# Les méthodes de descente

Guillaume Tochon
Laboratoire de Recherche de l'EPITA

# Rappels sur la géométrie du gradient

Dans toute la suite, on se donne une fonction $f : \mathbb{R}^n \to \mathbb{R}$

- supposée **différentiable**
- pas nécessairement convexe (mais on appréciera particulièrement celles qui le sont...)
- dont on cherche un minimiseur : $\boxed{\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})}$ et $f^\star = f(\mathbf{x}^\star) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

# Rappels sur la géométrie du gradient

Dans toute la suite, on se donne une fonction $f : \mathbb{R}^n \to \mathbb{R}$

- supposée **différentiable**
- pas nécessairement convexe (mais on appréciera particulièrement celles qui le sont...)
- dont on cherche un minimiseur : $\boxed{\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\, f(\mathbf{x})}$ et $f^\star = f(\mathbf{x}^\star) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

Si $f$ est convexe :

$\to$ $\boxed{\nabla f(\mathbf{x}^\star) = \mathbf{0} \Leftrightarrow \mathbf{x}^\star \text{ minimum global}}$ $\Rightarrow$ le minimiseur est un point critique de $f$

Si $f$ n'est pas convexe, cette garantie ne tient plus :

$\to$ $\nabla f(\mathbf{x}^\star) = \mathbf{0} \Leftrightarrow \mathbf{x}^\star$ est un extremum (minimum *ou* maximum) *local* (donc pas nécessairement global) *ou* un point selle.

# Rappels sur la géométrie du gradient

Dans toute la suite, on se donne une fonction $f : \mathbb{R}^n \to \mathbb{R}$

- supposée **différentiable**
- pas nécessairement convexe (mais on appréciera particulièrement celles qui le sont...)
- dont on cherche un minimiseur : $\boxed{\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\, f(\mathbf{x})}$ et $f^\star = f(\mathbf{x}^\star) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

Si $f$ est convexe :

$\rightarrow$ $\boxed{\nabla f(\mathbf{x}^\star) = \mathbf{0} \Leftrightarrow \mathbf{x}^\star \text{ minimum global}}$ $\Rightarrow$ le minimiseur est un point critique de $f$

Si $f$ n'est pas convexe, cette garantie ne tient plus :

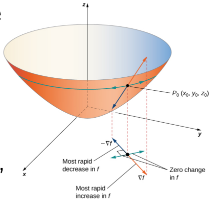
$\rightarrow$ $\nabla f(\mathbf{x}^\star) = \mathbf{0} \Leftrightarrow \mathbf{x}^\star$ est un extremum (minimum *ou* maximum) *local* (donc pas nécessairement global) *ou* un point selle.

Mais en un point $\mathbf{x}$ donné, $\nabla f(\mathbf{x})$ pointe dans la direction de plus forte pente montante

$\rightarrow$ faire un petit pas à l'opposée de $\nabla f(\mathbf{x})$ doit permettre de faire décroître la valeur de $f$.

$\rightarrow$ si $\mathbf{d}_k$ est une direction opposée à $\nabla f(\mathbf{x}_k)$ et $\eta_k$ un petit pas, on peut construire $\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k$ tel que $\boxed{f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)}$

$\rightarrow$ chaque itération de cette procédure fait décroître la valeur de $f$, on peut espérer atteindre un minimum (à minima local).



# Procédure générale d'une méthode de descente

Étant donnée une fonction $f : \mathbb{R}^n \to \mathbb{R}$ différentiable dont on cherche un minimiseur :

**Algorithme :** Procédure générale d'une méthode de descente

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$

**tant que** *critère d'arrêt non satisfait* **faire**

- $\to$ calcul d'une direction de descente $\mathbf{d}_k$
- $\to$ calcul d'un pas de descente "acceptable" $\eta_k > 0$ dans la direction $\mathbf{d}_k$
- $\to$ calcul du nouvel itéré $\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k}$

**fin**

**Sortie :** Minimum global $\mathbf{x}^\star$ (du moins, on l'espère 🤞)

# Procédure générale d'une méthode de descente

Étant donnée une fonction $f : \mathbb{R}^n \to \mathbb{R}$ différentiable dont on cherche un minimiseur :

