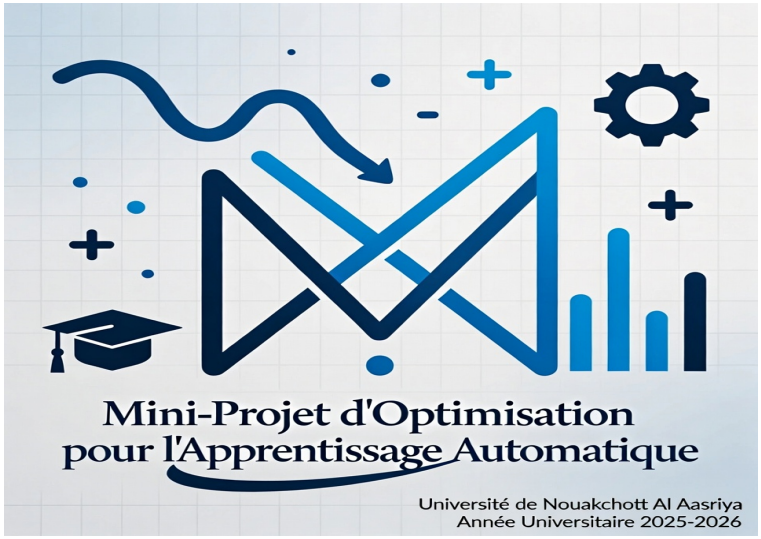
Université de Nouakchott Al Aasriya
Faculté des Sciences et Techniques
Département Mathématiques et Informatiques

Mini-Projet d'Optimisation pour l'Apprentissage Automatique



Modélisation, Gradient Stochastique (SGD) et Méthodes Proximales pour l'Optimisation en Apprentissage Automatique

Réalisé par :

Mohamed Lemine Abdallahi Tah C12896

Encadré par :
Dr El Benany Mohamed Mahmoud

Année universitaire : 2025 – 2026

Projet réalisé dans le cadre du Master SSD – Statistiques et Sciences des Données

Table des matières

Note

Les objectifs principaux sont les suivants :
- Étudier les propriétés mathématiques de la perte logistique régularisée.
- Comparer des méthodes déterministes d'optimisation (DG et Gradient Conjugué).
- Analyser les algorithmes stochastiques (SGD avec pas décroissant, RMSProp, Adam) et l'impact du momentum.
- Introduire la régularisation ℓ_1 favorisant la parcimonie et étudier les méthodes proximales ISTA et FISTA.

Note

L'ensemble du travail (notebook Jupyter, figures générées et rapport final) est disponible dans le dépôt GitHub suivant : `https://github.com/MoLemine/Optimization`
Le dossier principal du projet est **Exam/** et contient :
- **optim.ipynb** : notebook Jupyter regroupant toutes les implémentations, simulations et figures ;
- **rapport.pdf** : rapport final présentant l'analyse théorique et l'interprétation des résultats.

1 Introduction

L'apprentissage automatique repose sur la résolution de problèmes d'optimisation de grande dimension, souvent définis comme la minimisation d'une fonction objectif construite à partir de données observées. Dans ce contexte, la compréhension théorique des algorithmes d'optimisation est essentielle afin de garantir la stabilité numérique, la convergence et l'efficacité computationnelle.

Ce travail s'inscrit dans le cadre de l'optimisation différentiable et étudie plusieurs méthodes de descente de gradient appliquées à un problème de régression linéaire régularisée. L'accent est mis sur les propriétés théoriques des fonctions considérées et sur les garanties de convergence associées aux algorithmes étudiés.

2 Formulation du problème

On considère un ensemble de données $\{(x_i, y_i)\}_{i=1}^n$, avec $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$. Le problème étudié consiste à minimiser la fonction objectif suivante :

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}(x_i^\top w - y_i)^2 + \frac{\mu}{2}\|w\|^2, \tag{1}$$

où $\mu > 0$ est un paramètre de régularisation.

