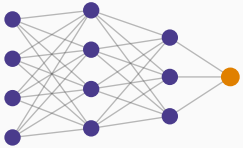
Apprentissage profond

Chapitre 3 : Fonctions de perte et d'activation



Jaouad Madkour

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

Plan du chapitre

La fonction de perte

Les activations

Ce qu'il faut retenir

La fonction de perte

Ce que le réseau cherche à minimiser

Les pertes usuelles

Tâche	Perte	Formule
Régression	erreur quadratique	$\frac{1}{n}\sum(y_i - \hat{y}_i)^2$
Régression robuste	erreur absolue	$\frac{1}{n}\sum\lvert y_i - \hat{y}_i\rvert$
Classification binaire	entropie croisée	$-\frac{1}{n}\sum\left[y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\right]$

L'entropie croisée n'est pas une nouveauté

Minimiser l'entropie croisée revient exactement à **maximiser la vraisemblance** du modèle logistique.

C'est le maximum de vraisemblance du chapitre 8 de *Statistique inférentielle*, sous un autre nom. Le vocabulaire change ; le critère est identique.

Pourquoi l'entropie croisée plutôt que l'erreur quadratique

En classification, l'erreur quadratique produit des gradients qui **s'annulent** quand le réseau se trompe avec assurance — précisément quand il faudrait corriger le plus fort.

L'entropie croisée, au contraire, pénalise très lourdement une prédiction confiante et fausse :

Les activations

La sigmoïde et son défaut

La sigmoïde

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Elle ramène toute valeur dans $]0, 1[$. Sa dérivée vaut $\sigma(z)\,(1 - \sigma(z))$.

Le problème, chiffré

La dérivée de la sigmoïde est **maximale en** $z = 0$, où elle vaut $0{,}25$. Elle est plus faible partout ailleurs.

Or la rétropropagation **multiplie** ces dérivées couche après couche :

Couches	Facteur au mieux
1	$0{,}25$
5	$9{,}8 \times 10^{-4}$
10	$9{,}5 \times 10^{-7}$
20	$9{,}1 \times 10^{-13}$

ReLU, et pourquoi elle a tout changé

ReLU

$$\mathrm{ReLU}(z) = \max(0, z)$$

Sa dérivée vaut 1 pour $z > 0$ et 0 sinon.

La dérivée de 1 est décisive

Multiplier des 1 ne diminue rien. Le gradient traverse donc les couches sans s'atténuer, et l'entraînement de réseaux profonds devient possible.

À cela s'ajoutent deux avantages : le calcul est trivial, et la **parcimonie** — les neurones à sortie nulle ne participent pas, ce qui allège le réseau.

Le passage de la sigmoïde à ReLU est l'une des raisons principales du succès de l'apprentissage profond, davantage qu'une astuce d'implémentation.

Son défaut, et ses variantes

Définition

Un neurone dont l'entrée reste négative a un gradient nul en permanence : il est **mort**, définitivement.

Variante	Correction
Leaky ReLU	pente faible du côté négatif
ELU	transition douce, sortie négative bornée
GELU	version lissée, usuelle dans les transformeurs

Ce qu'il faut faire en pratique

Application en gestion

Commencer par **ReLU** dans les couches cachées. Passer à une variante seulement si l'on observe des neurones morts.

Ne jamais mettre de sigmoïde ou de tangente hyperbolique dans les couches cachées d'un réseau profond — sauf dans les portes des architectures récurrentes du chapitre 9, où leur bornage est justement ce qu'on recherche.

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. La **perte** encode l'objectif : quadratique, absolue, ou entropie croisée.
2. L'**entropie croisée** est le maximum de vraisemblance logistique, sous un autre nom.
3. La dérivée de la sigmoïde plafonne à $0{,}25$: après dix couches, le gradient est divisé par un million.
4. **ReLU** a une dérivée de 1 : le gradient traverse sans s'atténuer.
5. ReLU dans les couches cachées ; sigmoïde ou softmax **uniquement en sortie**.

La suite

Perte et activations sont choisies. Le chapitre 4 aborde l'algorithme qui minimise effectivement cette perte.