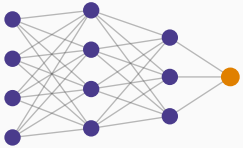
Apprentissage profond

Chapitre 4 : La descente de gradient et ses variantes



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

Le principe

Descendre dans le sens de la pente

La règle de mise à jour

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

Le gradient indique la direction de plus forte **montée** ; on va donc dans le sens opposé. η est le **taux d'apprentissage**.

C'est le chapitre 7 de Mathématiques pour la gestion

Chercher un minimum en annulant les dérivées partielles est exactement la condition du premier ordre à plusieurs variables.

La différence est qu'ici le système ne se résout pas : il y a des millions de paramètres, et l'on **descend pas à pas** au lieu de résoudre.

Le taux d’apprentissage

Le paramètre le plus important du cours

Une descente entièrement calculable

Minimisons $L(w) = w^2$, de dérivée $2w$, en partant de $w = 1$.

La règle donne $w \leftarrow w(1 - 2\eta)$: chaque pas multiplie w par un facteur fixe.

η	Facteur	Trajectoire
$0{,}1$	$+0{,}8$	$1 \to 0{,}8 \to 0{,}64 \to 0{,}512 \to 0{,}41$
$1{,}1$	$-1{,}2$	$1 \to -1{,}2 \to 1{,}44 \to -1{,}73 \to 2{,}07$

Le premier **converge** vers zéro. Le second **oscille en s'éloignant** : la descente diverge.

La condition de convergence, ici

$$|1 - 2\eta| < 1 \qquad \Longleftrightarrow \qquad 0 < \eta < 1$$

Les trois régimes de descente

Combien d'observations par pas

Batch, stochastique, mini-batch

Régime	Observations par pas	Caractère
Batch	toutes	coûteux, trajectoire lisse
Stochastique	une seule	rapide, très bruité
Mini-batch	32 à 256	**le compromis employé partout**

Pourquoi le bruit est utile

On pourrait croire que la descente batch, plus précise, est préférable.

Le bruit du mini-batch aide en réalité à **sortir des minima locaux médiocres** et des plateaux. Il agit comme une régularisation légère.

De plus, un pas approximatif calculé cent fois vaut mieux qu'un pas exact calculé une fois : à temps égal, le mini-batch progresse davantage.

Époque

Définition

Une **époque** est un passage complet sur les données d'apprentissage.

Avec $1\,000$ observations et des lots de 32, une époque compte environ 32 mises à jour.

Les optimiseurs

Trois améliorations successives

Le momentum

On accumule une **inertie** : le pas courant garde une part du précédent.

Cela lisse les oscillations dans les directions instables et accélère dans les directions constantes — comme une bille qui prend de la vitesse dans une vallée.

Le taux adaptatif

RMSProp ajuste le taux **par paramètre**, en divisant par une moyenne des gradients récents.

Les paramètres à gradients forts avancent prudemment, ceux à gradients faibles plus franchement.

Adam

Adam combine les deux : inertie et taux adaptatif par paramètre.

C'est l'optimiseur par défaut, robuste aux réglages approximatifs. On le lance typiquement avec $\eta = 10^{-3}$.

Le choix pratique, en une phrase

Application en gestion

Commencer par **Adam**. Il fonctionne presque toujours correctement du premier coup.

La descente avec momentum, mieux réglée, atteint parfois une meilleure solution finale sur les grands modèles de vision — c'est une optimisation de dernier ressort, pas un point de départ.

Faire décroître le taux

Un taux élevé au début explore ; un taux faible à la fin affine.

Stratégie	Principe
Par paliers	diviser par dix à intervalles fixés
Décroissance cosinus	descente douce et continue
Sur plateau	réduire quand la validation stagne

Ce que le chapitre a établi

Le taux d'apprentissage est le paramètre auquel un réseau est **le plus sensible**. Architecture, activation et optimiseur comptent moins que lui.

Devant un réseau qui n'apprend pas, c'est la première chose à regarder — et souvent la seule à corriger.

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. $w \leftarrow w - \eta\, \partial L/\partial w$: on descend dans le sens opposé au gradient.
2. Sur $L(w) = w^2$: $\eta = 0{,}1$ converge, $\eta = 1{,}1$ **diverge** — la condition est $0 < \eta < 1$.
3. Le **mini-batch** de 32 à 256 est le régime standard ; son bruit est utile.
4. **Adam** — momentum et taux adaptatif — est l'optimiseur par défaut, à 10^{-3}.
5. Le **taux d'apprentissage** est le paramètre le plus sensible de tous.

La suite

Il reste à savoir comment le gradient est calculé pour des millions de paramètres. C'est la rétropropagation, au chapitre 5.