Cette formulation correspond à une régression linéaire régularisée de type Ridge. La régularisation quadratique permet de contrôler la norme de la solution et d'améliorer le conditionnement du problème.

2.1 Propriétés théoriques de la fonction objectif

La fonction f vérifie les propriétés suivantes :
— f est convexe, en tant que somme de fonctions convexes ;
— f est continûment différentiable ;
— f est μ-fortement convexe ;
— le gradient de f est Lipschitzien.

Le gradient de f est donné par :

$$\nabla f(w) = \frac{2}{n}X^\top(Xw - y) + \mu w, \tag{2}$$

et la Hessienne est constante :

$$\nabla^2 f(w) = \frac{2}{n}X^\top X + \mu I_d. \tag{3}$$

La forte convexité garantit l'existence et l'unicité du minimiseur global $w^\star$.

3 Descente de gradient batch

La descente de gradient batch consiste à itérer :

$$w_{k+1} = w_k - \alpha\nabla f(w_k), \tag{4}$$

où α est le pas d'apprentissage.

3.1 Convergence

Si le gradient de f est L-Lipschitzien, alors pour tout pas $\alpha \in (0, 2/L)$, la suite (w_k) converge vers $w^\star$. De plus, en présence de forte convexité, la convergence est linéaire :

$$\|w_k - w^\star\| \leq C\rho^k, \tag{5}$$

avec $\rho \in (0, 1)$.

Cette méthode est théoriquement robuste, mais chaque itération nécessite le calcul du gradient sur l'ensemble des données, ce qui limite son applicabilité aux grands jeux de données.

Résultats numériques

La Figure 1 illustre la convergence monotone de la fonction objectif, en accord avec les résultats théoriques.

4 Descente de gradient stochastique

La descente de gradient stochastique (SGD) remplace le gradient exact par une approximation basée sur une seule observation :

$$\nabla f_i(w) = 2x_i(x_i^\top w - y_i) + \mu w. \tag{6}$$

Analyse théorique

Le gradient stochastique est un estimateur non biaisé du gradient exact :

$$\mathbb{E}[\nabla f_i(w)] = \nabla f(w). \tag{7}$$

Cependant, la variance du gradient stochastique est non nulle, ce qui entraîne des oscillations autour du minimum et empêche une convergence monotone.

Illustration expérimentale

La Figure 2 montre que le SGD présente des fluctuations importantes par rapport à la descente de gradient batch, ce qui est conforme aux résultats théoriques.

5 Méthode mini-batch

La méthode mini-batch généralise le SGD en utilisant un sous-ensemble de taille b à chaque itération.

5.1 Justification théorique

L'augmentation de la taille du batch permet de réduire la variance du gradient tout en conservant un coût de calcul inférieur à celui du gradient batch complet. Cette méthode constitue donc un compromis naturel entre stabilité et efficacité.

