\mathrm{H}f(x)$ est inversible pour tout x dans une boule ouverte U centrée en $x^\star$. Définissons alors sur U l'application

$$\mathrm{N}\colon x \longmapsto x - \big(\mathrm{H}f(x)\big)^{-1}\nabla f(x),$$

Cette application vérifie deux propriétés importante pour notre étude :

$$x_{k+1} = \mathrm{N}(x_k) \qquad \text{et} \qquad x^\star = \mathrm{N}(x^\star).$$

Par définition de N, nous avons une relation entre $\mathrm{N}(x)$ et x

$$\mathrm{H}f(x)(\mathrm{N}(x) - x) + \nabla f(x) = 0.$$

Écrivons le développement de Taylor de la fonction $x \mapsto \nabla f(x)$ de classe C^1 entre un point $x \in \mathrm{U}$ et $x^\star$

$$0 = \nabla f(x^\star) = \nabla f(x) + \mathrm{H}f(x)(x^\star - x) + \mathrm{R}(x, x^\star)$$

où $\mathrm{R}(x, x^\star)$ est le reste de Taylor associé. En soustrayant les deux égalités, on obtient

$$\mathrm{H}f(x)(\mathrm{N}(x) - x^\star) = \mathrm{R}(x, x^\star)$$

d'où
$$\|\mathrm{N}(x) - x^\star\| = \|(\mathrm{H}f(x))^{-1}\mathrm{R}(x, x^\star)\| \leqslant \|(\mathrm{H}f(x))^{-1}\|\,\|\mathrm{R}(x, x^\star)\|.$$

Les fonctions $x \mapsto (\mathrm{H}f(x))^{-1}$ et $x \mapsto \mathrm{D}^3 f(x)$ étant continues, elles sont bornées sur tout compact de U. Fixons K une boule fermée contenant $x^\star$ et incluse dans U et notons

$$\mathrm{C}_1 = \max_{x \in \mathrm{K}} \|(\mathrm{H}f(x))^{-1}\| \qquad \text{et} \qquad \mathrm{C}_2 = \max_{x \in \mathrm{K}} \|\mathrm{D}^3 f(x)\|,$$

ainsi que $\mathrm{C} = \mathrm{C}_1\mathrm{C}_2$. On obtient donc la majoration

$$\|\mathrm{N}(x) - x^\star\| \leqslant \mathrm{C}\|x - x^\star\|^2. \tag{$*$}$$

Définition. Cette inégalité garantit tout d'abord que la suite de Newton est bien définie au voisinage de $x^\star$. Pour voir ceci, on peut choisir un réel $\alpha > 0$ vérifiant $\mathrm{C}\alpha < 1$, et assez petit pour que la boule fermée $\mathrm{B}(x^\star, \alpha)$ de centre $x^\star$ et rayon α soit incluse dans K. Toutes ces précautions sont prises pour que la fonction N envoie la boule $\mathrm{B}(x^\star, \alpha)$ dans elle-même. En effet, nous avons

$$\|\mathrm{N}(x) - x^\star\| \leqslant (\mathrm{C}\|x - x^\star\|)\|x - x^\star\| \leqslant \mathrm{C}\alpha\|x - x^\star\|.$$

Puisque $\mathrm{C}\alpha < 1$, ceci nous garantit que si $x \in \mathrm{B}(x^\star, \alpha)$ alors $\mathrm{N}(x) \in \mathrm{B}(x^\star, \alpha)$. Ainsi, comme $x_{k+1} = \mathrm{N}(x_k)$, on en déduit par récurrence que si $x_0 \in \mathrm{B}(x^\star, \alpha)$ alors la suite de Newton est bien définie et $x_k \in \mathrm{B}(x^\star, \alpha)$ pour tout k. Pour la fin de la preuve, nous nous plaçons dans cette situation.