---

**Algorithme :** Procédure générale d'une méthode de descente

---

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$

**tant que** *critère d'arrêt non satisfait* **faire**

- $\to$ calcul d'une direction de descente $\mathbf{d}_k$
- $\to$ calcul d'un pas de descente "acceptable" $\eta_k > 0$ dans la direction $\mathbf{d}_k$
- $\to$ calcul du nouvel itéré $\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k}$

**fin**

**Sortie :** Minimum global $\mathbf{x}^\star$ (du moins, on l'espère 🤞)

---

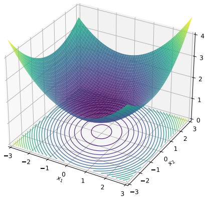
Évidemment, cette procédure vient avec son lot de questions, parmi lesquelles :

- $\to$ Comment choisir la condition initiale $\mathbf{x}_0$ ?
- $\to$ Comment choisir la direction de descente $\mathbf{d}_k$ ?
- $\to$ Comment choisir un pas de descente acceptable $\eta_k$ ?
- $\to$ Comment choisir le critère d'arrêt ?
- $\to$ Comment garantir la convergence de la méthode ?

# Choix de la condition initiale

## La question qui vaut chère 💰💰

Si $f$ est convexe :



Les facteurs critiques pour garantir la convergence sont la direction de descente $\mathbf{d}_k$ et le pas de descente $\eta_k$

→ s'ils sont bien choisis, alors l'impact de $\mathbf{x}_0$ est limité

→ limité, mais pas nul pour autant... (cf plus loin)

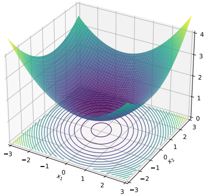
# Choix de la condition initiale

## La question qui vaut chère 💰💰

**Si $f$ est convexe :**

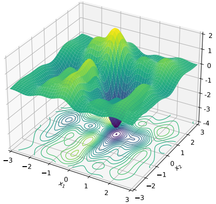
Les facteurs critiques pour garantir la convergence sont la direction de descente $\mathbf{d}_k$ et le pas de descente $\eta_k$

- $\rightarrow$ s'ils sont bien choisis, alors l'impact de $\mathbf{x}_0$ est limité
- $\rightarrow$ limité, mais pas nul pour autant... (cf plus loin)



**Si $f$ n'est pas convexe :**

On peut, dans certains cas, écrire des résultats de convergence, mais vers quoi ?



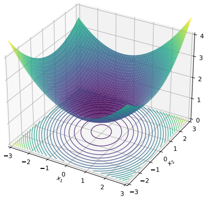
# Choix de la condition initiale

## La question qui vaut chère 💰💰

**Si $f$ est convexe :**

Les facteurs critiques pour garantir la convergence sont la direction de descente $\mathbf{d}_k$ et le pas de descente $\eta_k$

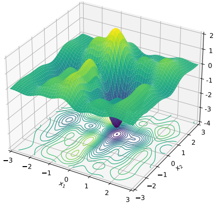
- $\rightarrow$ s'ils sont bien choisis, alors l'impact de $\mathbf{x}_0$ est limité
- $\rightarrow$ limité, mais pas nul pour autant... (cf plus loin)



**Si $f$ n'est pas convexe :**

On peut, dans certains cas, écrire des résultats de convergence, mais vers quoi ?

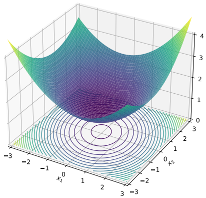
- $\rightarrow$ impact de $\mathbf{x}_0$ sur la position du minimum local atteint ?
- $\rightarrow$ choix de $\mathbf{x}_0$ pour trouver le meilleur minimum local ?



# Choix de la condition initiale

## La question qui vaut chère 💰💰

**Si $f$ est convexe :**

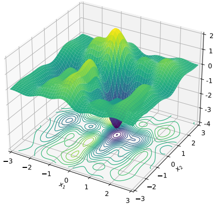


Les facteurs critiques pour garantir la convergence sont la direction de descente $\mathbf{d}_k$ et le pas de descente $\eta_k$

→ s'ils sont bien choisis, alors l'impact de $\mathbf{x}_0$ est limité

→ limité, mais pas nul pour autant... (cf plus loin)

**Si $f$ n'est pas convexe :**



On peut, dans certains cas, écrire des résultats de convergence, mais vers quoi ?

