

# La descente de gradient

*Comment une IA apprend en descendant une pente*

## Document fondation

Ce document introduit la descente de gradient sur une fonction simple à une variable. C'est l'outil mathématique central que réutilisent tous les autres documents de cette série (régression, classification, réseaux de neurones) pour minimiser une fonction de coût.

**Niveau : Première et Terminale (spécialité mathématiques) — la partie A est accessible dès la Seconde**

## Introduction : pourquoi une IA a-t-elle besoin de « descendre une pente » ?

« Je suis une intelligence artificielle. Pour répondre correctement à tes questions ou résoudre un problème, je dois d'abord apprendre — c'est-à-dire ajuster des paramètres internes jusqu'à ce que mes erreurs soient les plus petites possibles. »

Concrètement, à chaque tâche (reconnaître un chiffre écrit à la main, prédire un mot, estimer un prix...), une IA compare sa prédiction à la bonne réponse et mesure l'écart grâce à une fonction de coût (ou fonction d'erreur). Plus cet écart est grand, plus le coût est élevé.

**Tout l'enjeu de l'apprentissage est donc de trouver les paramètres qui minimisent cette fonction de coût.**

La méthode la plus utilisée pour cela s'appelle la descente de gradient : on part d'un point quelconque, et on se déplace petit à petit dans la direction qui fait diminuer le coût, jusqu'à s'approcher du minimum — un peu comme une bille qui roule au fond d'une vallée.

Dans ce document, on étudie cette méthode sur une fonction simple à une seule variable. Les documents suivants montreront comment le même principe, exactement, permet à une IA d'ajuster les coefficients d'une droite, les poids d'un réseau de neurones, ou les paramètres d'un modèle de langage.

## Partie A — Étudier une fonction simple

**Niveau : Seconde / Première**

On considère la fonction définie sur  $\mathbb{R}$  par :

$$f(x) = x^2 - 4x + 5$$

On cherche à comprendre la forme de cette fonction avant de chercher à en minimiser la valeur — exactement comme une IA doit « visualiser » sa fonction de coût avant de l'optimiser.

1. Complète le tableau de valeurs suivant.

| x    | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
|------|----|---|---|---|---|---|---|
| f(x) | ?  | ? | ? | ? | ? | ? | ? |

2. Trace la courbe représentative de  $f$  sur l'intervalle  $[-1 ; 5]$  (à la main ou avec un outil numérique).
3. Quelle est la forme de cette courbe ? Quel est, d'après le graphique, le minimum de  $f$  et pour quelle valeur de  $x$  est-il atteint ?
4. Vérifie, en développant, que pour tout réel  $x$  :  $f(x) = (x - 2)^2 + 1$ . En déduire, sans utiliser la dérivée, le minimum de  $f$  et la valeur de  $x$  pour laquelle il est atteint.

## Partie B — Une première idée : descendre « à tâtons »

**Niveau : Première**

Imagine une bille posée sur la courbe de  $f$ , au point d'abscisse  $x_0 = 5$ . Elle va rouler vers le bas, en se rapprochant du minimum. On va simuler ce mouvement pas à pas, sans utiliser de dérivée pour l'instant : à chaque étape, on avance de 0,5 dans la direction où  $f$  diminue.

5. En partant de  $x_0 = 5$  avec un pas de 0,5, indique les valeurs successives  $x_1, x_2, x_3, \dots$  en te basant sur le tableau de la partie A (à chaque étape, choisis parmi  $x - 0,5$  et  $x + 0,5$  celle qui donne la plus petite valeur de  $f$ ). À quelle étape peut-on estimer qu'on a atteint le minimum ?

6. Pourquoi cette méthode « à tâtons » n'est-elle pas idéale pour un ordinateur qui doit ajuster des millions de paramètres à la fois ?

## Partie C — La dérivée pour indiquer la direction

### Niveau : Première / Terminale

La dérivée  $f'(x)$  donne la pente de la tangente à la courbe au point d'abscisse  $x$ . Elle nous indique, bien mieux qu'un tâtonnement, dans quelle direction et à quelle vitesse la fonction varie.

$$f'(x) = 2x - 4$$

7. Résous l'équation  $f'(x) = 0$ . Que représente géométriquement cette solution pour la courbe de  $f$  ? Est-ce cohérent avec la question A.3 ?

