Rappel : Equation de la tangente au graphe de $f : \mathbb{R} \to \mathbb{R}$ au point $(x, f(x))$: $T_x : y \mapsto f(x) + (y-x) f'(x)$

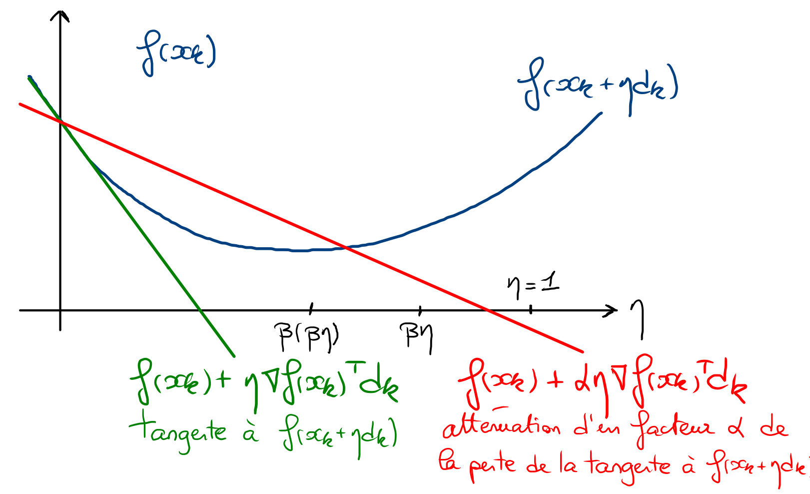
Ici, l'équation de la tangente à φ en 0 est donc $\eta \mapsto \varphi(0) + \eta \varphi'(0)$

Avec $\varphi(0) = f(x_k)$ et $\varphi'(0) = D_{d_k} f(x_k)$ la dérivée directionnelle de f en x_k selon la direction d_k

$\to \varphi'(0) = \nabla f(x_k)^T d_k$

L'équation de la tangente à f en x_k suivant la direction d_k est donc $\eta \mapsto f(x_k) + \eta \nabla f(x_k)^T d_k$

Pour $\alpha < 1$, la règle d'Armijo permet donc de trouver un pas η_k qui assure une certaine fraction de décroissance par rapport à la prédiction linéaire $\boxed{f(x_k + \eta_k d_k) \leqslant f(x_k) + \alpha (\eta_k \nabla f(x_k)^T d_k)}$



Pour trouver η_k en pratique, on a besoin d'un deuxième paramètre $\beta < 1$, et on applique à chaque itération k de la descente la sous boucle suivante :

- $\to \eta = 1$
- $\to$ Tant que $f(x_k + \eta d_k) > f(x_k) + \alpha \eta \nabla f(x_k)^T d_k$ (règle d'Armijo non respectée)

  $\eta = \beta \eta$

Soit la règle d'Armijo est respectée pour $\eta = 1$ (auquel cas $\eta_k = \eta = 1$)

Soit on retente en remplaçant η par $\beta\eta$ $\to$ η diminue par une progression géométrique de raison β

En pratique, on limite en général $\alpha < \frac{1}{2}$ (typiquement $\alpha = 0{,}1$ ou $\alpha = 0{,}3$), et β est typiquement pris dans l'intervalle $[0{,}1 , 0{,}8]$

L'utilisation du critère d'Armijo pour trouver un pas approximant le pas optimal ne permet cependant pas d'accélérer la convergence : avec les mêmes hypothèses que précédemment (f convexe et ∇f M-Lipschitzien), après k itérations de descente de gradient avec un pas calculé par la méthode d'Armijo, on a $\boxed{f(x_k) - f(x^*) \leqslant \dfrac{\|x_0 - x^*\|_2^2}{2 \eta_{min} k}}$ avec $\eta_{min} = \min(1, \frac{\beta}{M})$ et β la constante de mise à jour du pas

La méthode d'Armijo a tendance à fournir un pas trop petit : on peut la coupler à une autre méthode pour contrebalancer cet effet, par exemple la règle de Wolfe

Critère de Wolfe

Pour une direction de descente d_k, on dit que le pas η_k satisfait la règle de Wolfe pour α_1, α_2 (avec $0 < \alpha_1 < \alpha_2 < 1$) s'il satisfait la règle d'Armijo pour α_1 et $\boxed{\nabla f(x_k + \eta_k d_k)^T d_k \geqslant \alpha_2 (\nabla f(x_k)^T d_k)}$

- $\to \nabla f(x_k)^T d_k < 0$ car c'est la dérivée directionnelle de f en x_k dans la direction d_k (qui est une direction de descente)