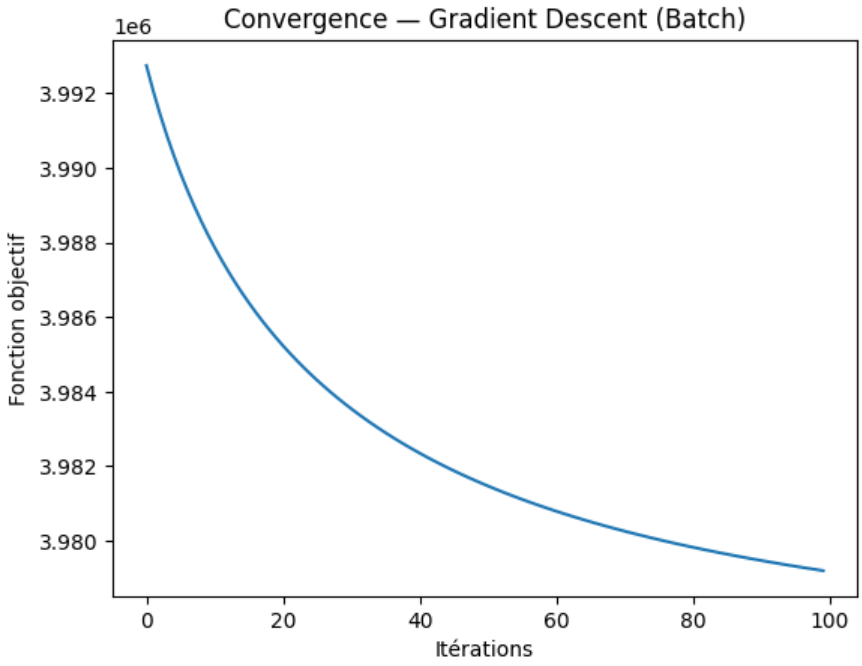


FIGURE 1 – Convergence de la descente de gradient batch.

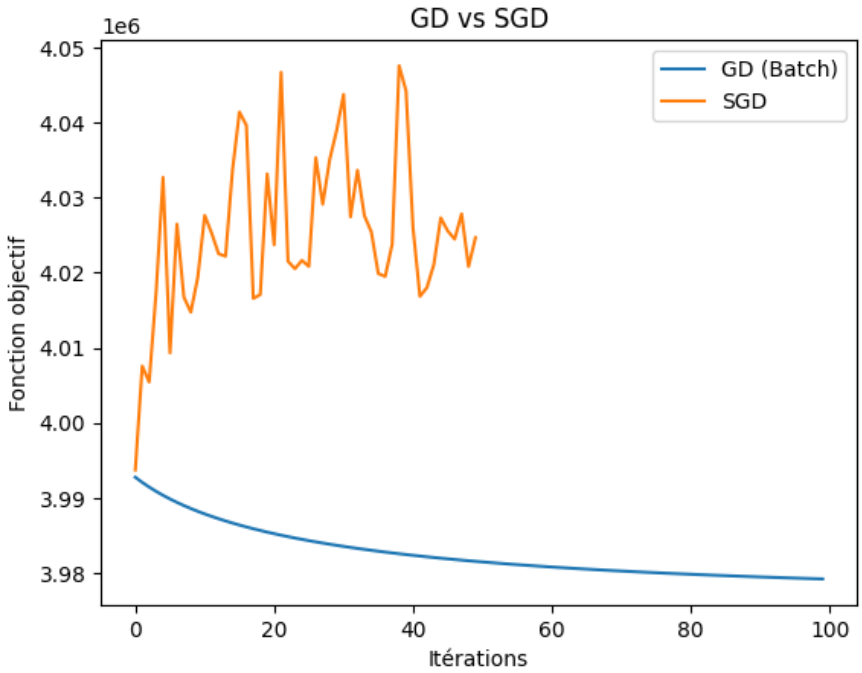


FIGURE 2 – Comparaison GD vs SGD.

5.2 Résultats expérimentaux

La Figure 3 illustre ce compromis, avec une convergence plus stable que le SGD et un coût de calcul réduit par rapport à GD.

6 Méthode adaptative Adam

Adam est une méthode adaptative qui ajuste automatiquement le pas d'apprentissage à partir des moments du gradient.

6.1 Analyse théorique

L'utilisation conjointe du premier et du second moment du gradient permet de réduire l'impact du bruit et d'améliorer la stabilité numérique, en particulier pour les problèmes mal conditionnés.

6.2 Illustration expérimentale

La Figure 4 montre une convergence rapide et stable, confirmant l'intérêt pratique des méthodes adaptatives.

7 Parcimonie et Algorithmes Proximaux

7.1 a) Analyse géométrique : L1 vs L2

La régularisation L2 (ridge) pénalise la norme euclidienne du vecteur des paramètres :

$$\|w\|_2^2 \leq c$$

dont la boule associée est sphérique. Lorsque cette contrainte intersecte les lignes de niveau de la fonction de perte, la solution optimale est rarement située sur les axes, ce qui conduit à des coefficients généralement non nuls.

