ADAM : Une méthode pour l'optimisation stocchastique

Guillaume BERNARD-REYMOND, Guillaume BOULAND,
Camille MOTTIER, Abel SILLY

Octobre 2024

L'algorithme

Entrées : $f(\theta)$, α, β_1, β_2, ε, θ_0

$m_0 \longleftarrow 0$

$v_0 \longleftarrow 0$

$t \longleftarrow 0$

Tant que θ_t ne converge pas

$t \longleftarrow t + 1$

$g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$

$m_t \longleftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

$v_t \longleftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$

$\widehat{m}_t \longleftarrow m_t / (1 - \beta_1^t)$

$\widehat{v}_t \longleftarrow v_t / (1 - \beta_2^t)$

$\theta_t \longleftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \varepsilon)$

Sorties : θ_t

Explications

$f_1, \ldots, f_T$: fonction f restreinte à des mini-batches aléatoires

Explications

$f_1, \ldots, f_T$: fonction f restreinte à des mini-batches aléatoires
$g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$: mise à jour du gradient

Explications

$f_1, \ldots, f_T$: fonction f restreinte à des mini-batches aléatoires
$g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$: mise à jour du gradient

$$\left.\begin{array}{l} m_t \longleftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t \\ v_t \longleftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2)\cdot g_t^2 \end{array}\right\} : \text{mise à jour des moments}$$

d'ordre 1 et 2 par moyenne mobile

Explications

$f_1, \ldots, f_T$: fonction f restreinte à des mini-batches aléatoires
$g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$: mise à jour du gradient
$$\left.\begin{array}{l} m_t \longleftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \\ v_t \longleftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \end{array}\right\} \text{ : mise à jour des moments}$$
d'ordre 1 et 2 par moyenne mobile
$$\left.\begin{array}{l} \widehat{m}_t \longleftarrow m_t/(1-\beta_1^t) \\ \widehat{v}_t \longleftarrow v_t/(1-\beta_2^t) \end{array}\right\} \text{ : débiaisage des moments d'ordre 1 et 2}$$

Explications

$f_1, \dots, f_T$: fonction f restreinte à des mini-batches aléatoires
$g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$: mise à jour du gradient

$$\left.\begin{array}{l} m_t \longleftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t \\ v_t \longleftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2)\cdot g_t^2 \end{array}\right\} \text{ : mise à jour des moments}$$

d'ordre 1 et 2 par moyenne mobile

$$\left.\begin{array}{l} \widehat{m}_t \longleftarrow m_t/(1-\beta_1^t) \\ \widehat{v}_t \longleftarrow v_t/(1-\beta_2^t) \end{array}\right\} \text{ : débiaisage des moments d'ordre 1 et 2}$$

$\theta_t \longleftarrow \theta_{t-1} - \alpha\cdot \widehat{m}_t/(\sqrt{\widehat{v}_t} + \varepsilon)$: Mise à jour de θ_t

Résultats de convergence

On définit le regret par :

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x)$$

Theorem

Sous certaines hypothèses de majoration du gradient de f_t, et de l'écart entre les valeurs de θ_n on a :

$$\frac{R(T)}{T} = O\left(\frac{1}{\sqrt{T}}\right)$$

Simulations numériques

Mettre images et commenter