Convergence. L'inégalite $(*)$ permet alors de montrer que $(x_k)_{k\in\mathbb{N}}$ converge vers $x^\star$. On a en effet

$$\|x_{k+1} - x^\star\| \leqslant \mathrm{C}\|x_k - x^\star\|^2,$$

ce qui entraine
$$\mathrm{C}\|x_{k+1} - x^\star\| \leqslant (\mathrm{C}\|x_k - x^\star\|)^2.$$

Par récurrence, on obtient donc
$$\mathrm{C}\|x_k - x^\star\| \leqslant (\mathrm{C}\|x_0 - x^\star\|)^{2^k},$$

ce qui donne finalement
$$\|x_k - x^\star\| \leqslant (\mathrm{C}\alpha)^{2^k}/\mathrm{C}.$$

Comme $\mathrm{C}\alpha < 1$, ceci implique que $\|x_k - x^\star\|$ tend (très vite) vers 0. ■

3. Commentaires

Nous nous contentons ici de quelques remarques sur le théorème 1. Après ces remarques (et après avoir beaucoup parlé des avantages de la méthode de Newton), nous allons aussi parler un peu de ses défauts.

Remarque 2 — Minimisation vs équation. Rappelons que la méthode de Newton pour minimiser f correspond à la méthode de Newton pour résoudre l'equation $g(x) = 0$ avec $g = \nabla f$; voir les quelques mots à ce propos à l'exemple 1.53 de [BPM]. Ces « deux » méthodes de Newton consistent donc en fait au même processus. Il est ainsi normal de pouvoir traduire ce théorème et cette preuve pour la méthode de Newton pour résoudre $g(x) = 0$ (avec $g\colon \mathbb{R}^n \to \mathbb{R}^n$).

Remarque 3 — Convexité cachée. Il est intéressant de noter que, sous les hypothèses du théorème, la fonction f est convexe, et même « très convexe ». On démontre en effet au lemme 9 la α-convexité de f sur U qui entraîne sa convexité.

Remarque 4 — Convergence quadratique. La preuve du théorème établit que la vitesse de convergence asymptotique de $(x_k)_{k\in\mathbb{N}}$ est quadratique : si x_0 est proche de $x^\star$ alors la suite converge et on a

$$\|x_{k+1} - x^\star\| \leqslant \mathrm{C}\|x_k - x^\star\|^2.$$

Le théorème garantit donc d'une part la pertinence de la méthode de Newton, et d'autre part son efficacité. La convergence quadratique est en effet gage de qualité : dès que les itérés seront proches de la solution $x^\star$, il vont s'en rapprocher très vite. Ce phénomème se voit bien en une dimension : le nombre de chiffres significatifs double à chaque itération, puisque si $|x_k - x^\star| \leqslant 10^{-p}$, alors $|x_{k+1} - x^\star| \leqslant 10^{-2p}$.

Remarque 5 — Défauts de la méthode de Newton. Pointons du doigt les défauts majeurs (et parfois rédhibitoires) de la méthode de Newton.

- L'information sur le second ordre de f n'est pas toujours disponible en pratique. En effet, la plupart du temps, les fonctions à minimiser ne sont pas connues explicitement, mais simplement via un algorithme calculant $f(x)$ et $\nabla f(x)$ pour un x donné. C'est le cas en particulier quand les fonctions sont issues de calculs en amont (par exemple après la résolution d'une équation différentielle). Dans ces cas, même si la régularité de la fonction peut être garantie par ailleurs, on ne peut pas appliquer la méthode de Newton car toute l'information n'est pas disponible.
- Même en supposant être dans une situation où on peut calculer le hessien, la méthode de Newton doit parfois être proscrite pour des raisons numériques. En effet, chaque itération nécessite de résoudre un système linéaire pour calculer la direction de descente. Si la matrice hessienne est mal conditionnée ou, même pire, si elle n'est pas inversible (ce qui peut tout à fait arriver loin de $x^\star$), la direction de Newton n'est pas pertinente.