→ impact de $\mathbf{x}_0$ sur la position du minimum local atteint ?

→ choix de $\mathbf{x}_0$ pour trouver le meilleur minimum local ?

⇒ Utilisation d'heuristiques telles que de multiples initialisations aléatoires ou méthodes de descente accélérées...

⇒ ou simplement accepter le fait que la solution sera potentiellement sous-optimale.

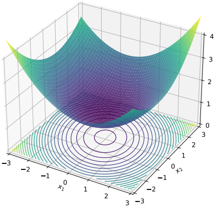
La définition d'une direction de descente $\mathbf{d}$ à partir d'un point $\mathbf{x}$ donné nécessite d'évaluer le comportement de $f$ lorsqu'on évolue "proche" de $\mathbf{x}$, selon l'axe engendré par $\mathbf{d}$ :

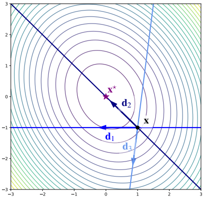
# Choix de la direction de descente

## Restriction de $f$ le long d'un axe

La définition d'une direction de descente $\mathbf{d}$ à partir d'un point $\mathbf{x}$ donné nécessite d'évaluer le comportement de $f$ lorsqu'on évolue "proche" de $\mathbf{x}$, selon l'axe engendré par $\mathbf{d}$ :

$\rightarrow$ Définition (paramétrique) de la droite $\mathcal{D}_{\mathbf{x},\mathbf{d}}$ passant par $\mathbf{x}$ et de vecteur directeur $\mathbf{d}$ : $\mathcal{D}_{\mathbf{x},\mathbf{d}} = \{x + t\mathbf{d}, t \in \mathbb{R}\}$.



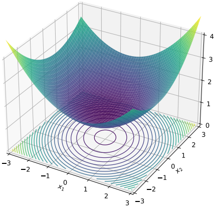



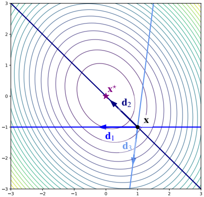
# Choix de la direction de descente

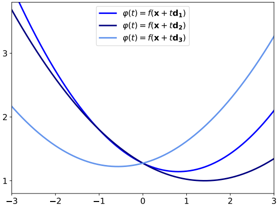
## Restriction de $f$ le long d'un axe

La définition d'une direction de descente $\mathbf{d}$ à partir d'un point $\mathbf{x}$ donné nécessite d'évaluer le comportement de $f$ lorsqu'on évolue "proche" de $\mathbf{x}$, selon l'axe engendré par $\mathbf{d}$ :

- $\rightarrow$ Définition (paramétrique) de la droite $\mathcal{D}_{\mathbf{x},\mathbf{d}}$ passant par $\mathbf{x}$ et de vecteur directeur $\mathbf{d}$ : $\mathcal{D}_{\mathbf{x},\mathbf{d}} = \{x + t\mathbf{d}, t \in \mathbb{R}\}$.
- $\rightarrow$ Définition de la fonction $\varphi : t \mapsto f(\mathbf{x} + t\mathbf{d})$ comme restriction de $f$ à $\mathcal{D}_{\mathbf{x},\mathbf{d}}$.



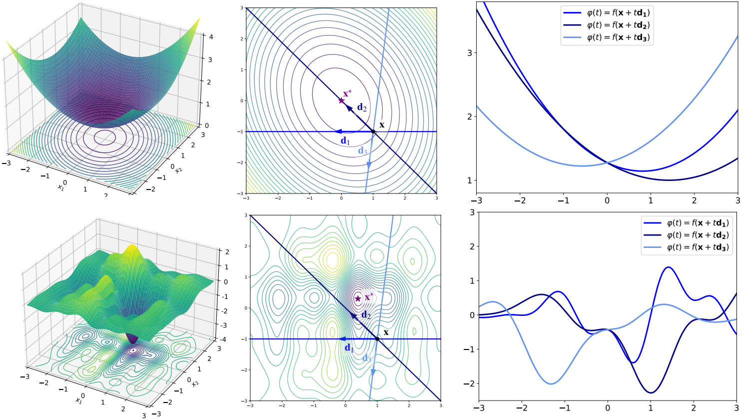



# Choix de la direction de descente

## Restriction de $f$ le long d'un axe

La définition d'une direction de descente $\mathbf{d}$ à partir d'un point $\mathbf{x}$ donné nécessite d'évaluer le comportement de $f$ lorsqu'on évolue "proche" de $\mathbf{x}$, selon l'axe engendré par $\mathbf{d}$ :

