Optimisation stochastique

Dieng Babakar
C14050

Université FST

January 16, 2026

Plan

Introduction

Intro

L'optimisation stochastique constitue une réponse efficace à ces limitations en introduisant des approximations du gradient permettant de réduire significativement le coût de calcul par itération.
Ce TP, réalisé dans le cadre du Chapitre 3, a pour objectif d'étudier et de comparer différentes méthodes d'optimisation stochastique, telles que la descente de gradient classique, le gradient stochastique (SGD), la méthode mini-batch et les optimiseurs adaptatifs comme Adam.

Descente de gradient classique

Objectif

On considère un ensemble de données :

$$\{(x_i, y_i)\}_{i=1}^{n}$$

avec :

- $x_i \in \mathbb{R}^d$: variables explicatives
- $y_i \in \{-1, +1\}$ ou $y_i \in \mathbb{R}$

L'objectif est de minimiser :

$$\min_{w \in \mathbb{R}^d} F(w)$$

Fonction objectif

Exemple & Propriété

Exemple : **perte logistique**

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-y_i x_i^T w}\right)$$

Propriétés :

- Fonction convexe
- Gradient lipschitzien
- Minimisation par descente de gradient

Gradient Stochastique

Définition

Le gradient de F est donné par :

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

Coût de calcul :

$$\mathcal{O}(nd)$$

Très coûteux lorsque n est grand.

Descente de gradient (Batch GD)

Schéma itératif :

$$w_{k+1} = w_k - \alpha \nabla F(w_k)$$

Condition de convergence :

$$\alpha \leq \frac{1}{L}$$

Inconvénient majeur :

- Calcul du gradient sur tout le jeu de données

Motivation du SGD

Idée clé :

- Approximater le gradient par un seul échantillon

$$\nabla F(w) \approx \nabla f_i(w)$$

Avantage :

- Coût par itération : $\mathcal{O}(d)$

Algorithme SGD

À l'itération k :

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

avec :

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

Avantages :

- Très rapide
- Adapté aux grands jeux de données

Inconvénients du SGD

- Gradient bruité
- Oscillations autour du minimum
- Convergence non monotone
- Nécessité d'améliorations

Principe du Mini-batch

Compromis entre :

- Batch Gradient Descent
- Gradient Stochastique

Gradient calculé sur un sous-ensemble B_k :

$$\nabla F_{B_k}(w) = \frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(w)$$

Avantages du Mini-batch

- Réduction de la variance
- Utilisation efficace du parallélisme
- Convergence plus stable que SGD

Adam

Adam combine :

- Momentum
- RMSProp

Mises à jour :

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_k$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2$$

Mise à jour Adam

$$w_{k+1} = w_k - \alpha \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \varepsilon}$$

Avantages :

- Très rapide au début
- Stable numériquement
- Très utilisé en pratique

Synthèse des méthodes

Comparaison :

- SGD : rapide mais très bruité
- Mini-batch : meilleur compromis variance / vitesse
- Adam : convergence la plus rapide

Bonnes pratiques :

- Pas trop grand $\Rightarrow$ divergence
- Shuffling $\Rightarrow$ indépendance statistique

Merci pour votre attention