À l'inverse, la régularisation L1 (lasso) repose sur la norme :

$$\|w\|_1 \leq c$$

dont la boule est un polytope avec des sommets alignés sur les axes de coordonnées. Géométriquement, les points d'intersection avec les niveaux de la perte se produisent fréquemment sur ces sommets, ce qui force de nombreux coefficients à être exactement nuls.

Cette propriété rend la régularisation L1 particulièrement adaptée aux problèmes de grande dimension, car elle induit naturellement une sélection de variables et améliore l'interprétabilité du modèle.

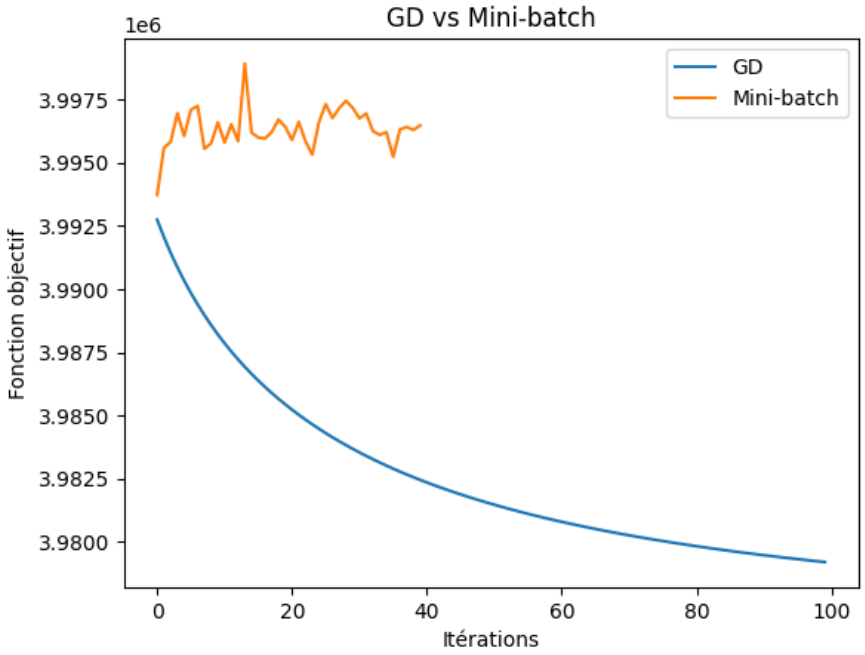


FIGURE 3 – Comparaison GD vs Mini-batch.

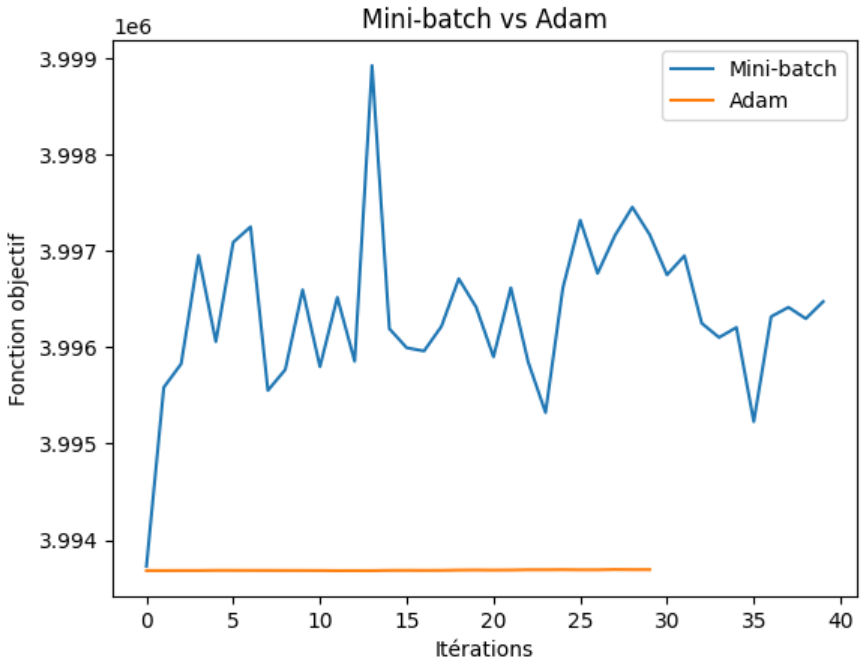


FIGURE 4 – Comparaison Mini-batch vs Adam.

