ADAM: A method for Stochastic Optimization
Résumé d'article

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

26 septembre 2024

1 Introduction

L'article étudié ici, intitulé « ADAM : A method for Stochastic Optimization », a été écrit en 2015 par Diederik P. Kingma (Université d'Amsterdam, OpenAI) et Jimmy Lei Ba (Université de Toronto), dans le cadre d'une conférence à l'International Conference on Learning Representations (ICLR).

Cet article présente l'algorithme Adam, algorithme d'optimisation stochastique, basée sur une descente de gradient, dans le cadre d'un espace de paramètres à grande dimension. Outre le fait que cet algorithme est simple à implémenter, efficace computationnellement et nécessite peu de mémoire, il offre une méthode qui marche bien dans un large panel de cas, y compris dans les cas problématiques de gradients éparses ou de fonctions-objectifs non stationnaires. En cela, il combine les qualités d'algorithmes existants au préalable, tels que AdaGrad et RMSProp.

Cet article présente une description précise de l'algorithme Adam, fournit un résultat de convergence de la méthode et détaille l'apport de l'algorithme Adam vis-à-vis d'autres algorithmes.

2 Algorithme

On considère une fonction-objectif stochastique $f(\theta)$ de paramètres θ, qu'on suppose différentiable. On souhaite optimiser les paramètres θ afin de minimiser l'espérance $\mathbb{E}[f(\theta)]$.
Outre la fonction $f(\theta)$, l'algorithme nécessite la donnée d'un pas α, de taux $\beta_1, \beta_2 \in [0, 1[$, d'une constante $\varepsilon > 0$ et de paramètres initiaux θ_0. Il exécute alors le schéma suivant :

Algorithme 1 : Adam

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* **faire**
  $t \longleftarrow t + 1$
  $g_t \longleftarrow \nabla_\theta f_t(\theta_{t-1})$
  $m_t \longleftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
  $v_t \longleftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
  $\widehat{m}_t \longleftarrow m_t / (1 - \beta_1^t)$
  $\widehat{v}_t \longleftarrow v_t / (1 - \beta_2^t)$
  $\theta_t \longleftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \varepsilon)$
fin
Sorties : θ_t

3 Efficacité et points forts de la méthode

4 Amélioration des méthodes AdaGrad et RMSProp

+ visu