TP 3 : Optimisation Stochastique

Classification (Iris) et Régression (California Housing)

SSD : Optimisation Différentiable

2025/2026

1 Introduction et Objectifs

Ce TP vise à illustrer les concepts du **Chapitre 3** sur le Gradient Stochastique (SGD). Contrairement au Gradient de Batch (GD) qui utilise tout le dataset, le SGD permet de traiter de grands volumes de données en mettant à jour le modèle de manière itérative.

— **Partie 1 :** Classification binaire sur *Iris* (Comprendre la mise à jour).

— **Partie 2 :** Régression sur *California Housing* (Analyse de la convergence et des variantes).

2 Partie 1 : Classification sur Iris (Le "Petit" Pas)

Le dataset **Iris** contient 150 exemples. Nous allons simplifier le problème en une classification binaire (Iris-Setosa vs autres).

2.1 Exercice 1 : Descente de Gradient Stochastique (SGD) "From Scratch"

L'objectif est de coder la règle de mise à jour vue en section 3.2 : $w \leftarrow w - \alpha \nabla f_i(w)$.

1. Chargez les données via `sklearn.datasets.load_iris`.
2. Implémentez la fonction de coût Logistique (Cross-Entropy).
3. Écrivez une boucle qui tire un échantillon aléatoire à chaque itération et met à jour les poids w.
4. **Question :** Tracez la courbe de coût. Pourquoi est-elle très instable par rapport à une descente de gradient classique ?

3 Partie 2 : Régression sur California Housing (Le Passage à l'Échelle)

Ce dataset contient plus de 20 000 lignes. Calculer le gradient complet ici est coûteux, ce qui motive l'usage du SGD.

3.1 Exercice 2 : Importance de la Standardisation

Le SGD est extrêmement sensible à l'échelle des données.

1. Entraînez un modèle `SGDRegressor` sur les données brutes.
2. Utilisez `StandardScaler` pour normaliser les données et ré-entraînez le modèle.
3. **Analyse :** Comparez les temps de convergence. Pourquoi la normalisation aide-t-elle l'algorithme à "descendre" plus vite ? (Indice : regardez la forme des lignes de niveau de la fonction de coût).

3.2 Exercice 3 : Mini-batch et Optimiseurs Modernes

Référence Section 3.3 et 3.4 du cours.

1. Comparez l'évolution de la MSE (Mean Squared Error) pour :
   — SGD Pur ($batch\_size = 1$).
   — Mini-batch SGD ($batch\_size = 32$).
   — Adam (Optimiseur adaptatif).
2. **Observation :** Lequel de ces algorithmes atteint le "plateau" de performance le plus rapidement ?

4 Synthèse et Rapport

Dans votre compte-rendu, vous devrez répondre aux points suivants :

1. Pourquoi ne faut-il jamais utiliser un pas α trop grand avec le SGD ? (Section 3.2).
2. Quel est l'avantage computationnel du Mini-batch par rapport au SGD pur sur une carte graphique (GPU) ?
3. Expliquez le concept de **Shuffling** (mélange) : pourquoi est-il crucial pour le gradient stochastique ?