7.2 b) ISTA et opérateur proximal de la norme L1

On considère le problème d'optimisation convexe :

$$\min_w \ f(w) + \lambda \|w\|_1$$

où f est une fonction différentiable à gradient Lipschitzien.

L'opérateur proximal associé à la norme L1 est donné par le *soft-thresholding* :

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

L'algorithme ISTA (Iterative Shrinkage-Thresholding Algorithm) est alors défini par :

$$w^{k+1} = \text{prox}_{\alpha\lambda\|\cdot\|_1}\left(w^k - \alpha\nabla f(w^k)\right)$$

Sous des hypothèses standard de convexité et de Lipschitzianité du gradient, ISTA converge avec un taux sublinéaire :

$$f(w^k) - f(w^\star) = \mathcal{O}\left(\frac{1}{k}\right)$$

7.3 c) Accélération par FISTA

FISTA (Fast ISTA) améliore ISTA en introduisant un terme d'extrapolation inspiré des méthodes de Nesterov. Il repose sur les itérations suivantes :

$$y^k = w^k + \frac{t_{k-1} - 1}{t_k}(w^k - w^{k-1})$$
$$w^{k+1} = \text{prox}_{\alpha\lambda\|\cdot\|_1}\left(y^k - \alpha\nabla f(y^k)\right)$$

avec

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}$$

Cette accélération permet d'atteindre un taux de convergence optimal pour les méthodes du premier ordre :

$$f(w^k) - f(w^\star) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

7.4 Comparaison empirique ISTA / FISTA

La Figure 5 illustre la convergence empirique des deux méthodes sur le dataset RCV1. On observe que FISTA converge significativement plus rapidement qu'ISTA, ce qui est en parfait accord avec les résultats théoriques.

d) Sélection de variables et effet de λ

La régularisation L1 permet de contrôler directement le niveau de parcimonie via le paramètre λ. Lorsque λ augmente, la pénalisation devient plus forte et davantage de coefficients sont annulés.