- La méthode diverge souvent fortement, si x_0 n'est pas bien choisi. En effet, même si le minimum est un bon minimum, on ne peut pas garantir d'être dans le bassin d'attraction où la fonction est bien convexe. Enfin, il se peut tout à fait que les itérés soient attirés par un maximum local...

Ainsi, si la principale qualité de la méthode de Newton est bien connue (quand elle converge, elle converge vite), ses défauts le sont aussi. C'est justement pour répondre à ces griefs que différentes améliorations de la méthode de Newton ont été développées (notamment les méthodes de Newton tronqué et les méthodes de quasi-Newton). Aiguillons les lecteurs courageux vers le livre de référence sur ces questions : « Numerical Optimization » de J.F. Bonnans, J.C. Gilbert, C. Lemaréchal et C. Sagastizábal, édité chez Springer.

4. Appendice

Ouvrons une parenthèse pour mettre en avant un résultat qui nous sera utile dans la suite.

Lemme 6 – $\mathcal{S}_n^{++}$ est un ouvert. L'ensemble $\mathcal{S}_n^{++}$ des matrices definies positives de taille n forme un ouvert de l'espace des matrices symétriques. En fait, pour une matrice $\mathrm{A} \in \mathcal{S}_n^{++}$, on a

$$\exists\,\beta > 0, \quad \forall\,\mathrm{B} \text{ proche de A}, \quad \forall\, h \in \mathbb{R}^n, \qquad {}^t h\,\mathrm{B}\,h \geqslant \beta \|h\|^2$$

Preuve. Rappelons tout d'abord que pour une matrice symétrique S, on a l'équivalence

$$\text{S définie positive} \iff \exists\,\alpha > 0, \quad \alpha\|x\|^2 \leqslant \langle x, \mathrm{S}x\rangle.$$

Ceci s'obtient en invoquant l'équivalence des normes en dimension finie : une matrice symétrique définie positive induit une norme sur $\mathbb{R}^n$ où toutes les normes sont équivalentes. On peut aussi avoir un résultat plus précis en utilisant directement le théorème de réduction en base orthonormée : en écrivant $\langle x, \mathrm{S}x\rangle$ et en décomposant x dans une base orthonormée de vecteurs propres de S, on montre que le meilleur α possible est $\lambda_{\min}(\mathrm{S})$.

Démontrons à présent le lemme en lui-même. Soient S une matrice dans $\mathcal{S}_n^{++}$ et T une matrice symétrique de norme inférieure à $\alpha/2$ (pour la norme subordonnée) ; alors S+T est aussi dans $\mathcal{S}_n^{++}$, puisqu'on peut écrire

$$\langle x, (\mathrm{S}+\mathrm{T})x\rangle \geqslant \alpha\|x\|^2 + \langle x, \mathrm{T}x\rangle \geqslant \|x\|^2 - \|x\|\|\mathrm{T}x\| \geqslant \alpha\|x\|^2 - \|\mathrm{T}\|\|x\|^2 \geqslant \alpha/2\|x\|^2$$

en utilisant Cauchy-Schwarz et la définition de la norme subordonnée. ■

Remarque 7 – Avec la convexité en plus. Nous donnons au passage une autre (rapide) preuve du lemme précédent utilisant les deux résultats suivants :

- L'application 1.69 de [BPM] démontre que la fonction $\mathrm{A} \mapsto \lambda_{\max}(\mathrm{A})$ qui, à une matrice symétrique, associe sa plus grande valeur propre est une fonction convexe.
- La remarque 1.70 de [BPM] explique en particulier qu'une fonction convexe est continue à l'intérieur de son domaine de définition.

