Mini-Projet d'Optimisation en Machine Learning

Solutions mathématiques détaillées

Introduction

On considère un problème de classification binaire à partir d'un jeu de données $\{(x_i, y_i)\}_{i=1}^n$, avec $x_i \in \mathbb{R}^d$ et $y_i \in \{-1, +1\}$.

L'objectif est d'étudier rigoureusement, d'un point de vue mathématique, les propriétés du problème d'optimisation, les algorithmes de gradient (déterministes et stochastiques) ainsi que les méthodes proximales pour la régularisation parcimonieuse.

1 Phase 1 – Fondements et Gradient déterministe

1.1 Analyse de la fonction objectif

La fonction à minimiser est :

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i x_i^T w}) + \frac{\lambda}{2} \|w\|_2^2$$

a) Régularité (classe C^2)

- La fonction scalaire $t \mapsto \log(1 + e^{-t})$ est C^∞ sur $\mathbb{R}$.
- La fonction $w \mapsto y_i x_i^T w$ est linéaire donc C^∞.
- La composition d'une fonction C^∞ avec une application linéaire reste C^∞.
- Le terme $\|w\|_2^2$ est un polynôme de degré 2 donc C^∞.

Conclusion : F est de classe C^2.

b) Convexité

- La fonction $t \mapsto \log(1 + e^{-t})$ est convexe.
- La composition avec une fonction linéaire conserve la convexité.
- Une somme de fonctions convexes est convexe.
- Le terme $\frac{\lambda}{2} \|w\|_2^2$ est strictement convexe.

Conclusion : F est convexe sur $\mathbb{R}^d$.

c) Forte convexité

Le Hessien du terme quadratique vaut :

$$\nabla^2 \left(\frac{\lambda}{2} \|w\|_2^2 \right) = \lambda \mathrm{Id}$$

Ainsi, pour tout w :

$$\nabla^2 F(w) \succeq \lambda \mathrm{Id}$$

Conclusion : F est λ-fortement convexe, ce qui garantit :

— l'unicité du minimiseur w^*,

— une convergence linéaire de la descente de gradient.

1.2 Calcul du gradient et Lipschitzianité

a) Gradient

On pose $\sigma(t) = \frac{1}{1+e^{-t}}$. On obtient :

$$\nabla F(w) = \frac{1}{n} \sum_{i=1}^{n} -y_i x_i \sigma(-y_i x_i^T w) + \lambda w$$

b) Gradient Lipschitzien

Le Hessien s'écrit :

$$\nabla^2 F(w) = \frac{1}{n} X^T D(w) X + \lambda \mathrm{Id}$$

où $D(w)$ est diagonale avec

$$0 \leq D_{ii}(w) = \sigma(z_i)(1 - \sigma(z_i)) \leq \frac{1}{4}$$

Ainsi :

$$\|\nabla^2 F(w)\| \leq \frac{1}{4n} \|X\|^2 + \lambda$$

Conclusion : le gradient est L-Lipschitzien avec

$$L = \frac{1}{4n} \|X\|^2 + \lambda$$

1.3 Descente de gradient vs Gradient conjugué

— La descente de gradient converge linéairement avec un pas $\alpha \in (0, 2/L)$.

— Le gradient conjugué exploite la structure quadratique locale et converge en un nombre fini d'itérations dans le cas quadratique.

Comparaison théorique :

— DG : simple mais lente si $\kappa = L/\lambda$ est grand.

— GC : convergence plus rapide mais coût par itération plus élevé.

2 Phase 2 – Passage à l'échelle stochastique

2.1 Gradient stochastique (SGD)

On approxime le gradient complet par :

$$\nabla F_{i_k}(w) = -y_{i_k} x_{i_k} \sigma(-y_{i_k} x_{i_k}^T w) + \lambda w$$

La mise à jour est :

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

avec un pas décroissant :

$$\alpha_k = \frac{\alpha_0}{1+k}$$

Cette condition assure la convergence presque sûre.

2.2 RMSProp et Adam (analyse conceptuelle)

— RMSProp normalise le gradient par une moyenne exponentielle de ses carrés.
— Adam combine RMSProp et momentum.

Résultat théorique : meilleure stabilité et convergence plus rapide dans les premières itérations.

2.3 Rôle du momentum

Le momentum introduit une mémoire des gradients passés :

$$v_{k+1} = \beta v_k + (1-\beta) g_k$$

Effets :

— réduction des oscillations,
— accélération dans les vallées étroites,
— meilleure stabilité numérique.

3 Phase 3 – Non-lisse et méthodes proximales

3.1 Non-lissité du problème

La fonction

$$\Phi(w) = f(w) + \lambda \|w\|_1$$

n'est pas différentiable car $|t|$ n'est pas dérivable en 0.

3.2 Opérateur proximal de la norme L1

L'opérateur proximal est défini par :

$$\text{prox}_{\lambda\|\cdot\|_1}(z) = \arg\min_u \left(\frac{1}{2}\|u-z\|^2 + \lambda \|u\|_1 \right)$$

Il s'agit du seuil doux :

$$(\text{prox}_{\lambda\|\cdot\|_1}(z))_j = \text{sign}(z_j) \max(|z_j| - \lambda, 0)$$

3.3 ISTA

$$w_{k+1} = \mathrm{prox}_{\alpha\lambda\|\cdot\|_1}(w_k - \alpha \nabla f(w_k))$$

Convergence en $O(1/k)$.

3.4 FISTA

FISTA introduit une extrapolation :

$$\begin{cases} w_{k+1} = \mathrm{prox}(z_k - \alpha \nabla f(z_k)) \\ z_{k+1} = w_{k+1} + \frac{t_k - 1}{t_{k+1}}(w_{k+1} - w_k) \end{cases}$$

Convergence accélérée : $O(1/k^2)$.

3.5 Interprétation de λ

- λ petit : solution dense.
- λ grand : nombreux coefficients nuls.

Conclusion : la régularisation L1 effectue une sélection automatique de variables.

Conclusion générale

Ce projet met en évidence :

- le lien entre propriétés mathématiques (convexité, Lipschitz) et convergence,
- l'intérêt des méthodes stochastiques pour les grands jeux de données,
- la puissance des méthodes proximales pour la parcimonie.

Ces outils constituent le socle théorique de l'optimisation moderne en machine learning.