Optimisation pour l'apprentissage automatique

11 décembre 2024

Séance 2/8 : Théorie descente de gradient

* Régression linéaire
* Cas général : vitesses de convergence / complexités
* Un résultat important dans le cas non convexe

Séance 3/8 (demain)

* TP descente de gradient (cas convexe)
* Exercices descente de gradient

Théorie de la descente de gradient

Cadre général

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \; f(w) \qquad f : \mathbb{R}^d \to \mathbb{R} \; C^1$$

Algorithme de descente de gradient

$$\forall k \in \mathbb{N}, \quad w_{k+1} = w_k - \alpha_k \nabla f(w_k) \quad \text{avec} \quad \alpha_k > 0$$

Deux types de résultats théoriques

- Convergence asymptotique : garanties à la limite, quand $k \to \infty$

  Ex) $f(w_k) \xrightarrow[k \to \infty]{} \min_{w \in \mathbb{R}^d} f(w)$

  ↑ Valeur optimale du problème

- Convergence non-asymptotique / vitesse de convergence / complexité : Garanties en "temps" fini

  Ex) Après K itérations, que peut-on dire de $f(w_K)$? De $f(w_K) - \min_{w \in \mathbb{R}^d} f(w)$?

Cadre le plus simple : Descente de gradient pour la régression linéaire

Problème :

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \; \frac{1}{2n} \|Xw - y\|^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(x_i^T w - y_i \right)^2$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

Soit $f(w) = \frac{1}{2n}\|Xw - y\|^2$ et supposons que $\underbrace{\underbrace{X^T}_{d\times n}\underbrace{X}_{n\times d}}_{d\times d}$ est inversible

Sous cette hypothèse, le problème a une unique solution donnée par $w^* = (X^TX)^{-1}X^Ty$.

En effet, $\nabla f(w) = 0 \iff \frac{1}{n}X^T(Xw - y) = 0$

$\iff X^TXw - X^Ty = 0$

$\iff X^TXw = X^Ty$

X^TX inversible $\rightarrow \iff w = (X^TX)^{-1}X^Ty = w^*$

Donc w^* est l'unique point en lequel le gradient est nul.

On f est une fonction convexe, donc $\nabla f(w^*) = 0 \iff w^* \in \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} f(w)$

On applique la descente de gradient au problème avec une taille de pas constante $\alpha_k = \frac{n}{\|X^TX\|} > 0$

$\|X^TX\| = \max_{\|v\|\neq 0} \frac{\|X^TXv\|}{\|v\|}$

$= \lambda_{\max}(X^TX)$ plus grande valeur de X^TX

λ valeur propre de X^TX $\iff \exists v \neq 0,\ X^TXv = \lambda v$

Pour tout $k \in \mathbb{N}$, on a alors

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

$= w_k - \frac{n}{\|X^TX\|} \times \frac{1}{n}X^T(Xw_k - y)$

$= w_k - \frac{1}{\|X^TX\|}\left(X^TXw_k - X^Ty\right)$

$w_{k+1} = \left(I_d - \frac{X^TX}{\|X^TX\|}\right)w_k + \frac{1}{\|X^TX\|}X^Ty$

$I_d = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}$

$I_d w_k = w_k$

D'où

$$w_{k+1} - w^* = \left(Id - \frac{X^TX}{\|X^TX\|}\right) w_k + \frac{1}{\|X^TX\|} \underbrace{X^Ty}_{X^TXw^*} - w^*$$

$$= \left(Id - \frac{X^TX}{\|X^TX\|}\right) w_k + \left(\frac{X^TX}{\|X^TX\|} - Id\right) w^*$$

$$= \left(Id - \frac{X^TX}{\|X^TX\|}\right)(w_k - w^*)$$

Pour montrer que l'algorithme converge, il faut montrer que $w_k \to w^*$, c'est-à-dire $\|w_k - w^*\| \to 0$

D'après ce qui précède, on a

$$\|w_{k+1} - w^*\| = \left\| \left(Id - \frac{X^TX}{\|X^TX\|}\right)(w_k - w^*) \right\|$$

$$\leq \left\| Id - \frac{X^TX}{\|X^TX\|} \right\| \|w_k - w^*\|$$

$$= \underbrace{\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)}_{\in (0,1)} \|w_k - w^*\|$$

plus petite valeur propre de X^TX (>0 car X^TX inversible)

$$\|w_{k+1} - w^*\| < \|w_k - w^*\|$$

A chaque itération, l'algorithme se rapproche de la solution.

