

ADAM : Une méthode pour l'optimisation stochastique

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

14 octobre 2024

L'algorithme

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

$$m_0 \leftarrow 0$$

$$v_0 \leftarrow 0$$

$$t \leftarrow 0$$

Tant que θ_t ne converge pas

$$t \leftarrow t + 1$$

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

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

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

$$\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$$

$$\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$$

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

Sorties : θ_t

Explications

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

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

Explications

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

$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$: mise à jour du gradient

$\left. \begin{array}{l} m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \\ v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \end{array} \right\}$: mise à jour des moments d'ordre 1 et 2 par
moyenne mobile

Explications

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

$\mathbf{g}_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$: mise à jour du gradient

$\left. \begin{array}{l} \mathbf{m}_t \leftarrow \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t \\ \mathbf{v}_t \leftarrow \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot \mathbf{g}_t^2 \end{array} \right\}$: mise à jour des moments d'ordre 1 et 2 par

moyenne mobile

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

Explications

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

$\mathbf{g}_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$: mise à jour du gradient

$\left. \begin{array}{l} \mathbf{m}_t \leftarrow \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t \\ \mathbf{v}_t \leftarrow \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot \mathbf{g}_t^2 \end{array} \right\}$: mise à jour des moments d'ordre 1 et 2 par

moyenne mobile

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

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \varepsilon)$: 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