OUTILS D'OPTIMISATION POUR LES SCIENCES DES DONNÉES ET DE LA DÉCISION

20 octobre 2023

Aujourd'hui : Accélération (cours + TD avec notebook d'illustration)

NB : Polycopié mis à jour avec le contenu des séances 4-6

OPTIMISATION CONVEXE ET ACCÉLÉRATION

Cadre

minimiser $f(w)$ $w \in \mathbb{R}^d$

$f : \mathbb{R}^d \to \mathbb{R}$ C^1_L et convexe

∇f existe en tout w

∇f est lipschitzien

∇f est L-lipschitz

$\|\nabla f(v) - \nabla f(w)\| \leq L\|v - w\|$

f convexe :

- Tout minimum local est global
- $\|\nabla f(\bar{w})\| = 0 \iff \bar{w}$ est un minimum global

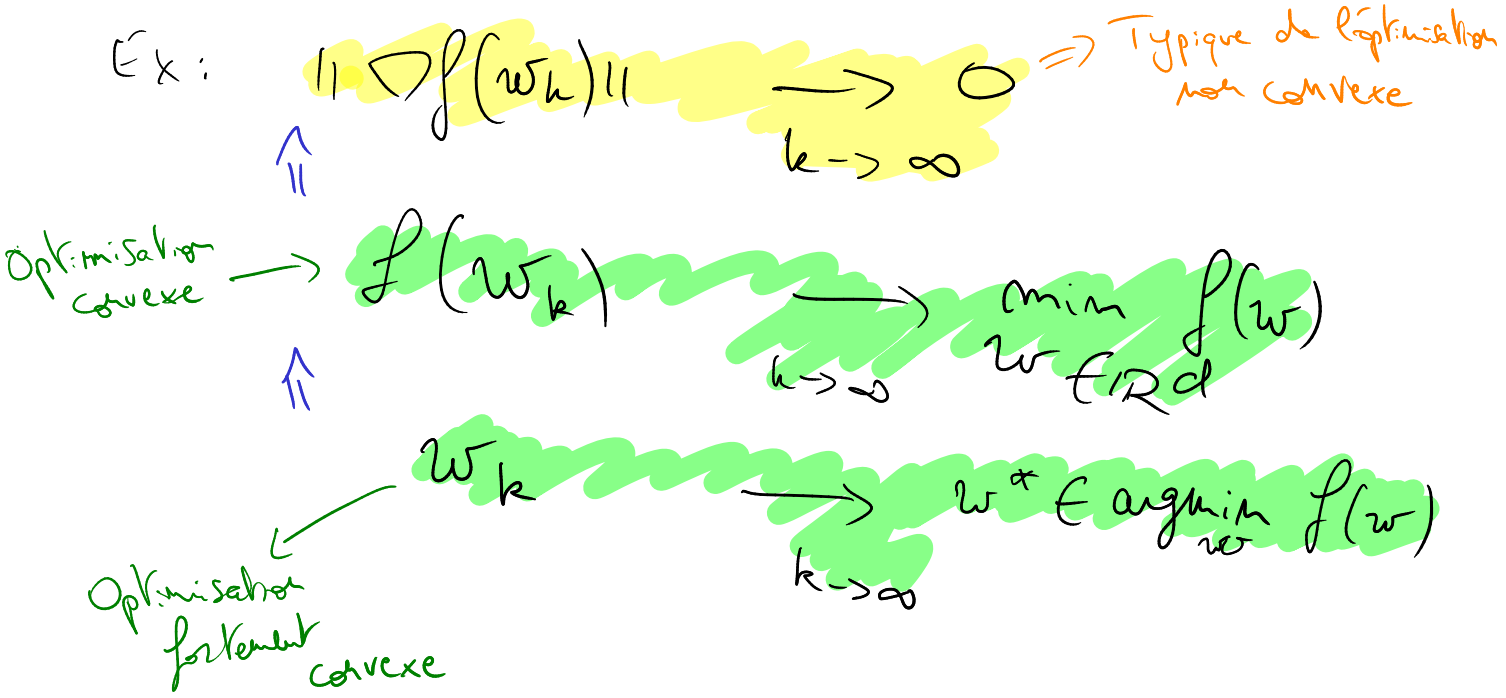
Objectif : Montrer la convergence de l'algorithme de descente de gradient vers une solution du problème