$\forall K \in \mathbb{N}$,

$$\|w_K - w^*\| \leq \left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)^K \|w_0 - w^*\|$$

(application de $\|w_{k+1} - w^*\| \leq \left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right) \|w_k - w^*\|$ pour $k = 0 \dots K-1$)

Ce résultat s'appelle une vitesse de convergence : on dit que la descente de gradient converge linéairement (ou en vitesse linéaire) sur ce problème

Linéaire : $\ln\left(\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)^K \|w_0 - w^*\| \right) = K \ln\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right) + \ln(\|w_0 - w^*\|)$

Fonction linéaire/affine de K

→ Pour un nombre fixé d'itérations K, on a des garanties sur $0 \leq \|w_K - w^*\| \quad \left(\leq \left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)^K \|w_0 - w^*\| \right)$

→ Cela implique que $\|w_K - w^*\| \xrightarrow[K \to \infty]{} 0$ car $\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)^K \to 0$

→ Cette vitesse de convergence se traduit aussi en (borne de) complexité :

Pour tout $\varepsilon > 0$, on peut garantir $\|w_K - w^*\| \leq \varepsilon$

· en un nombre d'itérations de l'ordre de $|\ln \varepsilon|$

* Pour avoir $\|w_K - w^*\| \leq \varepsilon$, il suffit que

$$\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)^K \|w_0 - w^*\| \leq \varepsilon$$

$$K \underbrace{\ln\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)}_{<0} + \ln\left(\|w_0 - w^*\|\right) \leq \ln \varepsilon$$

$$K \geq \frac{\ln \varepsilon}{\ln\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)} - \frac{\ln\left(\|w_0 - w^*\|\right)}{\ln\left(1 - \frac{\lambda_{min}(X^TX)}{\lambda_{max}(X^TX)}\right)}$$

$$= O\left(|\ln \varepsilon|\right)$$

$O(A)$: une constante $\times A$, et la constante ne dépend pas des quantités qui apparaissent dans A

Choix de ε

- Typiquement dans $(0,1)$
- Peut aussi être "normalisé" par rapport au problème

$$\varepsilon = \eta \|2v_0 - v^*\|$$

$$\eta \in (0,1)$$

Analyse pour des classes plus générales de fonctions

↳ Les résultats ci-dessus pour la régression linéaire s'étendent aux fonctions $C^{1,1}$ et fortement convexes

Def : • $f : \mathbb{R}^d \to \mathbb{R}$ est $C^{1,1}_L$ si elle est C^1 et ∇f est L-lipschitzien avec $L > 0$, càd

$$\forall (v,w) \in (\mathbb{R}^d)^2, \quad \|\nabla f(v) - \nabla f(w)\| \leq L \|v - w\|$$

On dit parfois que f est L-lisse ("L-smooth")

Si f est $C^{1,1}_L$, alors pour tous $(v,w) \in (\mathbb{R}^d)^2$,

$$(1) \quad f(v) \leq f(w) + \nabla f(w)^T (v - w) + \frac{L}{2} \|v - w\|^2$$

(Version exacte de $f(v) \approx f(w) + \nabla f(w)^T (v - w)$)

Conséquence : Si $w = w_k$ et $v = w_k - \alpha_k \nabla f(w_k)$

- $f\left(w_k - \alpha_k \nabla f(w_k)\right) < f(w_k)$ pour $0 < \alpha_k < \frac{2}{L}$
- $f\left(w_k - \frac{1}{L} \nabla f(w_k)\right) \leq f(w_k) - \frac{1}{2L} \|\nabla f(w_k)\|^2$, et

c'est la plus forte décroissance pour $\alpha_k \in \left(0, \frac{2}{L}\right)$ $\Rightarrow$ Justifie de choisir $\alpha_k = \frac{1}{L}$ $\forall k$

Ex) Régression linéaire

$$f(w) = \frac{1}{2n} \|Xw - y\|^2 \qquad \nabla f(w) = \frac{1}{n} X^T (Xw - y)$$

f est $C^{1,1}_L$ avec $L = \frac{\|X^T X\|}{n}$

Déf : $f : \mathbb{R}^d \to \mathbb{R}$ C^1 est μ-fortement convexe si : $(\mu > 0)$

(2) $\forall (v, w) \in (\mathbb{R}^d)^2$, $f(v) \geq f(w) + \nabla f(w)^T (v - w) + \frac{\mu}{2} \|v - w\|^2$

