Donc $-\nabla f(x)^T d \leqslant \nabla f(x)^T \nabla f(x) \iff \underbrace{\nabla f(x)^T(-\nabla f(x))}_{D_{-\nabla f(x)} f(x)} \leqslant \underbrace{\nabla f(x)^T d}_{D_d f(x)}$

La direction de descente selon $-\nabla f(x)$ est donc plus pentue que selon d : $-\nabla f(x)$ est bien la direction de plus forte pente

3) Choix du critère d'arret

→ à partir de quel moment peut-on stopper la descente ?

Rappel : condition d'optimalité : f admet un minimum local en x^* si :

- $\nabla f(x^*) = 0$ x^* est point critique (optimalité du premier ordre)
- $H_f(x^*) \succ 0$ La hessienne de f en x^* est définie positive (optimalité du second ordre)

Si f est convexe, alors $H_f(x) \succeq 0$ en tout point (caractérisation à l'ordre 2 de la convexité) et
$f(x) = f(x^*) + \underbrace{\nabla f(x^*)^T}_{=0}(x - x^*) = f(x^*)$ donc x^* point critique est un minimum global

Le but du critère d'arret est que l'algorithme de descente s'arrête en un point suffisamment proche d'un minimum local (ou global dans le cas convexe)

→ Un critère d'arret naturel basé sur la condition d'optimalité du premier ordre est donc :

$\boxed{\|\nabla f(x_k)\| < \varepsilon}$ avec $\varepsilon > 0$ une tolérance fournie par l'utilisateur

Si f est convexe, ce critère est à priori suffisant. Dans le cas contraire, on pourrait vérifier que la hessienne est bien définie positive en x_k : $H_f(x_k) \succ 0$

Mais c'est cher en calcul, sauf dans des cas particuliers (fonctions quadratiques). On préfère donc en général d'autres heuristiques en pratique :

→ Stagnation (relative) de l'itéré : $\dfrac{\|x_{k+1} - x_k\|}{\|x_k\|} < \varepsilon$

→ Stagnation (relative) de la valeur courante : $\dfrac{|f(x_{k+1}) - f(x_k)|}{|f(x_k)|} < \varepsilon$

→ Nombre d'itérations : $k < MaxIter$

En général : on choisit toujours de limiter le nombre d'itérations, combiné au test d'optimalité ou à un critère de stagnation

4) Choix du pas de descente

Une fois la direction de descente d_k choisie, il faut déterminer un pas de descente η_k tq $f(x_k + \eta_k d_k) < f(x_k)$

Cette étape s'appelle la phase de recherche linéaire

Objectif : trouver le meilleur pas en faisant le moins de calculs possible (un peu contradictoire ...)

4-1) Descente de gradient à pas fixe

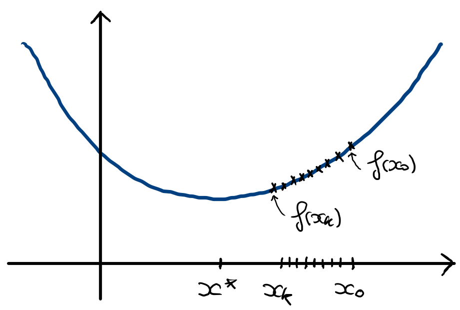
Dans un schéma de descente à pas fixe, l'itération est donnée par $\boxed{x_{k+1} = x_k + \eta d_k}$ avec η un pas constant et d_k la direction de descente

Si $d_k = -\nabla f(x_k)$, on parle alors de descente de gradient à pas fixe

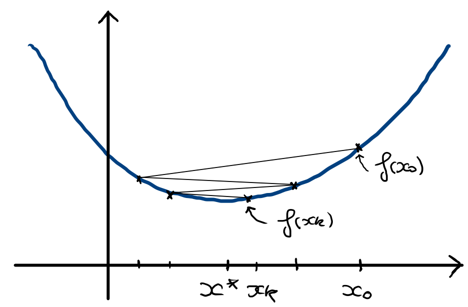
→ Est-il possible de choisir le pas η pour garantir la convergence de la méthode ?

→ Existe-t-il des valeurs du pas qui permettent de converger plus vite ?

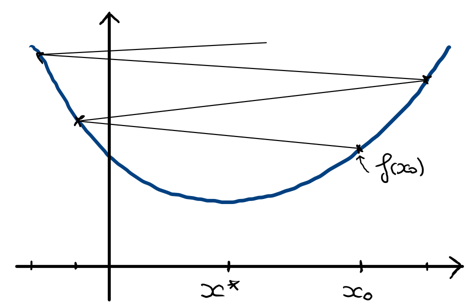
Exemple : pour une fonction $f : \mathbb{R} \to \mathbb{R}$



Pas trop petit, mais la descente converge



Pas trop grand, mais la descente converge



Pas trop grand, et la descente diverge

Dans les deux premiers cas, la descente est sous-optimale, bien que convergente. Au delà d'une certaine limite pour la valeur du pas, la descente ne converge plus.

Définition : Soit $(x_k)_{k\in\mathbb{N}}$ une suite d'itérés qui converge vers x^* : $x_k \underset{k\to+\infty}{\longrightarrow} x^*$

On dit que la convergence est linéaire si l'erreur $e_k = \|x_k - x^*\|$ décroît linéairement. Autrement dit :

* il existe $\tau \in]0,1[$ tq $\lim\limits_{k\to+\infty} \dfrac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \tau$

* il existe $c > 0$ et $\tau \in]0,1[$ tq $\|x_k - x^*\| \leq c\,\tau^k$ (les deux formulations sont équivalentes)

La plus petite valeur de τ qui vérifie la condition précédente est appelée taux de convergence

Exemple : On considère une fonction quadratique convexe $f : \begin{array}{l} \mathbb{R}^n \to \mathbb{R} \\ x \mapsto \frac{1}{2} x^T A x + b^T x \end{array}$ avec A symétrique définie positive

Puisque $\nabla f(x) = Ax + b$, l'itération de la descente de gradient à pas constant s'écrit $x_{k+1} = x_k - \eta \nabla f(x_k)$

→ $x_{k+1} = x_k - \eta(Ax_k + b) = (I_d - \eta A)x_k - \eta b$

Les conditions sur la valeur de η qui garantissent la convergence sont données par la propriété suivante

Propriété (convergence dans le cas quadratique)

Soient $\ell > 0$ et $L > \ell$ la plus petite et la plus grande valeur propre de A. Alors, pour $\eta \in]0, \frac{2}{L}[$, la descente de gradient converge vers le minimum global, et la convergence est linéaire

Le meilleur (plus petit) taux de convergence vaut $\tau = \dfrac{L-\ell}{L+\ell}$, et le pas (optimal) associé est $\eta = \dfrac{2}{L+\ell}$