Rappel : Itération de la descente de gradient

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k)$$

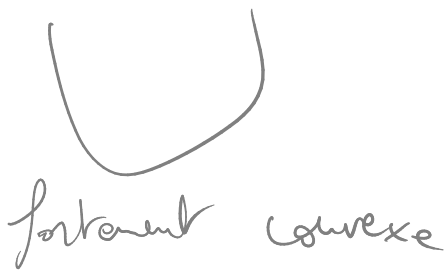
$\alpha_k > 0$

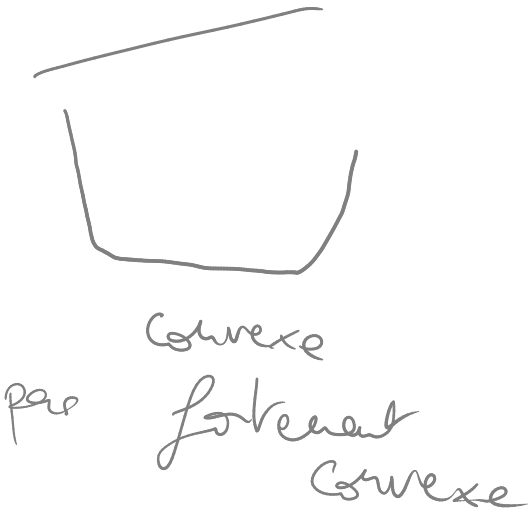
↳ Analyse classique en optimisation : Garanties asymptotiques



Rappel : f est μ-fortement convexe

$\Rightarrow f$ possède un unique minimum global


fortement convexe


convexe mais pas fortement convexe

↳ En pratique, on dispose d'un temps fini (ex : nombre max d'itérations)

et on recherche donc des garanties dites non-asymptotiques

- **Complexité** : Étant donné $\varepsilon > 0$, combien de temps pour atteindre la précision ε ?

Ex) Quel est le nombre d'itérations de descente de gradient nécessaires pour obtenir $f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \varepsilon$?

⇒ But : Obtenir un nombre d'itérations qui est une fonction de ε

- **Vitesse de convergence** : Étant donné un budget, quelle précision puis-je espérer atteindre ?

Ex) Si on effectue K itérations de descente de gradient, quelle est la valeur au pire cas de $f(w_K) - \min_{w \in \mathbb{R}^d} f(w)$?

⇒ Fonction de K

Remarque : On peut toujours transformer une vitesse de convergence en complexité, et vice-versa

(1) Résultats pour la descente de gradient

Cas convexe : f convexe qui atteint son minimum

On note $f^* = \min_{w \in \mathbb{R}^d} f(w)$

Théorème (Vitesse de convergence de la descente de gradient sur problèmes convexes)

Supposons que l'on effectue K itérations de descente de gradient avec $K \geq 1$ et $\alpha_k = \frac{1}{L}$ (>0). Alors

$$f(w_K) - f^* \leq \frac{C}{K}$$

où $C > 0$ est une constante qui dépend de w_0, L mais pas de K

On dit que la descente de gradient a une vitesse de convergence en $O\left(\frac{1}{K}\right)$

N.B. : Le résultat reste vrai pour d'autres choix de α_k (ex. recherche linéaire)

Interprétation

- Garantie au pire cad sur la valeur de $f(w_k) - f^*$ ("distance à la valeur optimale")
- Implique la convergence asymptotique
$0 \leq f(w_k) - f^* \leq O\left(\frac{1}{k}\right)$ $\Rightarrow f(w_k) - f^* \to 0$ $k \to \infty$

$f(w_k) \underset{k \to \infty}{\to} f^*$

- Garantie plus précise qui relie la nombre d'itérations à la précision

$f(w_k) - f^* \leq O\left(\frac{1}{k}\right)$ pour k fixé

Alors avec $10k$ itérations, $f(w_{10k}) - f^* \leq O\left(\frac{1}{10k}\right)$ équivalent.

- Résultat de complexité

Soit $\varepsilon > 0$. Alors on a $f(w_k) - f^* \leq \varepsilon$ en au plus

$$O\left(\frac{1}{\varepsilon}\right)$$ itérations

Cas fortement convexe

f $C^{1,1}_L$, μ-fortement convexe avec $\mu > 0$

On note $\{w^*\} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} f(w)$

et $f^* = f(w^*)$

NB: On a toujours $\mu \leq L$.

