Limites de suites
Problème de recherche : le réglage d'une intelligence artificielle

« Entraîner » une intelligence artificielle, c'est chercher les réglages qui rendent son erreur la plus petite possible. Pour cela, les ingénieurs utilisent presque toujours le même algorithme : la *descente de gradient*.
Dans une version simplifiée, l'erreur de l'IA dépend d'un seul réglage x et vaut

$$f(x) = 2(x-3)^2 + 1.$$

L'algorithme ne « voit » pas la courbe : il part d'un réglage $x_0 = 0$, puis, à chaque étape, il fait un petit pas dans le sens où l'erreur diminue, selon la règle

$$x_{n+1} = x_n - \eta\, f'(x_n),$$

où η (la lettre grecque « êta ») est un nombre strictement positif fixé à l'avance, appelé *taux d'apprentissage*.

La mission

1. Pour quelles valeurs de η la suite (x_n) converge-t-elle vers le meilleur réglage ?
2. Quelle valeur de η conseiller à l'équipe ? Que se passe-t-il pour les autres valeurs ?
3. Avec $\eta = 0{,}1$, à partir de quelle étape le réglage est-il à moins de 10^{-6} du meilleur réglage ?

Pour aller plus loin

Une IA plus réaliste a deux réglages x et y, et son erreur vaut $g(x\,;y) = x^2 + 10y^2$. L'algorithme part de $(x_0\,;y_0) = (1\,;1)$ et met à jour les deux réglages *avec le même taux* η :

$$x_{n+1} = x_n - \eta \times 2x_n \qquad \text{et} \qquad y_{n+1} = y_n - \eta \times 20y_n.$$

Pour quelles valeurs de η les deux réglages convergent-ils vers 0 ? Quelle valeur de η fait converger le plus vite *les deux à la fois* ?

Coups de pouce

Bloqué ? Des indices progressifs sont disponibles en ligne. N'en ouvrir qu'un à la fois, et seulement après avoir cherché !

Éléments de correction

D'autres démarches sont possibles (tableur, programme Python) : ce qui compte, c'est de justifier les conclusions.

Explorer d'abord

Avec $\eta = 0{,}1$, on obtient $x_1 = 1{,}2$; $x_2 = 1{,}92$; $x_3 \approx 2{,}35\ldots$: la suite semble croître vers 3. Avec $\eta = 0{,}4$: $x_1 = 4{,}8$; $x_2 = 1{,}92$; $x_3 \approx 3{,}65\ldots$: elle oscille autour de 3 en s'en rapprochant. Avec $\eta = 0{,}6$: 0 ; 7,2 ; −2,88 ; 11,23… : elle s'emballe.

1. Convergence

Le meilleur réglage est $x = 3$ (minimum de f). On a $f'(x) = 4(x-3)$, donc

$$x_{n+1} = x_n - 4\eta(x_n - 3).$$

On pose $d_n = x_n - 3$ (l'écart au meilleur réglage). Alors $d_{n+1} = x_{n+1} - 3 = (x_n - 3) - 4\eta(x_n - 3) = (1 - 4\eta)\, d_n$.

La suite (d_n) est géométrique de raison $q = 1 - 4\eta$ et de premier terme $d_0 = -3$: $d_n = -3(1-4\eta)^n$.

(x_n) converge vers 3 si et seulement si (d_n) converge vers 0, c'est-à-dire si et seulement si $-1 < q < 1$:

$$-1 < 1 - 4\eta < 1 \iff 0 < \eta < \frac{1}{2}.$$

L'algorithme trouve le meilleur réglage si et seulement si $0 < \eta < 0{,}5$**.**

2. Le bon réglage et les autres cas

- $\eta = 0{,}25$: $q = 0$, donc $x_1 = 3$. Le minimum est atteint **en une seule étape** : c'est la valeur idéale.
- $0 < \eta < 0{,}25$: $0 < q < 1$, la suite croît vers 3 ; d'autant plus lentement que η est petit.
- $0{,}25 < \eta < 0{,}5$: $-1 < q < 0$, la suite oscille autour de 3 en convergeant.
- $\eta = 0{,}5$: $q = -1$, la suite alterne entre 0 et 6 sans jamais converger.
- $\eta > 0{,}5$: $q < -1$, les écarts $|d_n|$ tendent vers $+\infty$: l'algorithme diverge.

En pratique, l'algorithme ne connaît pas la courbe (ici le coefficient 4) : on ne peut pas calculer le η idéal à l'avance. C'est pourquoi les ingénieurs testent plusieurs valeurs.

3. Seuil avec $\eta = 0{,}1$

$|d_n| = 3 \times 0{,}6^n$. On cherche le premier n tel que $3 \times 0{,}6^n < 10^{-6}$. La suite $(0{,}6^n)$ est décroissante ; un algorithme de seuil (ou le tableur) donne $3 \times 0{,}6^{29} \approx 1{,}1 \times 10^{-6}$ et $3 \times 0{,}6^{30} \approx 6{,}6 \times 10^{-7}$. **Il faut 30 étapes.**

Pour aller plus loin

$x_n = (1-2\eta)^n$ et $y_n = (1-20\eta)^n$: deux suites géométriques.

— Convergence de (x_n) : $0 < \eta < 1$. Convergence de (y_n) : $0 < \eta < 0{,}1$. Les deux convergent si et seulement si $0 < \eta < 0{,}1$: c'est la direction la plus « pentue » qui impose sa limite.

— La vitesse est fixée par la plus mauvaise des deux raisons (en valeur absolue). Pour $\eta \leqslant 0{,}05$, c'est $1-2\eta$, qui diminue quand η augmente ; pour $\eta \geqslant 0{,}05$, $|1-20\eta| = 20\eta - 1$ finit par l'emporter. Le meilleur compromis est obtenu quand $1-2\eta = 20\eta - 1$, soit $\eta = \frac{1}{11}$: les deux écarts sont alors multipliés par au plus $\frac{9}{11} \approx 0{,}82$ à chaque étape.

Bilan : avec un seul taux, on ne peut pas être rapide dans les deux directions à la fois.

Ce qu'on a utilisé du chapitre

— Suite auxiliaire géométrique (suite arithmético-géométrique).

— Limite de q^n selon les valeurs de q ; convergence et divergence.

— Algorithme de seuil.

Vers le supérieur (enrichissement, hors programme)

En prépa ou en école d'ingénieurs, f' est remplacée par le *gradient*, le vecteur des dérivées partielles. Le rapport $\frac{20}{2} = 10$ entre les courbures des deux directions s'appelle le *conditionnement* : plus il est grand, plus la descente de gradient est lente. Les algorithmes utilisés pour entraîner les grands modèles d'IA (Adam, par exemple) adaptent le taux à chaque direction pour contourner ce problème.