OUTILS D'OPTIMISATION LES SDD

28 octobre 2024

Aujourd'hui :

CM : Optimisation — mon convexe / Projet convexe

TD/TP : Optim

OPTIMISATION NON CONVEXE ET COURBURE NÉGATIVE

① Cadre de travail

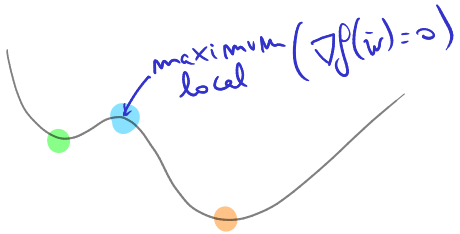
minimiser $f(w)$ avec $w \in \mathbb{R}^d$

$f : \mathbb{R}^d \to \mathbb{R}$ non convexe

↳ Difficultés en optimisation non convexe (par comparaison avec le cas convexe)

a) Distinction entre minima locaux et globaux

Ex)



Une fonction non convexe

b) Un point $\tilde{w} \in \mathbb{R}^d$ pour lequel $\|\nabla f(\tilde{w})\| = 0$ n'est pas forcément un minimum

c) Si on applique la descente de gradient à un problème non convexe, on ne peut garantir qu'une convergence vers un point en lequel le gradient est nul !

↳ Malgré ces difficultés théoriques, en pratique on observe souvent (et notamment en sciences des données), on observe souvent que la descente de gradient converge vers des minima locaux voire globaux

↳ En s'intéressant à des sous-classes de problèmes non convexe, on peut expliquer ce phénomène pratique, et même construire des algorithmes encore plus efficaces que la descente de gradient dans le cas non convexe.

(2) Paysages d'optimisation

↳ On s'intéresse à des fonctions

f C^2 $\Rightarrow$ $\forall w \in \mathbb{R}^d$,

f de classe C^2

on peut définir $\nabla f(w) \in \mathbb{R}^d$ (gradient)

et $\nabla^2 f(w) \in \mathbb{R}^{d \times d}$

(matrice hessienne, qui est symétrique)

Pour de telles fonctions, on

$[w^* \in \mathbb{R}^d$ minimum local de $f]$

sait que

$\Rightarrow [\nabla f(w^*) = 0$ et $\nabla^2 f(w^*) \geq 0]$

$v^T \nabla^2 f(w^*) v \geq 0$ $\forall v \in \mathbb{R}^d$

ce qui est équivalent à

$[\bar{w} \in \mathbb{R}^d$ n'est pas un minimum local $]$

$\Leftarrow [\nabla f(\bar{w}) \neq 0$ ou $\exists v \in \mathbb{R}^d$, $v^T \nabla^2 f(\bar{w}) v < 0]$

v est une direction de courbure négative pour f en $\bar{w}$

Def. Paysage d'optimisation de f

Le paysage de f est la fonction des valeurs de ∇f et $\nabla^2 f$ en ces points.

$\forall w \in \mathbb{R}^d$, on parle de

classification des points de $\mathbb{R}^d$ en

- Point non critique : si $\nabla f(w) \neq 0$
- Point critique d'ordre 1 : si $\nabla f(w) = 0$
- Point selle : si $\nabla f(w) = 0$ et $\nabla^2 f(w) \not\geq 0$
- Point critique d'ordre 2 : si $\nabla f(w) = 0$ et $\nabla^2 f(w) \geq 0$

NB: {Points critiques d'ordre 1} = {Points selles} ∪ {Points critiques d'ordre 2}

$\nabla^2 f(w) \geq 0 \iff \forall v \in \mathbb{R}^d, \quad v^T \nabla^2 f(w) v \geq 0$

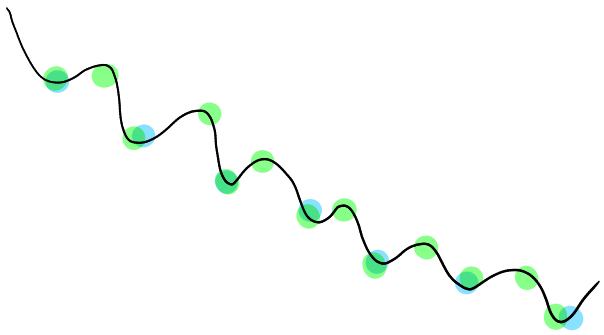
$\nabla^2 f(w) \not\geq 0 \iff \exists v \in \mathbb{R}^d, \quad v^T \nabla^2 f(w) v < 0$

$\nabla^2 f(w) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \not\geq 0$

$\begin{bmatrix} 1 & 0 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} > 0$

$\begin{bmatrix} 0 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} < 0$

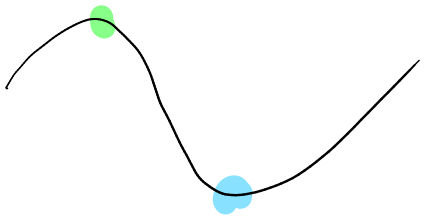
Mauvais paysage d'optimisation



→ Beaucoup de points critiques d'ordre 1 et d'ordre 2

→ Beaucoup de minima locaux non globaux

Bon paysage d'optimisation



→ Peu de points critiques

→ Tout minimum local est global

→ En sciences des données, il rencontrés ont en général un bon plus faciles à résoudre qu'un mauvais paysage.