- $\rightarrow$ Définition (paramétrique) de la droite $\mathcal{D}_{\mathbf{x},\mathbf{d}}$ passant par $\mathbf{x}$ et de vecteur directeur $\mathbf{d}$ : $\mathcal{D}_{\mathbf{x},\mathbf{d}} = \{x + t\mathbf{d}, t \in \mathbb{R}\}$.
- $\rightarrow$ Définition de la fonction $\varphi : t \mapsto f(\mathbf{x} + t\mathbf{d})$ comme restriction de $f$ à $\mathcal{D}_{\mathbf{x},\mathbf{d}}$.



### Dérivée directionnelle

On appelle *dérivée directionnelle* de $f$ en $\mathbf{x}$ selon le vecteur $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ la dérivée en 0, si elle existe, de la fonction $\varphi(t) = f(\mathbf{x} + t\mathbf{d})$ :

$$D_{\mathbf{d}} f(\mathbf{x}) = \varphi'(0) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

### Dérivée directionnelle

On appelle *dérivée directionnelle* de $f$ en $\mathbf{x}$ selon le vecteur $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ la dérivée en 0, si elle existe, de la fonction $\varphi(t) = f(\mathbf{x} + t\mathbf{d})$ :

$$D_{\mathbf{d}} f(\mathbf{x}) = \varphi'(0) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

- $\rightarrow$ si $\|\mathbf{d}\| = 1$, on parle de dérivée dans la direction de $\mathbf{d}$
- $\rightarrow$ si $D_{\mathbf{d}} f(\mathbf{x})$ existe, alors $D_{\alpha\mathbf{d}} f(\mathbf{x})$ existe et $D_{\alpha\mathbf{d}} f(\mathbf{x}) = \alpha D_{\mathbf{d}} f(\mathbf{x})$ pour $\alpha \in \mathbb{R}$
- $\rightarrow$ $D_{\mathbf{e}_i} f(\mathbf{x}) = \dfrac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x})$ : dérivée dans la direction du vecteur de la base canonique $\mathbf{e}_i$

## Dérivée directionnelle

On appelle *dérivée directionnelle* de $f$ en $\mathbf{x}$ selon le vecteur $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ la dérivée en 0, si elle existe, de la fonction $\varphi(t) = f(\mathbf{x} + t\mathbf{d})$ :

$$D_{\mathbf{d}} f(\mathbf{x}) = \varphi'(0) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

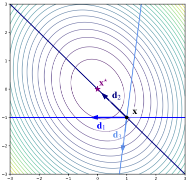
- $\rightarrow$ si $\|\mathbf{d}\| = 1$, on parle de dérivée dans la direction de $\mathbf{d}$
- $\rightarrow$ si $D_{\mathbf{d}} f(\mathbf{x})$ existe, alors $D_{\alpha\mathbf{d}} f(\mathbf{x})$ existe et $D_{\alpha\mathbf{d}} f(\mathbf{x}) = \alpha D_{\mathbf{d}} f(\mathbf{x})$ pour $\alpha \in \mathbb{R}$
- $\rightarrow$ $D_{\mathbf{e}_i} f(\mathbf{x}) = \dfrac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x})$ : dérivée dans la direction du vecteur de la base canonique $\mathbf{e}_i$
- $\rightarrow$ si $f$ est différentiable en $\mathbf{x}$, de différentielle $df_{\mathbf{x}}$, alors $f$ admet une dérivée directionnelle dans toute direction $\mathbf{d}$ et $\boxed{D_{\mathbf{d}} f(\mathbf{x}) = df_{\mathbf{x}}(\mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d}}$

