Les méthodes itératives de descente

Objectif: On se donne une fonction $f : \mathbb{R}^n \to \mathbb{R}$ supposée différentiable (mais pas nécessairement convexe) et on cherche à résoudre $\boxed{(\arg)\min_{x \in \mathbb{R}^n} f(x)}$

→ Si f est convexe, on sait que x^* minimum global $\Leftrightarrow \nabla f(x^*) = 0$, la solution du problème d'optimisation est donnée par les points critiques de f

→ Si f n'est pas convexe, on perd cette garantie. On sait que si x^* est un minimum (local), alors $\nabla f(x^*) = 0$, mais l'inverse est faux ($\nabla f(x^*) = 0$ implique que x^* est un extremum local ou un point selle).

MAIS l'étude géométrique du gradient nous dit qu'en un point x donné, $\nabla f(x)$ pointe dans la direction de plus forte pente (que f soit convexe ou non). Faire un petit pas à l'opposé de $\nabla f(x)$ doit permettre de faire décroître la valeur de f.

Si d_k est une direction opposée à $\nabla f(x_k)$ et η_k un petit pas, on peut construire $x_{k+1} = x_k + \eta_k d_k$ tel que $f(x_{k+1}) < f(x_k)$

→ Il "suffit" d'itérer cette procédure pour faire décroître la valeur de f, on peut donc espérer atteindre un minimum (à minima local)

Procédure générale d'une méthode de descente

Étant donné un point de départ x_0

Tant que : critère d'arrêt non satisfait

→ calcul d'une direction de descente d_k

→ calcul d'un pas de descente acceptable $\eta_k > 0$ dans la direction d_k

→ $\boxed{x_{k+1} = x_k + \eta_k d_k}$

Cette procédure soulève un certain nombre de questions:

→ Comment choisir la condition initiale x_0 ?

→ Comment choisir la direction de descente d_k ?

→ Comment choisir un pas de descente acceptable η_k ?

→ Comment choisir le critère d'arrêt ?

→ Comment garantir la convergence de la méthode ?

1) Choix de la condition initiale

Si f est convexe, on peut écrire des résultats de convergence (cf plus loin) en fonction des choix de la direction de descente d_k et du pas de descente η_k

→ En fonction de ces choix, l'impact de x_0 est limité (garantie de convergence peu importe la condition initiale)

Si f n'est pas convexe, on peut quand même écrire certains résultats de convergence, mais on perd la garantie que la méthode converge vers le même minimum local en fonction du choix de x_0 → choisir x_0 pour que la méthode converge vers le minimum global (s'il existe) est une question qui n'a malheureusement pas de réponse. Il faut dans ce cas se tourner vers des heuristiques telles que de multiples initialisations aléatoires de x_0 et comparaison des solutions obtenues, etc

2) Choix de la direction de descente

On formalise un peu la notion de direction de descente (car toutes les méthodes de descente ne suivent pas forcément la direction opposée à $\nabla f(x)$)

Rappel : on appelle dérivée directionnelle de f en x selon le vecteur $d \in \mathbb{R}^n \setminus \{0\}$ la dérivée en 0 (si elle existe) de la fonction $\varphi : t \mapsto f(x+td)$

$$\rightarrow \boxed{D_d f(x) = \varphi'(0) = \lim_{t \to 0} \frac{f(x+td) - f(x)}{t}}$$

$\varphi(t) = f(x+td) \equiv$ restriction de f à la droite passant par x et de vecteur directeur d

Si $\|d\| = 1$, on parle de dérivée dans la direction de d

Si $D_d f(x)$ existe, alors $D_{\alpha d} f(x)$ existe et $D_{\alpha d} f(x) = \alpha D_d f(x)$

La dérivée partielle $\frac{\partial f}{\partial x_i}(x)$ est la dérivée selon l'axe engendré par e_i (i^e vecteur de la base canonique)

$$\rightarrow \frac{\partial f}{\partial x_i}(x) = D_{e_i} f(x)$$

Rappel : si f est différentiable en x, de différentielle df_x, alors f admet une dérivée directionnelle dans toute direction d et $D_d f(x) = df_x(d) = \nabla f(x)^T d$

Définition : direction de descente

Soit $f : \mathbb{R}^n \to \mathbb{R}$ différentiable en x. On dit que $d \in \mathbb{R}^n \setminus \{0\}$ est une direction de descente en x si $\varphi : t \mapsto f(x+td)$ est strictement décroissante au voisinage de 0, c'est à dire $\varphi'(0) = D_d f(x) < 0$

→ il existe $c > 0$ tel que $\forall t \in [0, c[$, $f(x+td) < f(x)$

d est donc une direction de descente en x ssi $\boxed{\nabla f(x)^T d < 0}$ (avec $\nabla f(x) \neq 0$)

→ prendre $d = -\nabla f(x)$ donne bien une direction de descente puisque $\nabla f(x)^T d = \nabla f(x)^T (-\nabla f(x)) = -\|\nabla f(x)\|^2 < 0$

C'est même la direction de plus forte pente : prenons une autre direction de descente d (donc $\nabla f(x)^T d < 0$) de même norme que $\nabla f(x)$ (donc $d \neq -\nabla f(x)$ mais $\|d\| = \|\nabla f(x)\|$)

$\nabla f(x)^T d = \langle \nabla f(x), d \rangle$ et utilisons l'inégalité de Cauchy Schwartz ($|\langle x, y \rangle| \leq \|x\| . \|y\|$)

$\rightarrow |\nabla f(x)^T d| = |\langle \nabla f(x), d \rangle| \leq \|\nabla f(x)\| . \|d\| = \|\nabla f(x)\| . \|\nabla f(x)\| = \|\nabla f(x)\|^2 = \nabla f(x)^T \nabla f(x)$

Et $|\nabla f(x)^T d| = -\nabla f(x)^T d$ puisque $\nabla f(x)^T d < 0$ (d direction de descente)