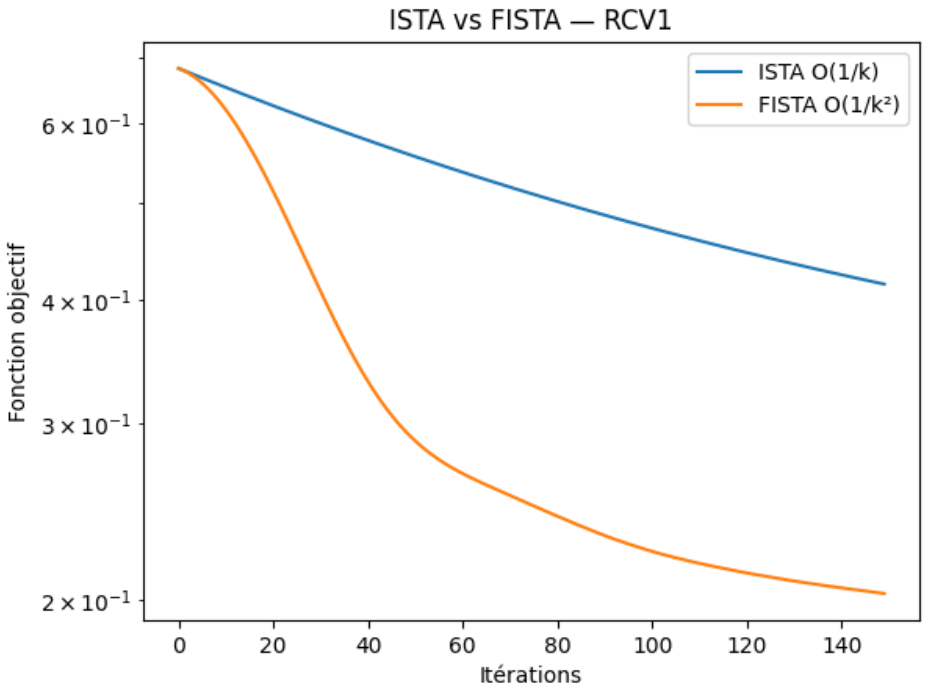


FIGURE 5 – Comparaison de la convergence d'ISTA ($\mathcal{O}(1/k)$) et FISTA ($\mathcal{O}(1/k^2)$) sur RCV1

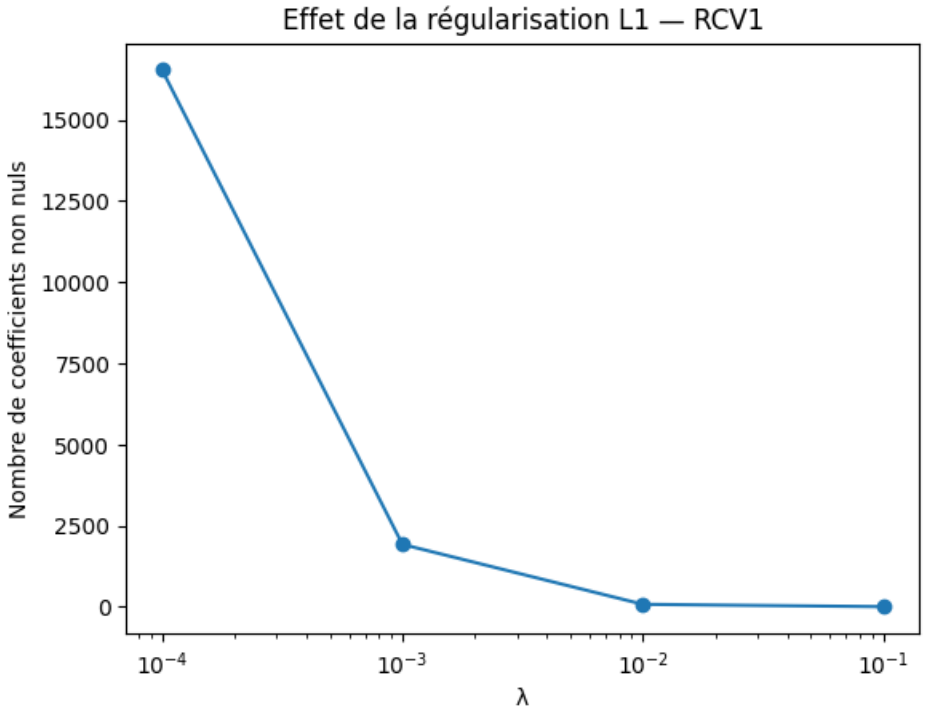


FIGURE 6 – Effet de la régularisation L1 sur la parcimonie du modèle

La Figure 6 montre l'évolution du nombre de coefficients non nuls en fonction de λ. On constate une chute brutale du nombre de variables actives, ce qui confirme la capacité du Lasso à effectuer une sélection automatique de mots discriminants.

Ces résultats sont particulièrement pertinents dans un contexte de classification de documents, où seuls quelques termes clés sont réellement informatifs.

8 Conclusion

L'étude théorique et expérimentale menée dans ce travail met en évidence l'importance des propriétés analytiques des fonctions objectives dans le choix des méthodes d'optimisation. Si la descente de gradient batch offre des garanties fortes de convergence, les méthodes stochastiques et adaptatives permettent un passage à l'échelle efficace, tout en conservant de bonnes propriétés de stabilité.