TP Chapitre 3
Optimisation stochastique

Fatimetou Limam Abeid
C16698

Université FST

January 14, 2026

Plan

Problème étudié

Le TP porte sur l'**optimisation stochastique** appliquée à des problèmes de **classification** et de **régression**.

Objectifs :

- Transformer un problème statistique en problème d'optimisation
- Comparer différentes méthodes de gradient
- Comprendre le compromis précision / vitesse / variance

Formulation mathématique

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 une fonction :

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

Fonction objectif

Exemple : perte logistique

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

Propriétés :

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

Gradient exact

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 quand n est grand

Descente de gradient (Batch GD)

Schéma itératif :

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

Convergence garantie si :

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

Inconvénient majeur :

- Calcul du gradient sur tout le jeu de données

Motivation du SGD

Idée clé :

- Approximater le gradient par un seul échantillon
- Réduire le coût par itération

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

Complexité :

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

Algorithme SGD

À l'itération k :

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

où :

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

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 GD
- SGD

Gradient calculé sur un sous-ensemble $\mathcal{B}_k$:

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

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

Conclusion

Ce TP montre que :

- L'optimisation stochastique est essentielle en grande dimension
- Les méthodes du Chapitre 3 sont complémentaires
- La théorie explique les comportements observés