On définit alors la règle de mise à jour de la descente de gradient :

#### Formule de la descente de gradient (une variable)

$$x_{n+1} = x_n - \alpha \times f'(x_n)$$

où  $\alpha$  (« alpha ») est le pas d'apprentissage (ou taux d'apprentissage) : un nombre réel strictement positif, en général petit, qui fixe la taille de chaque déplacement.

8. On choisit  $x_0 = 5$  et  $\alpha = 0,1$ . Calcule  $x_1$ ,  $x_2$  et  $x_3$  en utilisant la formule ci-dessus (arrondis à  $10^{-3}$  près). Complète le tableau :

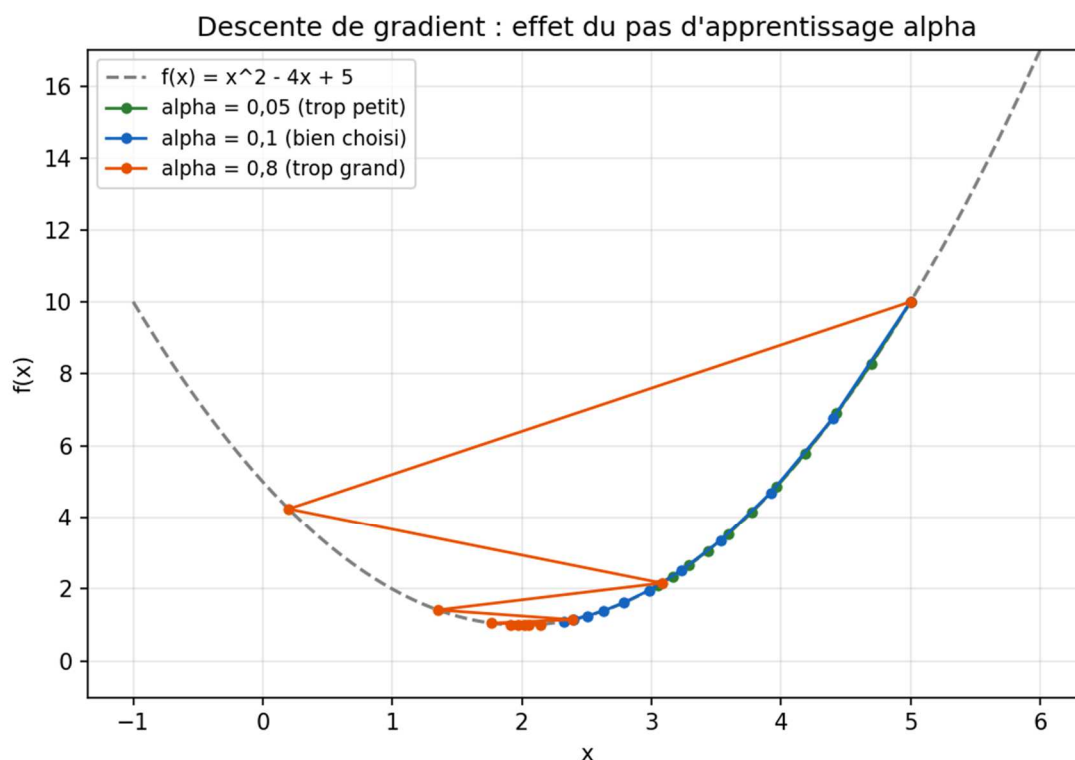
| n | $x_n$ | $f'(x_n) = 2x_n - 4$ | $x_{n+1} = x_n - 0,1 \times f'(x_n)$ |
|---|-------|----------------------|--------------------------------------|
| 0 | 5     | ...                  | ...                                  |
| 1 | ...   | ...                  | ...                                  |
| 2 | ...   | ...                  | ...                                  |

9. Que remarques-tu sur la suite  $(x_n)$  ? Vers quelle valeur semble-t-elle converger ? Est-ce cohérent avec les parties A et C.1 ?

## Partie D — Bien choisir son pas d'apprentissage

### Niveau : Terminale

Le graphique ci-dessous compare, à partir du même point de départ  $x_0 = 5$ , l'évolution de la suite  $(x_n, f(x_n))$  pour trois valeurs du pas  $\alpha$  : 0,05 ; 0,1 et 0,8.