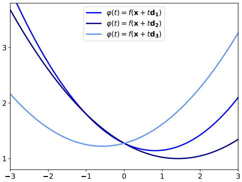
## Dérivée directionnelle

On appelle *dérivée directionnelle* de $f$ en $\mathbf{x}$ selon le vecteur $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ la dérivée en 0, si elle existe, de la fonction $\varphi(t) = f(\mathbf{x} + t\mathbf{d})$ :

$$D_{\mathbf{d}} f(\mathbf{x}) = \varphi'(0) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

- $\rightarrow$ si $\|\mathbf{d}\| = 1$, on parle de dérivée dans la direction de $\mathbf{d}$
- $\rightarrow$ si $D_{\mathbf{d}} f(\mathbf{x})$ existe, alors $D_{\alpha\mathbf{d}} f(\mathbf{x})$ existe et $D_{\alpha\mathbf{d}} f(\mathbf{x}) = \alpha D_{\mathbf{d}} f(\mathbf{x})$ pour $\alpha \in \mathbb{R}$
- $\rightarrow$ $D_{\mathbf{e}_i} f(\mathbf{x}) = \dfrac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x})$ : dérivée dans la direction du vecteur de la base canonique $\mathbf{e}_i$
- $\rightarrow$ si $f$ est différentiable en $\mathbf{x}$, de différentielle $df_{\mathbf{x}}$, alors $f$ admet une dérivée directionnelle dans toute direction $\mathbf{d}$ et $\boxed{D_{\mathbf{d}} f(\mathbf{x}) = df_{\mathbf{x}}(\mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d}}$




$$f(\mathbf{x}) = x_1^2 + x_1 x_2 + x_2^2 \rightarrow \nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 + x_2 \\ 2x_2 + x_1 \end{pmatrix}$$

En $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ selon $\mathbf{d} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ :

$$D_{\mathbf{d}} f(\mathbf{x}) = \nabla f(\mathbf{x})^T \mathbf{d} = -6$$

# Choix de la direction de descente

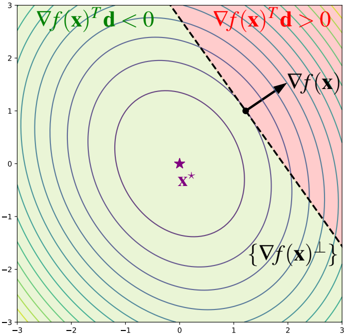
## Direction de descente

On dit que $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ est une *direction de descente* en $\mathbf{x}$ si $\varphi : t \mapsto f(\mathbf{x} + t\mathbf{d})$ est strictement décroissante au voisinage de 0, c'est-à-dire $\varphi'(0) = D_{\mathbf{d}} f(\mathbf{x}) < 0$

## Direction de descente

On dit que $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ est une *direction de descente* en $\mathbf{x}$ si $\varphi : t \mapsto f(\mathbf{x} + t\mathbf{d})$ est strictement décroissante au voisinage de 0, c'est-à-dire $\varphi'(0) = D_{\mathbf{d}} f(\mathbf{x}) < 0$

- $\rightarrow$ Il existe $c > 0$ tel que $\forall t \in [0, c[, \; f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$
- $\rightarrow$ Si $\nabla f(\mathbf{x}) \neq \mathbf{0}$, $\mathbf{d}$ est une direction de descente en $\mathbf{x}$ si et seulement si $\boxed{\nabla f(\mathbf{x})^T \mathbf{d} < 0}$

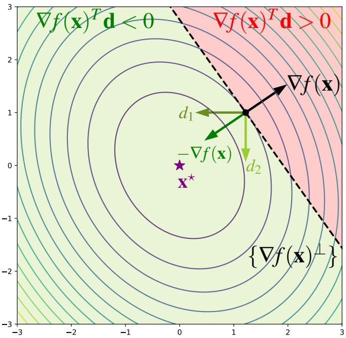


# Choix de la direction de descente

## Direction de descente

On dit que $\mathbf{d} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ est une *direction de descente* en $\mathbf{x}$ si $\varphi : t \mapsto f(\mathbf{x} + t\mathbf{d})$ est strictement décroissante au voisinage de 0, c'est-à-dire $\varphi'(0) = D_{\mathbf{d}} f(\mathbf{x}) < 0$

