Le taux de convergence est lié au nombre de conditionnement de la matrice A (qui vaut $K(A) = \frac{L}{\ell}$, rapport de la plus grande et la plus petite valeur propre)

Si $L \gg \ell$, alors $K(A) \gg 1$: la matrice A est mal conditionnée. Dans ce cas, le taux de convergence est $\tau \simeq 1$

$\Rightarrow$ La descente converge d'autant plus lentement que la matrice A est mal conditionnée

Dans le cas plus général d'une fonction f pas nécessairement quadratique, ça devient nettement plus difficile (pour ne pas dire impossible) de trouver des conditions sur la valeur du pas pour garantir la convergence de la descente : on a besoin d'imposer une régularité supplémentaire à f, à savoir que son gradient ∇f soit M-Lipschitzien

Rappel : On dit que f est k-Lipschitzienne, $k > 0$, ssi $\forall x, y$, $\|f(x) - f(y)\| \leqslant k\|x-y\|$

Propriété : Soit $f : \mathbb{R}^n \to \mathbb{R}$ différentiable avec ∇f M-Lipschitzien

Alors $\forall x, y$, $f(y) \leqslant f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|^2$

Si f est 2-fois différentiable, ∇f M-Lipschitzien $\Longleftrightarrow$ $H_f(x) \leqslant M I_d$

↳ au sens des formes quadratiques, donc $\forall y \in \mathbb{R}^n$, $y^T H_f(x) y \leqslant y^T (M I_d) y = M y^T y$

Grâce à cette propriété, on peut montrer que pour f pas nécessairement convexe, un pas $\eta < \frac{2}{M}$ garantit la convergence de l'algorithme de descente (au sens où $(f(x_k))_{k\in\mathbb{N}}$ converge vers une valeur finie si f est bornée inférieurement, et $\|\nabla f(x_k)\| \xrightarrow[k \to +\infty]{} 0$) et que le pas optimal est $\eta = \frac{1}{M}$

Mais ce résultat n'implique pas que la descente converge vers la valeur optimale ...

$\Rightarrow$ Il faut imposer la convexité de f comme condition supplémentaire

Théorème : Soit f une fonction convexe, de gradient ∇f M-Lipschitzien. Alors, après k itérations de descente de gradient avec un pas constant $\eta \leqslant \frac{1}{M}$, on a $\boxed{f(x_k) - f(x^*) \leqslant \frac{\|x_0 - x^*\|^2}{2\eta k}}$

- La condition initiale x_0 a donc un impact : plus x_0 est loin de x^* et plus la valeur $f(x_k)$ après k itérations sera loin de la valeur optimale $f(x^*)$
- Le pas optimal est $\eta = \frac{1}{M}$, auquel cas $f(x_k) - f(x^*) \leqslant \frac{M\|x_0 - x^*\|^2}{2k}$
- Le taux de convergence de la descente est en $O(\frac{1}{k})$: pour avoir $f(x_k) - f(x^*) \leqslant \varepsilon$, il faut $O(\frac{1}{\varepsilon})$ itérations, la convergence est sous linéaire

Si on suppose de plus que f est fortement convexe : $H_f(x) \geqslant m I_d$ avec $m > 0$ $\Longleftrightarrow$ la hessienne $H_f(x)$ de f est définie positive en tout point. Alors, après k itérations de descente avec un pas constant $\eta \leqslant \frac{1}{M}$, on a

$$\boxed{f(x_k) - f(x^*) \leqslant c^k \frac{M}{2}\|x_0 - x^*\|^2} \qquad \text{avec } c = 1 - \frac{m}{M} < 1$$

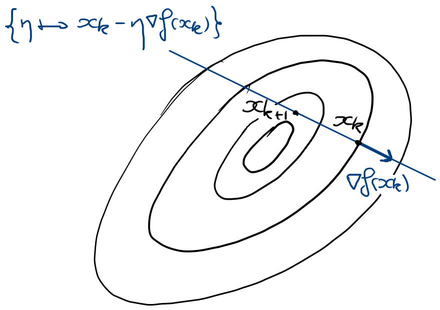
inverse du nombre de conditionnement de $H_f(x)$