10. Pour  $\alpha = 0,8$ , calcule à la main  $x_1$  et  $x_2$  à partir de  $x_0 = 5$ , puis compare avec le graphique. Que se passe-t-il ?
11. Analyse le graphique et réponds :
  - a) Quelle courbe converge le plus efficacement vers le minimum ?
  - b) Pourquoi la courbe orange ( $\alpha = 0,8$ ) oscille-t-elle autour du minimum au lieu de s'en approcher directement ?
  - c) Quel est l'inconvénient d'un pas trop petit ( $\alpha = 0,05$ ), même s'il ne pose pas de problème de divergence ?
12. Pourquoi appelle-t-on  $\alpha$  le « pas d'apprentissage » ? Propose une analogie avec un élève qui révise un cours (trop vite, trop lentement, ou à un bon rythme).

## Partie E — Généraliser : de $x$ à plusieurs paramètres

### Niveau : Terminale

Ici, on a ajusté un seul nombre  $x$ . Mais une IA doit en général ajuster plusieurs paramètres à la fois : par exemple les coefficients  $a$  et  $b$  d'une droite  $y = ax + b$ , ou les milliers (voire milliards) de poids d'un réseau de neurones.

Le principe reste rigoureusement le même : on remplace la dérivée par les dérivées partielles par rapport à chaque paramètre, et on met à jour chaque paramètre en parallèle :

#### Formule générale (plusieurs paramètres)

$$\mathbf{a} \leftarrow \mathbf{a} - \alpha \times \partial J / \partial \mathbf{a} \quad \mathbf{b} \leftarrow \mathbf{b} - \alpha \times \partial J / \partial \mathbf{b} \quad \dots$$

où  $J$  est la fonction de coût à minimiser (elle joue le rôle de  $f$ ), et  $\partial J / \partial a$ ,  $\partial J / \partial b$ ... sont les dérivées partielles de  $J$  par rapport à chacun des paramètres — c'est-à-dire la dérivée de  $J$  lorsque l'on ne fait varier qu'un seul paramètre à la fois, les autres étant fixés.

C'est exactement cette formule, appliquée à une droite de séparation puis à un modèle de classification, que tu retrouveras dans les documents suivants de cette série.

## Partie F — Ce que fait réellement une IA

### Niveau : Tous niveaux

Une IA comme un modèle de langage n'ajuste pas une seule variable  $x$ , mais des milliards de paramètres, en suivant exactement la même logique que dans ce document :

- 1. elle définit une fonction de coût  $J$  qui mesure son erreur (par exemple l'erreur quadratique moyenne pour prédire un nombre, ou l'entropie croisée pour prédire une catégorie ou un mot) ;

- 2. elle calcule le gradient de  $J$ , c'est-à-dire l'ensemble des dérivées partielles par rapport à chaque paramètre (cette étape, appelée rétropropagation dans un réseau de neurones, revient à appliquer très rapidement la règle de dérivation des fonctions composées à des millions de paramètres) ;
- 3. elle met à jour tous les paramètres simultanément avec la règle vue en partie E, avec un pas d'apprentissage  $\alpha$  choisi avec soin (ni trop grand, ni trop petit) ;
- 4. elle répète ces trois étapes des milliers, voire des millions de fois, sur d'énormes quantités de données, jusqu'à ce que le coût devienne suffisamment petit.

Ce document t'a fait vivre, sur un exemple à une seule variable qu'on peut représenter par une simple parabole, exactement ce que fait une IA à chaque instant de son entraînement.

## Exercices d'entraînement

### Exercice 1 — Application directe

#### Niveau : Première / Terminale

On considère  $g(x) = x^2 - 6x + 10$ .

- Écris  $g(x)$  sous forme canonique. En déduire le minimum de  $g$  et la valeur de  $x$  pour laquelle il est atteint.
- Calcule  $g'(x)$ .
- On part de  $x_0 = 6$  avec un pas  $\alpha = 0,2$ . Calcule  $x_1$ ,  $x_2$  et  $x_3$  à l'aide de la formule de la descente de gradient (arrondis à  $10^{-3}$  près).
- Vers quelle valeur la suite  $(x_n)$  semble-t-elle converger ? Est-ce cohérent avec la question 1 ?