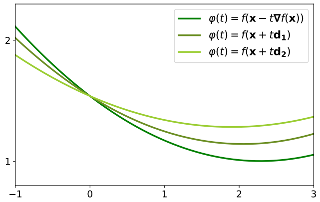
- $\rightarrow$ Il existe $c > 0$ tel que $\forall t \in [0, c[,\ f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$
- $\rightarrow$ Si $\nabla f(\mathbf{x}) \neq \mathbf{0}$, $\mathbf{d}$ est une direction de descente en $\mathbf{x}$ si et seulement si $\boxed{\nabla f(\mathbf{x})^T \mathbf{d} < 0}$




$\mathbf{d} = -\nabla f(\mathbf{x})$ donne bien une direction de descente :

$\rightarrow \nabla f(\mathbf{x})^T \mathbf{d} = \nabla f(\mathbf{x})^T(-\nabla f(\mathbf{x})) = -\|\nabla f(\mathbf{x})\|^2 < 0$

C'est même la direction de plus forte pente !



# Conditions d'optimalité d'un point

L'objectif d'une méthode de descente est de s'approcher itérativement d'un point *optimal* (minimum local)

## Conditions d'optimalité (cas général)

$f$ admet un minimum local en un point $\mathbf{x}^\star$ si

1. $\nabla f(\mathbf{x}^\star) = 0 \Leftrightarrow \mathbf{x}^\star$ est un point critique (condition d'optimalité du 1er ordre)
2. $\mathbf{H}_f(\mathbf{x}^\star) \succ \mathbf{0} \Leftrightarrow$ La hessienne de $f$ en $\mathbf{x}^\star$ est définie positive (condition d'optimalité du 2nd ordre)

## Apport de la convexité

Si $f$ est convexe

$\rightarrow$ $\mathbf{H}_f(\mathbf{x}) \succeq \mathbf{0}$ pour tout point $\mathbf{x} \in \mathbb{R}^n$ (caractérisation à l'ordre 2 de la convexité)

$\rightarrow$ $f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)^T(\mathbf{x} - \mathbf{x}^\star) = f(\mathbf{x}^\star)$ (caractérisation à l'ordre 1 de la convexité)

$\Rightarrow$ $\boxed{\text{Tout point critique } \mathbf{x}^\star \text{ de } f \text{ est un minimum global}}$

# Choix du critère d'arrêt

Le but du critère d'arrêt est de permettre à l'algorithme de descente de s'arrêter en un point $\mathbf{x}_k$ suffisamment proche d'un minimum local (ou global) $\mathbf{x}^\star$

---

**tant que** *critère d'arrêt non satisfait* **faire**
    $\rightarrow$ itération de descente
**fin**

---

Le but du critère d'arrêt est de permettre à l'algorithme de descente de s'arrêter en un point $\mathbf{x}_k$ suffisamment proche d'un minimum local (ou global) $\mathbf{x}^\star$

---

**tant que** *critère d'arrêt non satisfait* **faire**
  $\rightarrow$ itération de descente
**fin**

---

Critère d'arrêt naturel : $\|\nabla f(\mathbf{x}_k)\| \leq \varepsilon$ (condition d'optimalité du 1er ordre)

- 👍 À priori suffisant si $f$ est convexe.
- 👎 Nécessite de tester si la hessienne $\mathbf{H}_f(\mathbf{x}_k)$ est définie positive sinon.

# Choix du critère d'arrêt

Le but du critère d'arrêt est de permettre à l'algorithme de descente de s'arrêter en un point $\mathbf{x}_k$ suffisamment proche d'un minimum local (ou global) $\mathbf{x}^\star$

---

**tant que** *critère d'arrêt non satisfait* **faire**
  $\rightarrow$ itération de descente
**fin**

---

Critère d'arrêt naturel : $\|\nabla f(\mathbf{x}_k)\| \leq \varepsilon$ (condition d'optimalité du 1er ordre)

- 👍 À priori suffisant si $f$ est convexe.
- 👎 Nécessite de tester si la hessienne $\mathbf{H}_f(\mathbf{x}_k)$ est définie positive sinon.

