Projet Examen Final: Optimisation pour le Machine Learning

Modélisation, SGD et Méthodes Proximales

Auteur :
DIENG BABAKAR

Date :
February 1, 2026

Faculté des sciences et techniques (FST)

Introduction

Ce projet vise à illustrer les liens entre la théorie de l'optimisation convexe et les algorithmes modernes utilisés en Machine Learning à grande échelle. Nous étudions successivement la modélisation, les méthodes de gradient déterministes et stochastiques, puis l'optimisation non lisse via des algorithmes proximaux.

1 Exercice 1 : Modélisation et étude théorique

1.1 Modélisation

On considère le problème de régression linéaire régularisée :

$$\min_{w\in\mathbb{R}^d} f(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(x_i^T w - y_i)^2 + \frac{\mu}{2}\|w\|_2^2.$$

L'ajout du terme de régularisation rend la fonction μ-fortement convexe. D'après le Théorème du cours, une fonction fortement convexe admet un unique minimiseur global.

Justification de l'unicité du minimum global

Soit la fonction objectif de régression linéaire

$$f_0(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(x_i^T w - y_i)^2, \quad w \in \mathbb{R}^d.$$

La fonction f_0 est convexe et de classe C^2, mais elle n'est pas nécessairement strictement convexe lorsque la matrice des données $X \in \mathbb{R}^{n\times d}$ n'est pas de rang plein. Dans ce cas, la matrice $X^T X$ est seulement semi-définie positive et le minimum global de f_0 n'est pas unique.

On introduit alors la fonction régularisée

$$f(w) = f_0(w) + \frac{\mu}{2}\|w\|_2^2, \quad \mu > 0.$$

La Hessienne de f est donnée par

$$\nabla^2 f(w) = \frac{1}{n}X^T X + \mu I_d.$$

Pour tout vecteur non nul $v \in \mathbb{R}^d$, on a

$$v^T \nabla^2 f(w) v = \frac{1}{n}\|Xv\|_2^2 + \mu\|v\|_2^2 \geq \mu\|v\|_2^2 > 0.$$

Ainsi, la matrice $\nabla^2 f(w)$ est définie positive et la fonction f est μ-fortement convexe.

D'après le Théorème 1.2.9 du cours, toute fonction μ-fortement convexe admet un unique minimiseur global. Par conséquent, l'ajout du terme de régularisation $\frac{\mu}{2}\|w\|_2^2$ garantit l'unicité du minimum global du problème d'optimisation.

1.2 Gradient et Hessienne

Le gradient s'écrit :

$$\nabla f(w) = \frac{1}{n}X^T(Xw - y) + \mu w.$$

La Hessienne est donnée par :

$$\nabla^2 f(w) = \frac{1}{n}X^T X + \mu I_d,$$

qui est définie positive pour $\mu > 0$.

1.3 Constante de Lipschitz

Soit la décomposition en valeurs singulières $X = U\Sigma V^T$. La constante de Lipschitz du gradient est :

$$L = \frac{1}{n}\sigma_{\max}^2(X) + \mu.$$

Cette constante conditionne le pas maximal autorisé pour assurer la stabilité de la descente de gradient.

2 Exercice 2 : Méthodes stochastiques

2.1 Descente de gradient stochastique

La mise à jour SGD est donnée par :

$$w_{k+1} = w_k - \eta_k \nabla f_{i_k}(w_k).$$

On montre que $\mathbb{E}[\nabla f_i(w)] = \nabla f(w)$, ce qui garantit que le gradient stochastique est un estimateur sans biais.

2.2 Analyse comparative

Le gradient batch présente une convergence stable mais coûteuse, tandis que le SGD converge rapidement au début mais subit des oscillations dues au bruit de gradient.

2.3 Mini-batch et Adam

Les méthodes mini-batch offrent un compromis entre variance et coût de calcul. L'algorithme Adam adapte automatiquement les pas de descente grâce à des estimations des moments du gradient.

3 Exercice 3 : Optimisation parcimonieuse

3.1 Régularisation L^1

La géométrie de la boule L^1 favorise des solutions creuses, contrairement à la régularisation L^2 qui produit des solutions denses.

3.2 ISTA

L'opérateur proximal de la norme L^1 est le seuillage doux :

$$\text{prox}_{\lambda\|.\|_1}(z)_i = \text{sign}(z_i)\max(|z_i| - \lambda, 0).$$

3.3 FISTA

L'accélération de Nesterov permet d'améliorer le taux de convergence de $O(1/k)$ à $O(1/k^2)$.

3.4 Sélection de variables

Lorsque λ augmente, le nombre de coefficients non nuls diminue, ce qui permet d'identifier les variables les plus pertinentes.

Conclusion

Ce projet montre l'importance des propriétés théoriques (convexité, Lipschitz, proximalité) dans la conception d'algorithmes efficaces pour le Machine Learning moderne.