→ Exemples : reconstruction de réseaux de neurones

a été montré que les problèmes paysage d'optimisation, et sont donc problème d'optimisation avec un

phase, apprentissage de dictionnaire, linéaires, ACP (analyse en composantes principales)

Notre exemple pour le projet :

Données : $M \in \mathbb{R}^{d \times d}$ matrice de

⇒ On dispose d'un de coefficients de

⇒ Cadre typique

But : Trouver la meilleure

Approche naïve : minimiser $W \in \mathbb{R}^{d \times d}$

⊕ Problème convexe

⊕ Solution est M si $S = \{1, \dots, d\} \times \{1, \dots, d\}$

⊖ Pas d'information concernant qui ne sont pas observés

Approche moderne : On suppose que

⇒ Conceptuellement :

sans avoir à connaître

⇒ Mathématiquement :

Problème d'optimisation associé

Paramètre : Approximation de U^* → minimiser $U \in \mathbb{R}^{d \times r}$ $\frac{1}{2} \sum_{(i,j) \in S}$

Factorisation de matrice de rang faible

données que l'on observe partiellement

sous-ensemble $S \subseteq \{1, \dots, d\} \times \{1, \dots, d\}$

M

de systèmes de recommandation (ex : Netflix)

approximation possible de M

$$\frac{1}{2} \sum_{(i,j) \in S} (W_{ij} - M_{ij})^2$$

On cherche la matrice W qui colle le mieux possible aux observations de M

$\dots, d\} \times \{1, \dots, d\}$ (si on connaît tous les coefficients)

les coefficients de la matrice

M possède une structure de rang faible

On peut approcher M de manière précise beaucoup de ses coefficients

On suppose qu'il existe $U^* \in \mathbb{R}^{d \times r}$ avec $r \ll d$ telle que $M = U^* (U^*)^T$

$$\left([UU^T]_{ij} - M_{ij} \right)^2$$

UU^T : Approximation de M

⊕ Exploite la structure de M

⊕ Moins de variables dans cette d^2 pour l'approche naïve

⊖ Non convexe

⊖ Présence de points selles

→ On peut montrer pour ce que les points critiques d'ordre des minima globaux.

→ Par ailleurs, si on échantillonne un minimum global.

problème (résultats de 2016-2017) 1 sur soit des points selles, soit

$O(\log(d))$ valeurs, alors M est

suffisamment de valeurs, mais mais que la taille de M !

Bilan :

- Caractériser le paysage difficulté d'un problème difficulté à trouver ses
- De nombreux problèmes propriété { minima

d'optimisation permet de juger de la non convexe, et notamment de la minima globaux

en science des données possèdent la globaux } = { points critiques d'ordre 2 }

factorisation de matrice)

(3) Descente de gradient

et points critiques d'ordre 2

Q : Etant donné un problème non convexe avec un bon paysage, que peut-on dire de la descente de gradient sur ce problème ?

Descente de gradient : $w_{k+1} = w_k - \alpha_k \nabla f(w_k)$ avec $\alpha_k > 0$

partant de $w_0 \in \mathbb{R}^d$

Si w_0 est un point selle ce n'est pas un minimum et w_0 ! $w_1 = w_0$, $w_2 = w_1$, ...

$(\nabla f(w_0) = 0,\ \nabla^2 f(w_0) \not\succeq 0)$, alors la descente de gradient est bloquée en

$\Rightarrow$ Très facile de construire des exemples pour lesquels la descente de gradient ne converge pas

$\Rightarrow$ Mais en pratique, ces exemples ne semblent pas fréquents

Théorème (2015-2019) : La descente de gradient converge vers un point critique à l'ordre 2 pour presque tout $w_0 \in \mathbb{R}^d$.

(4) Le projet

→ En individuel ou en binôme

→ Point de départ :
- Notebook sur l'optimisation non convexe
- Sujet avec des questions

→ But : Etudier un algorithme de descente de gradient + courbure négative

Si $\nabla f(w_k) = 0$ et $\nabla^2 f(w_k) \not\succeq 0$, alors

$$w_{k+1} = w_k + \alpha_k v_k$$

avec $v_k^T \nabla^2 f(w_k) v_k < 0$

→ Converge plus vite vers des points critiques d'ordre 2 que la descente de gradient

→ Convexe même en partant d'un point selle !

↳ Le projet :
- Questions théoriques (démoissance, complexité)
- Application à la factorisation de matrice stochastique
- Cadre

↳ Deadline ≈ fin janvier 2025