### Exercice 2 — Un pas trop grand

#### Niveau : Terminale

On reprend  $g(x) = x^2 - 6x + 10$  et  $g'(x) = 2x - 6$ , avec cette fois  $x_0 = 6$  et  $\alpha = 1,5$ .

- Calcule  $x_1$  et  $x_2$ .
- Que remarques-tu sur le comportement de la suite  $(x_n)$  ? Compare avec le cas  $\alpha = 0,2$  de l'exercice 1.
- Propose une explication mathématique (à partir de l'expression de  $x_{n+1}$  en fonction de  $x_n$ ) de ce qui différencie un pas qui fait converger la suite d'un pas qui la fait diverger.

### Exercice 3 — Lien avec l'intelligence artificielle

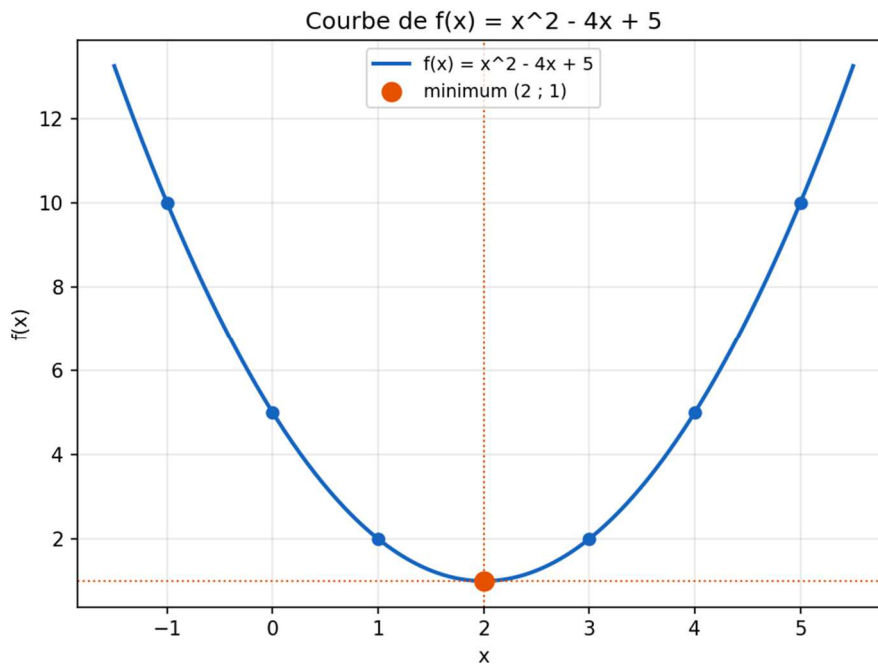
#### Niveau : Tous niveaux

- Une IA doit reconnaître si une image représente un chat ou un chien. Elle utilise une fonction de coût  $J$  qui dépend de plusieurs milliers de paramètres. Pourquoi ne peut-elle pas, comme dans la partie B, tester « à tâtons » toutes les directions possibles pour trouver le minimum de  $J$  ?
- Explique, avec tes mots, pourquoi un pas d'apprentissage  $\alpha$  mal choisi peut empêcher une IA d'apprendre correctement, en t'appuyant sur ce que tu as observé en partie D.
- En pratique, les ingénieurs commencent souvent l'entraînement d'une IA avec un pas d'apprentissage relativement grand, puis le diminuent progressivement au fil de l'entraînement. Propose une explication de cette stratégie à partir de ce que tu as observé dans ce document.

## Corrigé

### Partie A

| x    | -1 | 0 | 1 | 2 | 3 | 4 | 5  |
|------|----|---|---|---|---|---|----|
| f(x) | 10 | 5 | 2 | 1 | 2 | 5 | 10 |



A.2-3. La courbe est une parabole tournée vers le haut, symétrique par rapport à la droite  $x = 2$ . Le minimum semble être atteint en  $x = 2$ , avec  $f(2) = 1$ .

