

ADAM : Une méthode pour l'optimisation stochastique

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

14 octobre 2024

L'article

Adam : A Method for Stochastic Optimization

Diederik P. Kingma

Université d'Amsterdam, OpenAI

Jimmy Lei Ba

Université de Toronto

Première année de publication : 2014



Objectif de l'article : introduire un algorithme d'optimisation stochastique

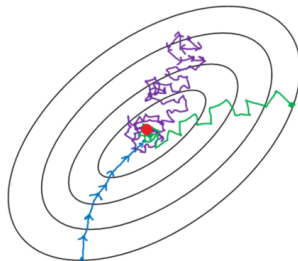
L'objectif

$f(\theta)$: fonction-objectif stochastique

Exemple :
$$f(\theta) = \sum_{i=1}^n L(x_i|\theta)$$

Mini-batch :
$$f_t(\theta) = \sum_{i \in I_t} L(x_i|\theta)$$

Objectif : minimiser $\mathbb{E}[f(\theta)]$



- Batch gradient descent
- Mini-batch gradient Descent
- Stochastic gradient descent

Méthode d'ordre 1

Qu'est-ce qu'une méthode d'ordre 1 ? \longrightarrow Évaluation de $f(\theta)$ et de $\nabla_{\theta}f(\theta)$

Algorithme : Descente de gradient stochastique

Entrées : $f(\theta)$, α , θ_0

$t \longleftarrow 0$

tant que θ_t ne converge pas **faire**

$t \longleftarrow t + 1$

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

$\theta_t \longleftarrow \theta_{t-1} - \alpha \cdot g_t(\theta_{t-1})$

fin

Sorties : θ_t

L'algorithme Adam

Algorithme : Adam

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* **faire**

$t \leftarrow t + 1$

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

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Estimateur de $\mathbb{E}[g_t]$)

fin

Sorties : θ_t

Illustration Guibou !

L'algorithme Adam

Algorithme : Adam

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* **faire**

$t \leftarrow t + 1$

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

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Estimateur de $\mathbb{E}[g_t]$)

fin

Sorties : θ_t

L'algorithme Adam

Algorithme : Adam

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 **faire**

$t \leftarrow t + 1$

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

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

(Estimateur de $\mathbb{E}[g_t]$)

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

(Estimateur de $\mathbb{E}[g_t^2]$)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot m_t / (\sqrt{v_t} + \varepsilon)$

(Mise à jour)

fin

Sorties : θ_t

Illustration extraite de la vidéo !!!!!!!

L'algorithme Adam

Algorithme : Adam

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 **faire**

$t \leftarrow t + 1$

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

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

(Estimateur de $\mathbb{E}[g_t]$)

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

(Estimateur de $\mathbb{E}[g_t^2]$)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot m_t / (\sqrt{v_t} + \varepsilon)$

(Mise à jour)

fin

Sorties : θ_t

L'algorithme Adam

Algorithme : Adam

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 faire

$t \leftarrow t + 1$

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

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

(Estimateur de $\mathbb{E}[g_t]$)

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

(Estimateur de $\mathbb{E}[g_t^2]$)

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

(« débiaisage »)

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

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

(Mise à jour)

fin

Sorties : θ_t

Correction de biais

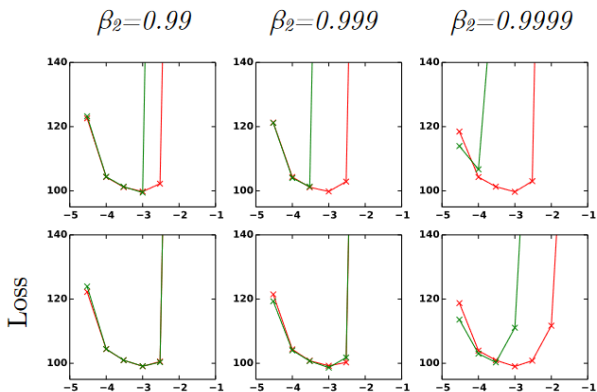
$$m_t = (1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \cdot g_i$$

$$\mathbb{E}[m_t] = \mathbb{E}[g_t] \underbrace{(1 - \beta_1^t)}_{<1} + \zeta$$

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

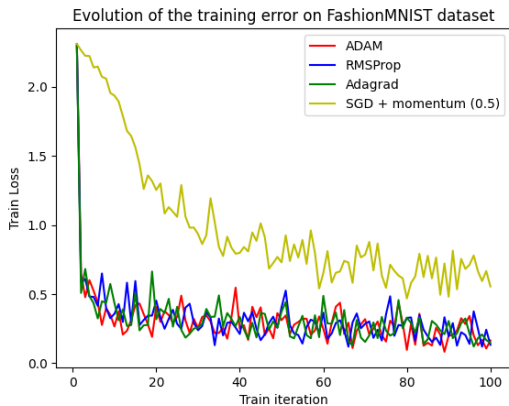
$\beta_1=0$

$\beta_1=0.9$

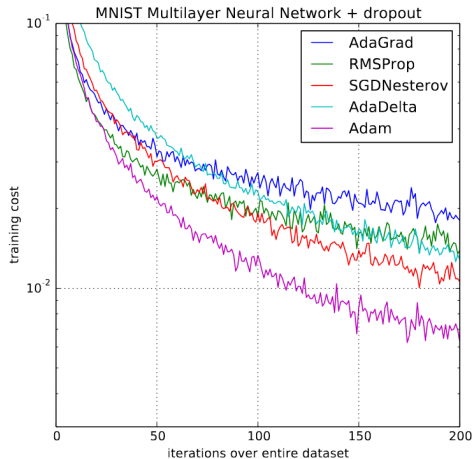
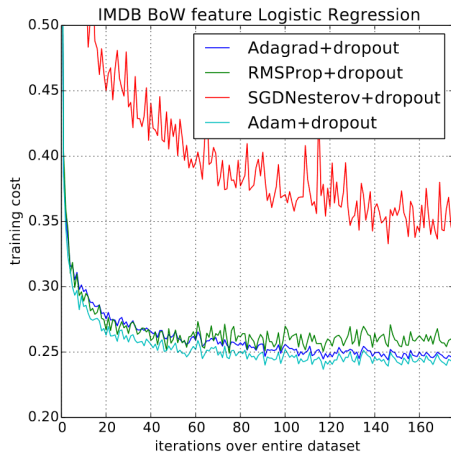


after 100 epochs

Simulations numériques



Simulations numériques



Résultat de convergence

On définit le regret par :

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

Théorème

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

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

Conclusion

ADAM est un algorithme :

- facilement implémentable
- peu gourmand en mémoire
- efficace dans de nombreux cas

D'où son succès !

Limite de l'article : manque de support théorique

MERCI !