Gradient et hyperplan d'appui

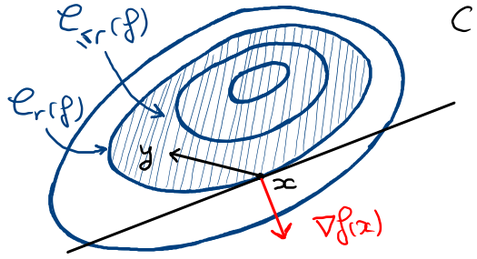
Le gradient d'une fonction convexe en un point donné fournit une information géométrique précieuse dans l'optique de minimiser cette fonction

Soit $x \in \mathcal{C}_r(f)$ ($f(x)=r$) et $y \in \mathcal{C}_{\leq r}(f)$ (y dans le lieu de sous niveau r de f)

La convexité à l'ordre 1 donne : $\underbrace{f(y)}_{\leq r} - \underbrace{f(x)}_{=r} \geq \nabla f(x)^T (y-x)$

$\rightarrow$ $\boxed{\nabla f(x)^T (y-x) \leq 0}$ C'est la définition d'un hyperplan d'appui en x à $\mathcal{C}_{\leq r}(f)$ de vecteur normal $\nabla f(x)$

$\Rightarrow$ Le gradient de f en x permet de définir un hyperplan d'appui au lieu de sous-niveau r $\rightarrow$ les valeurs plus petites que r se trouvent uniquement dans le demi espace opposé au gradient $\nabla f(x)$



C'est cette propriété géométrique qui est à la base des méthodes de descente de gradient

$\rightarrow$ En un point x_n donné, $\nabla f(x_n)$ pointe dans la direction de plus forte pente et définit un hyperplan d'appui à $\mathcal{C}_{\leq f(x_n)}(f)$

$\rightarrow$ On peut exclure tout le demi espace positif comme zone de recherche de x_{n+1} pour minimiser f. On part donc dans la direction opposée : $x_{n+1} = x_n - \alpha \nabla f(x_n)$, avec α le pas de la descente (appelé learning rate en machine learning)

Le choix de α est crucial pour la convergence de la méthode, mais s'il est bien choisi, alors $f(x_{n+1}) < f(x_n)$

$\rightarrow$ On itère comme cela, en faisant éventuellement varier α : $\boxed{x_{n+1} = x_n - \alpha_n \nabla f(x_n)}$

Si f est convexe, cette stratégie nous permet en théorie d'espérer trouver un point optimal x^* (donc $\nabla f(x^*)=0$)

On peut donc arrêter la descente au bout d'un certain nombre d'itérations, ou lorsque $\|\nabla f(x_n)\| < \varepsilon$ (auquel cas $x_n \simeq x^*$)

Si f n'est pas convexe, cette stratégie, si elle nous permet de trouver un point critique $\hat{x}$ tq $\nabla f(\hat{x}) = 0$, ne garantit pas la nature du point critique en question

Il y a donc besoin de caractériser plus finement les points critiques

3) Développement limité à l'ordre 2

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction différentiable en $x_0 \in \mathbb{R}^n$, de différentielle $df_{x_0} : h \longmapsto \nabla f(x_0)^T h$

On dit que f est <u>2-fois différentiable</u> en x_0 si l'application gradient de f : $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $x \longmapsto \nabla f(x)$ est elle même différentiable

$\rightarrow$ Comment s'écrit la différentielle de ∇f ?

∇f étant une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$, sa différentielle s'exprime donc via sa matrice jacobienne $\text{Jac}\,\nabla f(x_0)$ au point x_0 considéré : $d\nabla f_{x_0} : h \longmapsto (\text{Jac}\,\nabla f(x_0)) \times h$ et $\text{Jac}\,\nabla f(x_0) \in \mathbb{R}^{n \times n}$ est une matrice carrée.

Puisque $\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^T$, on peut écrire sa matrice jacobienne :

$$\text{Jac}\,\nabla f(x) = \begin{pmatrix} \nabla\left(\frac{\partial f}{\partial x_1}\right)(x)^T \\ \vdots \\ \nabla\left(\frac{\partial f}{\partial x_n}\right)(x)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1}(x)\right), \dots, \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_1}(x)\right) \\ \\ \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_n}(x)\right), \dots, \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_n}(x)\right) \end{pmatrix}$$

En notant $\frac{\partial^2 f}{\partial x_i^2}(x) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}(x)\right)$ et $\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}(x)\right)$, on obtient la matrice hessienne de f au point x c'est à dire la matrice jacobienne du gradient de f.

$$H_f(x) = \text{Jac}\,\nabla f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x), \dots, \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots \quad \ddots \quad \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x), \dots, \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}$$

Si toutes les fonctions $\frac{\partial^2 f}{\partial x_i \partial x_j}$ sont continues, alors $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ (théorème de Schwarz), et la matrice hessienne est symétrique

La hessienne d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ regroupe donc sous forme matricielle toutes les dérivées secondes de f

Propriété : développement limité à l'ordre 2

Soit $f : \mathbb{R}^n \to \mathbb{R}$ 2-fois différentiable en x_0

Alors $f(x_0+h) = f(x_0) + \nabla f(x_0)^T h + \frac{1}{2} h^T H_f(x_0) h + o_0(\|h\|^2)$

→ C'est la matrice hessienne qui apparait dans le terme quadratique du DL_2

Pour $f : \mathbb{R} \to \mathbb{R}$: $f(x_0+h) = f(x_0) + h f'(x_0) + \frac{1}{2} h^2 f''(x_0) + o_0(h^2)$

→ La hessienne est donc la généralisation de la dérivée d'ordre 2 (tout comme le gradient généralise la dérivée d'ordre 1).

4) Caractérisation des points critiques

Soit x^* un point critique d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ (donc $\nabla f(x^*) = 0$)

Si f est convexe, x^* est un minimum global

Si f n'est en revanche pas convexe, x^* peut être un minimum local ou global, un maximum local ou global, ou bien un point selle.