A.4. En développant  $(x - 2)^2 + 1 = x^2 - 4x + 4 + 1 = x^2 - 4x + 5 = f(x)$ . Comme  $(x - 2)^2 \geq 0$  pour tout réel  $x$ , on a  $f(x) \geq 1$ , avec égalité si et seulement si  $x = 2$ . Le minimum de  $f$  est donc 1, atteint en  $x = 2$  — ce qui confirme la lecture graphique, sans avoir eu besoin de dérivée.

## Partie B

B.1. En s'appuyant sur le tableau :  $f(4,5) < f(5)$ , on avance donc vers  $x = 4,5$  ; puis  $f(4) < f(4,5)$ , on continue vers  $x = 4$  ; puis  $f(3,5) < f(4)$ , etc. On se rapproche progressivement de  $x = 2$ , où  $f$  atteint son minimum. On peut estimer qu'on approche suffisamment le minimum dès que la valeur de  $f$  ne diminue quasiment plus d'une étape à l'autre.

B.2. Cette méthode nécessite de recalculer  $f$  en plusieurs points à chaque étape pour savoir dans quelle direction avancer. Avec un seul paramètre cela reste possible, mais avec des millions de paramètres, tester toutes les directions deviendrait bien trop coûteux en temps de calcul. La dérivée (partie C) donne directement la bonne direction, sans avoir besoin de tester plusieurs points.

## Partie C

C.1.  $f'(x) = 0 \Leftrightarrow 2x - 4 = 0 \Leftrightarrow x = 2$ . Cela signifie que la tangente à la courbe de  $f$  est horizontale au point d'abscisse 2 : c'est la caractérisation d'un minimum (ou d'un maximum) local pour une fonction dérivable. C'est bien cohérent avec la partie A.

| n | $x_n$ | $f'(x_n) = 2x_n - 4$ | $x_{n+1} = x_n - 0,1 \times f'(x_n)$ |
|---|-------|----------------------|--------------------------------------|
| 0 | 5     | 6                    | 4,4                                  |
| 1 | 4,4   | 4,8                  | 3,92                                 |
| 2 | 3,92  | 3,84                 | 3,536                                |

C.3. La suite  $(x_n)$  est strictement décroissante et semble se rapprocher progressivement de 2, qui est justement la valeur où  $f$  atteint son minimum (voir C.1 et A.4). Plus  $x_n$  se rapproche de 2, plus  $f'(x_n)$  se rapproche de 0, donc plus les déplacements deviennent petits : la suite ralentit en approchant du minimum, ce qui est cohérent avec le comportement observé sur le graphique de la partie D.

## Partie D

D.1. Avec  $\alpha = 0,8$  :  $x_1 = 5 - 0,8 \times f'(5) = 5 - 0,8 \times 6 = 5 - 4,8 = 0,2$ . Puis  $f'(0,2) = 2 \times 0,2 - 4 = -3,6$ , donc  $x_2 = 0,2 - 0,8 \times (-3,6) = 0,2 + 2,88 = 3,08$ . On observe bien, comme sur le graphique, que la suite « saute » d'un côté à l'autre du minimum au lieu de s'en approcher progressivement.

D.2. a) Le pas  $\alpha = 0,1$  est le plus efficace ici : il converge rapidement vers  $x = 2$  sans osciller. b) Avec  $\alpha = 0,8$ , chaque déplacement est trop grand : on dépasse le minimum et on se retrouve de l'autre côté de la parabole, ce qui crée un mouvement de va-et-vient (oscillation) autour du minimum. c) Avec  $\alpha = 0,05$ , la convergence est certes régulière et sans

oscillation, mais chaque pas est si petit qu'il faut beaucoup plus d'itérations pour s'approcher du minimum : l'apprentissage est très lent, ce qui coûte du temps de calcul.

D.3. On l'appelle « pas d'apprentissage » car il fixe la vitesse à laquelle le modèle progresse vers la solution, exactement comme le rythme de révision d'un élève : réviser trop vite et de façon trop « brutale » ( $\alpha$  trop grand) peut faire passer à côté de points importants sans les ancrer (oscillations, confusion) ; réviser trop prudemment ( $\alpha$  trop petit) permet d'ancrer chaque notion mais prend un temps très long pour progresser.

## Partie E et F

Ces parties sont surtout des parties de cours ; elles n'appellent pas de calculs supplémentaires. Elles seront directement réinvesties, avec des exemples chiffrés, dans les documents consacrés à la régression linéaire et à la classification.