Théorème (Vitesse de convergence de la descente de gradient sur les fonctions convexes)

Supposons qu'on effectue K itérations de descente de gradient avec $\alpha_k = \frac{1}{L}$

Alors

$$f(w_K) - f^* \leq C \left(1 - \frac{\mu}{L}\right)^K$$

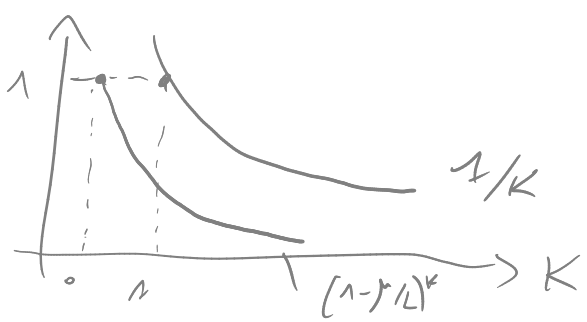
avec $C = f(w_0) - f^* \geq 0$

On dit que la descente de gradient converge en vitesse $O\left(\left(1 - \frac{\mu}{L}\right)^K\right)$

(vitesse exponentielle en K $\left(1 - \frac{\mu}{L}\right)^K = e^{K \ln\left(1 - \frac{\mu}{L}\right)}$)

Interprétation

- Meilleure vitesse de convergence que dans le cas convexe !



$0 < \mu \leq L \Rightarrow 1 - \frac{\mu}{L} < 1$

$\left(1 - \frac{\mu}{L}\right)^K \to 0$ plus vite que $\frac{1}{K}$

- La vitesse de convergence dépend des propriétés de f (L et μ)
- Implique la convergence asymptotique
- On peut aussi montrer que

$$\|w_k - w^*\| \leq O\left(\left(1-\frac{\mu}{L}\right)^k\right)$$

⇒ Convergence à la fois vers la valeur optimale et vers la solution optimale ! (Plus fort que les garanties dans le cas convexe)

(2) Accélération

↳ Que ce soit dans le cas convexe ou fortement convexe, les vitesses de convergence de la descente de gradient ne sont pas optimales ⇒ il existe des algorithmes qui calculent un gradient par itération et obtiennent de meilleures vitesses

	Descente de gradient	Optimal
Convexe	$O\left(\frac{1}{k}\right)$	$O\left(\frac{1}{k^2}\right)$
Fortement convexe	$O\left(\left(1-\frac{\mu}{L}\right)^k\right)$	$O\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^k\right)$

↳ les algorithmes avec vitesse de convergence optimale sont basés sur l'idée du momentum

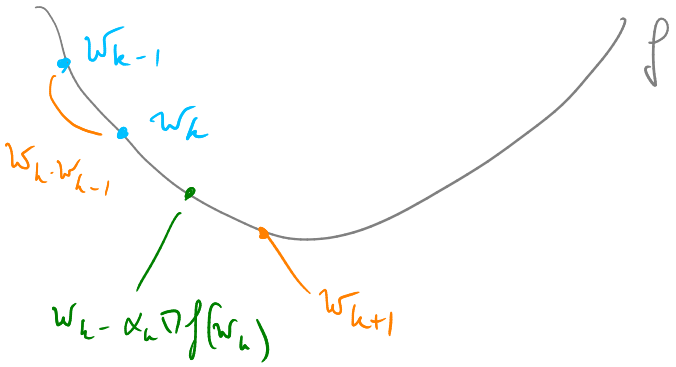
Première proposition : Heavy ball ("boule lestée") Polyak (~1964)

$\forall k \geq 1$

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k) + \beta_k (w_k - w_{k-1})$$

β_k : Paramètre de momentum

$(w_k - w_{k-1})$: Terme de momentum



- Polyak prouve la vitesse de convergence optimale pour f quadratique fortement convexe
- Mais cet algorithme peut ne pas converger sur des fonctions générales

Seconde proposition : Gradient accéléré (Y. Nesterov 1983)

$$w_{k+1} = w_k - \alpha_k \nabla f\left(w_k + \beta_k (w_k - w_{k-1})\right) + \beta_k (w_k - w_{k-1})$$

↳ ≠ avec Polyak : le gradient est évalué après le pas de momentum

↳ Atteint les bornes optimales de complexité

- pour f fortement convexe $\left(1-\sqrt{\frac{\mu}{L}}\right)^k$ avec $\beta_k = \beta > 0$ β dépend de L et μ
- pour f convexe $\frac{1}{k^2}$ avec $\{\beta_k\}$ choisi indépendamment de f et de x_0 (et de α_k)

NB: le gradient accéléré au car proximal peut se généraliser (ex: ISTA $\Rightarrow$ FAST ISTA)