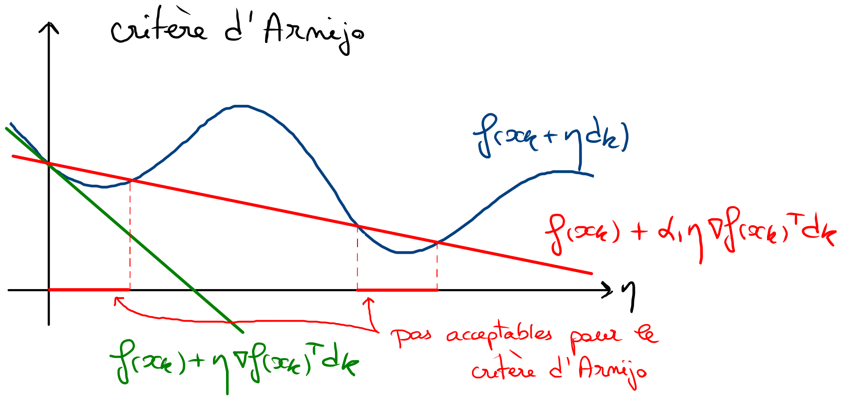
  $\to$ c'est la valeur en 0 de la dérivée de la fonction $\varphi : \eta \mapsto f(x_k + \eta d_k)$
- $\to \nabla f(x_k)^T d_k = \varphi'(0) < 0$

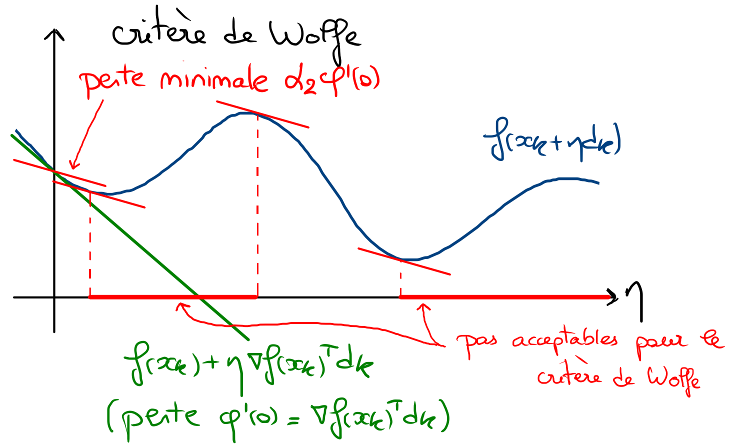
De son côté, le pas optimal $\eta^* = \min_\eta \varphi(\eta) = f(x_k+\eta d_k)$ vérifie $\varphi'(\eta^*) = 0$

$\Rightarrow$ φ' augmente (pas forcément de manière monotone) entre $\eta = 0$ et $\eta = \eta^*$

Pour un pas η_k donné, $\varphi'(\eta_k) = \nabla f(x_k+\eta_k d_k)^T d_k$ $\rightarrow$ le critère de Wolfe se réécrit donc $\boxed{\varphi'(\eta_k) \geq \alpha_2 \varphi'(0)}$

$\rightarrow$ le pas η_k doit être tel que la dérivée de $\eta \mapsto f(x_k+\eta d_k)$ ait suffisamment augmenté par rapport à sa valeur initiale (dérivée directionnelle négative)





Le critère de Wolfe (pris individuellement) est satisfait partout où la pente de la tangente en $f(x_k+\eta d_k)$ est supérieure à $\alpha_2 \times$ la pente de la tangente en $f(x_k)$

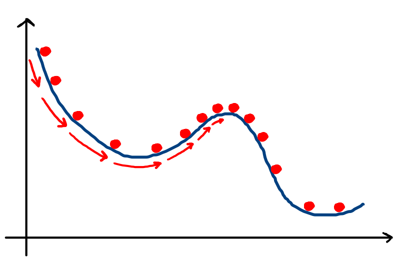
$\rightarrow$ la recherche se fait en pratique par dichotomie

5) Les méthodes d'accélération

Même dans le cas d'une fonction convexe, lorsque la convergence est garantie, il n'est pas toujours optimal de descendre en suivant la direction de plus forte pente $d_k = -\nabla f(x_k)$ (convergence lente, zigzag si la fonction est mal conditionnée, etc)

Si la fonction n'est pas convexe, la descente de gradient peut facilement se laisser capturer par des minimums locaux

Les méthodes d'accélération permettent (en partie) de résoudre ces problèmes)



L'analogie est celle d'une boule roulant sur un plan incliné : au fur et à mesure de sa descente elle va accumuler de l'énergie et de l'inertie (momentum), et cette inertie peut la faire remonter de l'autre côté d'un minimum local

$\rightarrow$ le mouvement en un point donné ne dépend pas que de la pente locale, mais aussi de la quantité d'inertie accumulée lors de la descente

Une manière équivalente de voir les choses est que si toutes les directions de descente précédentes pointent dans la même direction, alors on peut être confiant et y aller plus vite.

Descente avec momentum

On ajoute à chaque itération une fraction des itérations précédentes