(2) fournit un minorant de $f(v)$

→ Toute fonction fortement convexe est convexe

$$\left[f \ C^1 \text{ convexe}\right] \iff \forall (v, w) \in (\mathbb{R}^d)^2, \ f(v) \geq f(w) + \nabla f(w)^T (v - w)$$

mais l'inverse n'est pas vrai (contre ex : fonction cubique)

→ Toute fonction fortement convexe f possède un unique minimum, qui est l'unique vecteur w tel que $\nabla f(w) = 0$

↑ argmin $f(w)$ $w \in \mathbb{R}^d$ contient un seul vecteur

Ex) Régression linéaire : $f(w) = \frac{1}{2n}\|Xw - y\|^2$

Si X^TX est inversible, f est μ-fortement convexe avec

$$\mu = \frac{\lambda_{min}(X^TX)}{n}$$

Sinon, f est simplement convexe et le problème a une infinité de solutions.

Vitesses de convergence / Complexité dans le cas fortement convexe

Soit $f \in C^{1,1}_L$ et μ-fortement convexe.

On applique la descente de gradient avec $\alpha_k = \frac{1}{L}$.

Alors, si w^* est le minimum de f, on a :

$$\forall k \in \mathbb{N}, \quad \|w_k - w^*\| \leq C\left(1 - \frac{\mu}{L}\right)^k \|w_0 - w^*\|$$

avec $C > 0$

$\Rightarrow$ vitesse de convergence linéaire vers la solution

et $$f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(w_0) - \min_{w \in \mathbb{R}^d} f(w)\right)$$

$\Rightarrow$ vitesse de convergence linéaire vers la valeur optimale

Vitesse de convergence de la descente de gradient dans le cas fortement convexe linéaire en $O(t^k)$

avec $t = 1 - \frac{\mu}{L} \in (0,1)$

- Pour tout $\varepsilon > 0$, on a $f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \varepsilon$ après au plus $O(|\ln \varepsilon|)$ itérations

(idem pour $\|w_k - w^*\| \leq \varepsilon$)

$\rightarrow$ Se généralise au choix d'un α_k constant égal à $\alpha \in (0, \frac{2}{L})$

Comparaison avec le cas non fortement convexe

f $C^{1,1}_L$, f admet un minimum ($\operatorname{argmin}_{w \in \mathbb{R}^d} f(w) \neq \emptyset$) et on note

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

Descente de gradient avec $\alpha_k = 1/L$

f μ-fortement convexe	$f(w_k) - f^* \leq O\left((1 - \frac{\mu}{L})^k\right)$	$f(w_k) - f^* \leq \varepsilon$ en $O(\log(\frac{1}{\varepsilon}))$ itérations
f convexe	$f(w_k) - f^* \leq O\left(\frac{1}{k}\right)$	$f(w_k) - f^* \leq \varepsilon$ en $O\left(\frac{1}{\varepsilon}\right)$ itérations
f non convexe	$\min_{0 \leq k \leq K-1} \|\nabla f(w_k)\| \leq O\left(\frac{1}{\sqrt{K}}\right)$	$\|\nabla f(w_k)\| \leq \varepsilon$ en $O\left(\frac{1}{\varepsilon^2}\right)$ itérations

Analyse de ces résultats

- Le cas convexe est plus difficile que le cas fortement convexe, dans le sens où la descente de gradient converge moins vite dans le cas convexe (en vitesse sous-linéaire $\frac{1}{k}$)

Illustration : $\frac{\mu}{L} = \frac{1}{2}$, $K = 100$ $\quad \frac{1}{k} = 0.01 = 10^{-2}$

$$\left(1 - \frac{\mu}{L}\right)^K = \frac{1}{2^{100}} \approx 10^{-30}$$

$\Rightarrow$ Dans les deux cas cependant, on montre que $f(w_k) - f^* \to 0$

$\Rightarrow$ Dans le cas fortement convexe, comme le minimum est unique, on peut en plus garantir que w_k converge vers ce minimum

- Le cas non convexe est plus difficile que le cas convexe, et donne des garanties plus faibles :