En assemblant ces deux résultats, on montre facilement que l'application

$$\lambda_{\min} \colon \mathrm{A} \longmapsto \lambda_{\min}(\mathrm{A}) = -\lambda_{\max}(-\mathrm{A})$$

est une fonction concave continue sur l'espace des matrices symétriques. On en déduit directement que

$$\mathcal{S}_n^{++} = {\lambda_{\min}}^{-1}(\mathbb{R}^{+*})$$

est un ouvert de cet espace, et on obtient en prime que c'est un convexe (comme ensemble de niveau d'une fonction concave - cf exemple 1.64 de [BPM]). ■

Fermons la parenthèse et établissons la propriété de convexité de f sous les hypothèses du théorème 1. Rappelons tout d'abord (voir la courte remarque 1.74 de [BPM]) qu'une fonction f est dite α-convexe pour $\alpha > 0$ si elle s'ecrit comme la somme

$$f(x) = g(x) + \alpha\,\|x\|^2 \qquad \text{avec } g \text{ convexe.}$$

Ceci peut s'interpréter comme un ajout d'encore plus de convexité (le terme $\alpha\|\cdot\|^2$) à la fonction convexe g : on obtient donc une fonction f « très » convexe. En particulier, une fonction α-convexe est strictement convexe (car $\alpha\|\cdot\|^2$ est strictement convexe). Passer par la α-convexité est d'ailleurs une manière intéressante de démontrer la stricte convexité d'une fonction puisque l'on dispose de la caractérisation suivante.

Lemme 8 – α-convexité. Soit f une fonction deux fois différentiable sur un ouvert convexe $\mathcal{U}$; alors f est α-convexe sur $\mathcal{U}$ si et seulement si

$$\forall\, x \in \mathcal{U}, \quad \forall\, h \in \mathbb{R}^n, \qquad {}^t h\,\mathrm{H}f(x)\,h \geqslant \alpha\|h\|^2.$$

Preuve. Rappelons tout d'abord la caractérisation des fonctions convexes par leur hessien. Soit $f : \mathcal{U} \to \mathbb{R}$ une fonction deux fois différentiable, alors on a

$$f \text{ convexe sur } \mathcal{U} \iff \Big(\forall\, x \in \mathcal{U},\ \forall\, h \in \mathbb{R}^n, \quad {}^t h \, \mathrm{H} f(x)\, h \geqslant 0\Big)$$

Cette propriété nous permet de compléter les équivalences suivantes sur $\mathcal{U}$

$$\begin{aligned} f \text{ est } \alpha\text{-convexe} \quad &\iff \quad g = f - \alpha \|\cdot\|^2 \text{ est convexe} \\ &\iff \quad {}^t h \,\mathrm{H}\, g(x)\, h \geqslant 0 \quad \text{pour tout } h \in \mathbb{R}^n \\ &\iff \quad {}^t h \,\big(\mathrm{H} f(x) - \alpha\, \mathrm{Id}\big)\, h \geqslant 0 \quad \text{pour tout } h \in \mathbb{R}^n \\ &\iff \quad {}^t h \,\mathrm{H} f(x)\, h \geqslant \alpha\, \|h\|^2 \quad \text{pour tout } h \in \mathbb{R}^n \end{aligned}$$

ce qui donne l'équivalence souhaitée.

Lemme 9 – Convexité cachée 2. Soit f une fonction de classe C^2 autour d'un point $x^\star$ telle que $\mathrm{H} f(x^\star)$ est définie positive, alors f est α-convexe dans un voisinage de de $x^\star$.

Preuve. Comme f est C^2, la fonction $x \mapsto \mathrm{H} f(x)$ est continue. Or $\mathrm{H} f(x^\star) \in \mathcal{S}_n^{++}$, donc grâce au lemme 6, on établit que

$$\exists\, \alpha > 0, \quad \forall\, x \text{ proche de } x^\star, \quad \forall\, h \in \mathbb{R}^n, \qquad {}^t h\, \mathrm{H} f(x)\, h \geqslant \alpha \|h\|^2.$$

ce qui permet de conclure avec le lemme 8.

Références

[BPM] V. Beck, J. Malick, et G.Peyré. *Objectif Agrégation.* H & K, 2004.