Autres critères d'arrêt possibles :

- $\rightarrow$ stagnation relative de l'itéré : $\dfrac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|}{\|\mathbf{x}_k\|} \leq \varepsilon$
- $\rightarrow$ stagnation relative de la valeur courante : $\dfrac{|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|}{|f(\mathbf{x}_k)|} \leq \varepsilon$
- $\rightarrow$ limite du nombre d'itérations : $k < \textsc{MaxIter}$

# Choix du critère d'arrêt

Le but du critère d'arrêt est de permettre à l'algorithme de descente de s'arrêter en un point $\mathbf{x}_k$ suffisamment proche d'un minimum local (ou global) $\mathbf{x}^\star$

**tant que** *critère d'arrêt non satisfait* **faire**
  $\rightarrow$ itération de descente
**fin**

Critère d'arrêt naturel : $\|\nabla f(\mathbf{x}_k)\| \leq \varepsilon$ (condition d'optimalité du 1er ordre)

👍 À priori suffisant si $f$ est convexe.

👎 Nécessite de tester si la hessienne $\mathbf{H}_f(\mathbf{x}_k)$ est définie positive sinon.

Autres critères d'arrêt possibles :

- $\rightarrow$ stagnation relative de l'itéré : $\dfrac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|}{\|\mathbf{x}_k\|} \leq \varepsilon$
- $\rightarrow$ stagnation relative de la valeur courante : $\dfrac{|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|}{|f(\mathbf{x}_k)|} \leq \varepsilon$
- $\rightarrow$ limite du nombre d'itérations : $k < \textsc{MaxIter}$

En général : limite du nombre d'itérations + test d'optimalité ou stagnation (*early stopping*)

$\Rightarrow$ introduction d'hyperparamètres supplémentaires $\textsc{MaxIter}$ et tolérance $\varepsilon$...

# Descente de gradient à pas constant

**Algorithme :** Descente de gradient à pas constant

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, pas de descente $\eta$

**tant que** *critère d'arrêt non satisfait* **faire**

$$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$$

$$\eta_k = \eta$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k$$

**fin**

---

**Algorithme :** Descente de gradient à pas optimal

---

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$

**tant que** *critère d'arrêt non satisfait* **faire**

$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$

$\eta_k = \min_{\eta \in \mathbb{R}_*^+} f(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k))$

$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k$

**fin**

---

# Descente avec critère d'Armijo

**Algorithme :** Descente avec critère d'Armijo

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, $\alpha \in ]0, \frac{1}{2}[$, $\beta \in ]0, \frac{1}{2}[$

**tant que** *critère d'arrêt non satisfait* **faire**

  $\mathbf{d}_k =$ direction de descente  *pas nécessairement* $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$

  $\eta = 1$

  **tant que** $f(\mathbf{x}_k + \eta \mathbf{d}_k) > f(\mathbf{x}_k) + \alpha\eta\nabla f(\mathbf{x}_k)^T \mathbf{d}_k$ **faire**  *condition d'Armijo non respectée*

    $\eta = \beta\eta$

  **fin**

  $\eta_k = \eta$

  $\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k$

**fin**

→ Si $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, on parle de descente de gradient avec critère d'Armijo.

On utilise typiquement $0.1 \leq \alpha \leq 0.3$ et $0.2 \leq \beta \leq 0.8$

**Algorithme :** Descente avec *momentum*

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, $\alpha \in ]0, 1[$, pas de descente $\eta$

$\mathbf{v}_0 = 0$

**tant que** *critère d'arrêt non satisfait* **faire**

$\mathbf{v}_{k+1} = \alpha \mathbf{v}_k - \eta \nabla f(\mathbf{x}_k)$

$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}$

**fin**

**Algorithme :** Accélération de Nesterov

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, $\alpha \in ]0, 1[$, pas de descente $\eta$

$\mathbf{v}_0 = 0$

**tant que** *critère d'arrêt non satisfait* **faire**

$\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k$

$\mathbf{v}_{k+1} = \alpha \mathbf{v}_k - \eta \nabla f(\mathbf{y}_k)$

$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}$

**fin**

# Algorithme du gradient conjugué
## Cas quadratique

Dans le cas où $f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ avec $\mathbf{A} \in \mathbb{R}^{n\times n}$ symétrique définie positive