- En termes de vitesse de convergence

$\frac{1}{\sqrt{k}}$ tend vers 0 moins vite que $\frac{1}{k}$

$k = 100 \quad \frac{1}{k} = 10^{-2} \quad \frac{1}{\sqrt{k}} = 10^{-1}$

- En termes de quantité à laquelle s'applique cette vitesse

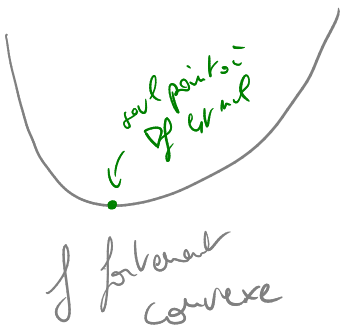
$f(w_k) - f^*$ pour f convexe / fortement convexe

$\min_{0 \le k \le K-1} \|\nabla f(w_k)\|$ pour f non convexe

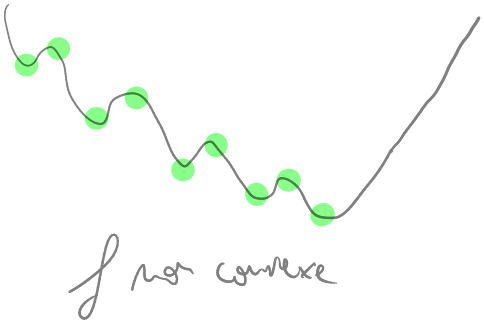
$\Rightarrow$ Cette différence s'explique par le fait que $\|\nabla f(\bar{w})\| = 0 \Leftrightarrow \bar{w} \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$ lorsque f est convexe, alors que on a seulement

$\left[\bar{w} \in \operatorname{argmin}_{w \in \mathbb{R}^d} f(w) \right] \Rightarrow \|\nabla f(\bar{w})\| = 0$

pour f non convexe en général.

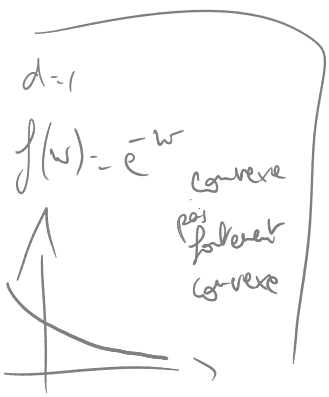


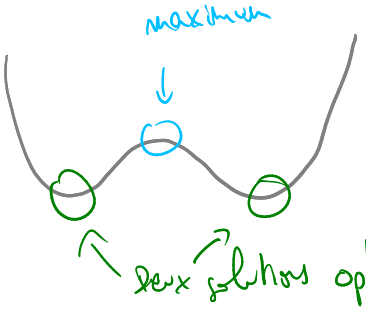
f fortement convexe



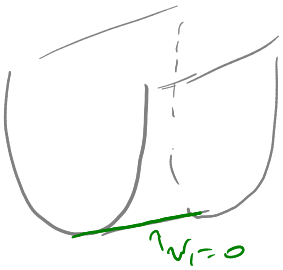
f non convexe

points où le gradient s'annule





f non convexe



$f\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}\right) = w_1^2$

convexe

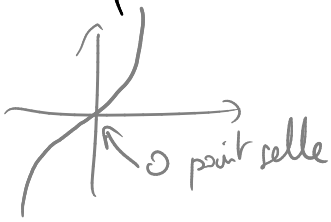
pas fortement convexe

→ La descente de gradient converge vers un point en lequel le gradient est nul.

Si la fonction est non convexe, alors un tel point peut être :

- Un minimum global (donc une solution de $\min_{w \in \mathbb{R}^d} f(w)$)
- Un minimum local non global ($f(\bar{w}) \leq f(w)$ $\forall w$ tel que $\|w - \bar{w}\|$ est faible)
- Un maximum local / global
- Un point selle, càd un point qui est un maximum local dans certaines directions et un minimum local pour d'autres

$d=1$, $f(w) = w^3$



→ On peut trouver des exemples de fonctions non convexes pour lesquelles la descente de gradient converge vers un maximum ou un point selle

→ Mais en pratique, et notamment sur des fonctions non convexes d'apprentissage (complétion de matrice, réseaux de neurones peu profonds, apprentissage de dictionnaire, ...), on observe que la descente de gradient converge vers un minimum local voire global.

(cf exercices)

Théorème (informel, 2015)

Si on applique la descente de gradient avec $w_0 \in \mathbb{R}^d$ choisi aléatoirement dans $\mathbb{R}^d$, alors on converge presque sûrement vers un minimum local.

$\mathbb{P}_{w_0}$ (on converge vers un maximum ou un point selle) $= 0$