### Exercice 1

- $g(x) = x^2 - 6x + 10 = (x - 3)^2 + 1$ . Le minimum de  $g$  est 1, atteint en  $x = 3$ .
- $g'(x) = 2x - 6$ .

| n | $x_n$ | $g'(x_n) = 2x_n - 6$ | $x_{n+1} = x_n - 0,2 \times g'(x_n)$ |
|---|-------|----------------------|--------------------------------------|
| 0 | 6     | 6                    | 4,8                                  |
| 1 | 4,8   | 3,6                  | 4,08                                 |
| 2 | 4,08  | 2,16                 | 3,648                                |

- La suite  $(x_n)$  décroît et se rapproche progressivement de 3, qui est bien la valeur trouvée à la question 1.

### Exercice 2

- $x_1 = 6 - 1,5 \times g'(6) = 6 - 1,5 \times 6 = 6 - 9 = -3$ . Puis  $g'(-3) = 2 \times (-3) - 6 = -12$ , donc  $x_2 = -3 - 1,5 \times (-12) = -3 + 18 = 15$ .
- Contrairement à l'exercice 1, la suite s'éloigne de plus en plus de  $x = 3$  : elle diverge, au lieu de converger. Un pas trop grand peut donc empêcher totalement l'algorithme de trouver le minimum, et même faire « exploser » les valeurs.
- En notant que  $g'(x) = 2x - 6$ , la formule de mise à jour s'écrit  $x_{n+1} = x_n - \alpha(2x_n - 6) = (1 - 2\alpha)x_n + 6\alpha$ . Le comportement de la suite dépend du coefficient  $(1 - 2\alpha)$  : si  $|1 - 2\alpha| < 1$  (soit  $0 < \alpha < 1$ ), la suite converge vers 3 ; si  $|1 - 2\alpha| \geq 1$  ( $\alpha \geq 1$ , comme ici avec  $\alpha = 1,5$ ), la suite s'éloigne de 3 au lieu de s'en rapprocher. Un pas d'apprentissage doit donc rester suffisamment petit pour garantir la convergence.

### Exercice 3

- Avec plusieurs milliers de paramètres, il faudrait tester un nombre de combinaisons de directions gigantesque (qui croît de façon exponentielle avec le nombre de paramètres) : c'est totalement impossible en un temps raisonnable, même pour un ordinateur puissant. La dérivée (ou plutôt le gradient, en partie E) donne directement, en un seul calcul, la meilleure direction locale pour tous les paramètres à la fois.
- Comme observé en partie D, un pas trop grand fait « sauter » les paramètres autour du minimum sans jamais s'en approcher précisément (voire les fait diverger, comme dans l'exercice 2) : l'IA n'apprend pas correctement, son erreur ne diminue pas ou augmente. Un pas trop petit fait progresser l'IA très lentement : l'entraînement devient extrêmement long et coûteux, sans réel gain de précision supplémentaire.
- Un grand pas au début permet de progresser rapidement lorsque l'on est encore loin du minimum (comme  $x_0 = 5$  loin de  $x = 2$ ) : les grands déplacements ne posent alors pas de problème car on est loin de la zone sensible. En diminuant progressivement le pas à mesure que l'on se rapproche du minimum, on évite les oscillations ou la divergence observées en partie D et en exercice 2, tout en conservant une convergence rapide au départ. Cette stratégie s'appelle la décroissance du pas d'apprentissage (learning rate decay), très utilisée dans l'entraînement réel des IA.

## Mathématiques du programme mobilisées dans ce document

### Seconde

Fonctions du second degré : tableau de valeurs, représentation graphique.

### Première (spécialité mathématiques)

Forme canonique d'un trinôme du second degré, minimum d'une fonction polynôme.

Dérivation : nombre dérivé, fonction dérivée d'une fonction polynôme, tangente.

Suites numériques définies par récurrence.

**Terminale (spécialité mathématiques)**

Étude du sens de variation d'une fonction à l'aide du signe de la dérivée.

Suites définies par récurrence : calcul de termes, étude de la convergence.

Notion de dérivée partielle et de gradient (généralisation à plusieurs variables), en lien avec les programmes d'informatique et de spécialité NSI / mathématiques expertes.