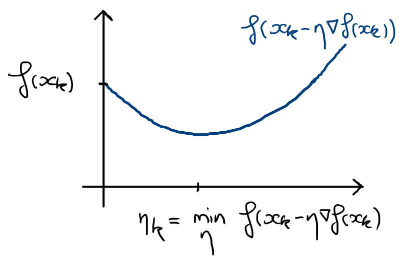
- le taux de convergence est maintenant en $O(c^k)$: pour avoir $f(x_k) - f(x^*) \leqslant \varepsilon$, il faut $O(\log(\frac{1}{\varepsilon}))$ itérations, la convergence devient linéaire (si tracée en échelle log

4-2) Descente de gradient à pas optimal

Dans un schéma de descente à pas optimal, on cherche à chaque itération le pas η_k qui est solution du sous problème
$\min_\eta f(x_k + \eta d_k)$ (restriction de f à la droite engendrée par la direction de descente d_k)

Si $d_k = -\nabla f(x_k)$, on parle de descente de gradient à pas optimal
→ elle nécessite, à chaque itération, de résoudre un problème d'optimisation unidimensionnel : $\min_\eta \varphi : \eta \mapsto f(x_k - \eta \nabla f(x_k))$





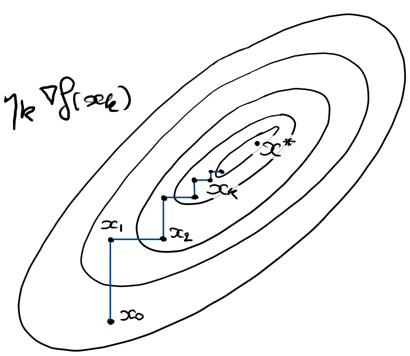
→ faisable, mais couteux (sauf dans le cas quadratique où on peut calculer une solution analytique)
→ se résoud en pratique de manière iterative, la solution ne sera qu'approchée quoi qu'il en soit

Propriété : Si η_k est le pas optimal solution de $\min_\eta f(x_k - \eta \nabla f(x_k))$ et l'itéré suivant est $x_{k+1} = x_k - \eta_k \nabla f(x_k)$
alors $\boxed{\langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle = 0}$

→ deux directions successives de descente sont orthogonales

Preuve : η_k est solution de $\min_\eta \varphi(\eta)$ avec $\varphi : \eta \mapsto f(x_k - \eta \nabla f(x_k))$
Donc $\varphi'(\eta_k) = 0$ (η_k est un point critique de φ puisque minimum global)

Or $\varphi'(\eta) = -\nabla f(x_k)^T f(x_k - \eta \nabla f(x_k))$ (dérivation en chaine)
→ $\varphi'(\eta_k) = -\nabla f(x_k)^T f(x_k - \eta_k \nabla f(x_k)) = -\nabla f(x_k)^T \nabla f(x_{k+1})$ avec $x_{k+1} = x_k - \eta_k \nabla f(x_k)$
→ $\varphi'(\eta_k) = 0 \iff \nabla f(x_k)^T \nabla f(x_{k+1}) = 0$
→ La descente de gradient à pas optimal itère en zigzaguant, ce qui est particulièrement lent lorsque la matrice hessienne de la fonction est mal conditionnée



4-3) Règles d'Armijo et de Wolfe

Pour une direction de descente d_k donnée, plutôt que de chercher le pas optimal η_k qui assure la meilleure décroissance de f le long de l'axe $\eta \mapsto x_k + \eta d_k$, on peut se contenter d'un pas "acceptable" qui garantit une décroissance suffisante de f le long de cette direction de descente : c'est le critère d'Armijo

Critère d'Armijo

Pour une direction de descente d_k, on dit que le pas η_k satisfait la règle d'Armijo pour $\alpha \in]0,1[$ si on a :

$$\boxed{f(x_k + \eta_k d_k) \leq f(x_k) + \alpha \eta_k \nabla f(x_k)^T d_k}$$

Pour comprendre la signification de ce critère, il faut écrire l'équation de la tangente au graphe de f en x_k dans la direction d_k, c'est à dire la tangente en 0 de la fonction $\varphi : \eta \mapsto f(x_k + \eta d_k)$