**Algorithme :** Gradient conjugué - cas quadratique

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$

**pour** $k = 0, \dots, n-1$ **faire**

- $\mathbf{g}_k = \nabla f(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b}$ *gradient au point actuel*
- **si** $k = 0$ **alors**
  - $\mathbf{d}_k = -\mathbf{g}_k$ *direction de descente initiale*
- **fin**
- **sinon**
  - $\beta_k = \dfrac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_{k-1}{}^T\mathbf{g}_{k-1}}$ *coefficient pour* $\mathbf{A}-$*conjuguer la nouvelle direction de descente à la précédente*
  - $\mathbf{d}_k = -\mathbf{g}_k + \beta_k\mathbf{d}_{k-1}$ *nouvelle direction de descente* $\mathbf{A}-$*conjuguée à la précédente* ($\mathbf{d}_k^T\mathbf{A}\mathbf{d}_{k-1} = 0$)
- **fin**
- $\alpha_k = \dfrac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}$ *pas de descente optimal dans le cas quadratique*
- $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$

**fin**

**Algorithme :** Gradient conjugué - cas général

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, $\alpha \in ]0, \frac{1}{2}[$, $\beta \in ]0, \frac{1}{2}[$

**tant que** *critère d'arrêt non satisfait* **faire**

$\mathbf{g}_k = \nabla f(\mathbf{x}_k)$ — *gradient au point actuel*

**si** $k = 0$ **alors**

$\mathbf{d}_k = -\mathbf{g}_k$ — *direction de descente initiale*

**fin**

**sinon**

$$\beta_k = \begin{cases} \dfrac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_{k-1}{}^T \mathbf{g}_{k-1}} & \textit{méthode de Fletcher-Reeves} \\ & \textit{ou} \\ \dfrac{(\mathbf{g}_k - \mathbf{g}_{k-1})^T \mathbf{g}_k}{\mathbf{g}_{k-1}{}^T \mathbf{g}_{k-1}} & \textit{méthode de Polack-Ribière} \end{cases}$$

$\mathbf{d}_k = -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}$

**fin**

$\alpha_k = \textsc{Armijo}(\mathbf{x}_k, \alpha, \beta)$ — *pas selon la méthode d'Armijo*

$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$

**fin**

**Algorithme :** Méthode de Newton

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$

**tant que** *critère d'arrêt non satisfait* **faire**

$$\mathbf{d}_k = -\mathbf{H}_f(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$$ *pas de Newton*

$$\lambda(\mathbf{x}_k) = \sqrt{\mathbf{d}_k^T \mathbf{H}_f(\mathbf{x}_k)\mathbf{d}_k} = \sqrt{\nabla f(\mathbf{x}_k)^T \mathbf{H}_f(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)}$$ *décrément de Newton*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$$

**fin**

Le critère d'arrêt de la méthode de Newton fait en général intervenir le *décrément* de Newton :

$\rightarrow$ stagnation du décrément de Newton: $\lambda(\mathbf{x}_k)^2 \leq \varepsilon$

# Méthode de quasi-Newton

**Algorithme :** Méthode de quasi-Newton (Broyden-Fletcher-Goldfarb-Shanno)

**Entrée :** Point de départ $\mathbf{x}_0 \in \mathbb{R}^n$, $\alpha \in ]0, \frac{1}{2}[$, $\beta \in ]0, \frac{1}{2}[$

$\mathbf{H}_0 = I_n$

**tant que** *critère d'arrêt non satisfait* **faire**

$\mathbf{d}_k = -\mathbf{H}_k \nabla f(\mathbf{x}_k)$ *approximation du pas de Newton*

$\eta_k = \textsc{Armijo}(\mathbf{x}_k, \alpha, \beta)$ *pas selon la méthode d'Armijo*

$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{d}_k$

$\boldsymbol{\delta}\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$

$\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$

$\mathbf{H}_{k+1} = \left(I_n - \frac{\boldsymbol{\delta}\mathbf{x}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \boldsymbol{\delta}\mathbf{x}_k}\right)\mathbf{H}_k\left(I_n - \frac{\mathbf{y}_k \boldsymbol{\delta}\mathbf{x}_k^T}{\mathbf{y}_k^T \boldsymbol{\delta}\mathbf{x}_k}\right) + \frac{\boldsymbol{\delta}\mathbf{x}_k \boldsymbol{\delta}\mathbf{x}_k^T}{\mathbf{y}_k^T \boldsymbol{\delta}\mathbf{x}_k}$ *mise à jour de l'approximation de* $\mathbf{H}_f(\mathbf{x}_{k+1})^{-1}$

**fin**