

# 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



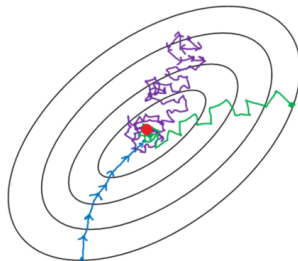
# L'objectif

$f(\theta)$  : 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

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

---

**Algorithme 1** : Descente de gradient stochastique

---

**Entrées** :  $f(\theta)$ ,  $\alpha$ ,  $\theta_0$

$t \leftarrow 0$

**tant que**  $\theta_t$  ne converge pas **faire**

$t \leftarrow t + 1$

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

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

**fin**

**Sorties** :  $\theta_t$

---

# L'algorithme d'ADAM

---

## Algorithme 2 : Adam

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\varepsilon$ ,  $\theta_0$

$m_0 \leftarrow 0$

$v_0 \leftarrow 0$

$t \leftarrow 0$

**tant que**  $\theta_t$  *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 :**  $\theta_t$

---

Illustration Guibou !

# L'algorithme d'ADAM

---

## Algorithme 3 : Adam

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\varepsilon$ ,  $\theta_0$

$m_0 \leftarrow 0$

$v_0 \leftarrow 0$

$t \leftarrow 0$

**tant que**  $\theta_t$  *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  $\mathbb{E}[g_t^2]$  )

**fin**

**Sorties :**  $\theta_t$

---

# L'algorithme d'ADAM

---

## Algorithme 4 : Adam

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\varepsilon$ ,  $\theta_0$

$m_0 \leftarrow 0$

$v_0 \leftarrow 0$

$t \leftarrow 0$

**tant que**  $\theta_t$  *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 » )

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

**fin**

**Sorties :**  $\theta_t$

---



# L'algorithme d'ADAM

---

## Algorithme 5 : Adam

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\varepsilon$ ,  $\theta_0$

$m_0 \leftarrow 0$

$v_0 \leftarrow 0$

$t \leftarrow 0$

**tant que**  $\theta_t$  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 » )

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

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

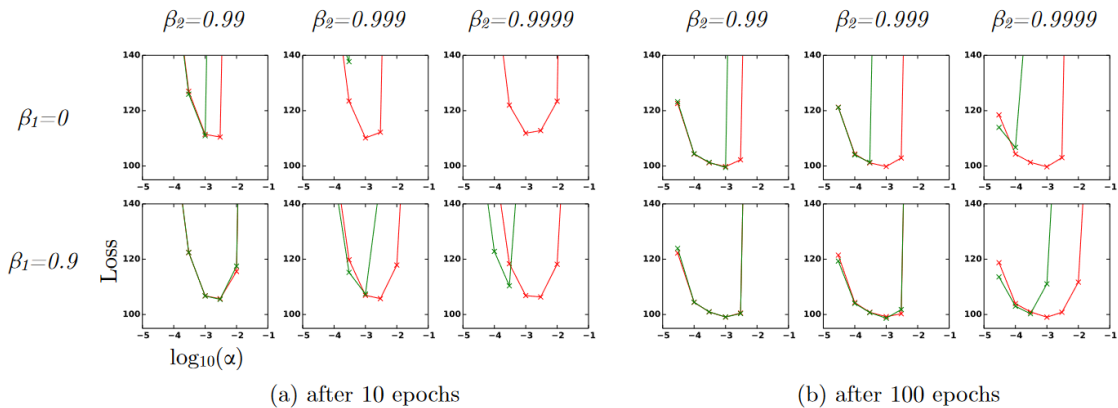
( Mise à jour )

**fin**

**Sorties :**  $\theta_t$

---

# Biais



mettre que pour 100 epochs ? rouge/vert : peu visible

# Résultats de convergence

On définit le regret par :

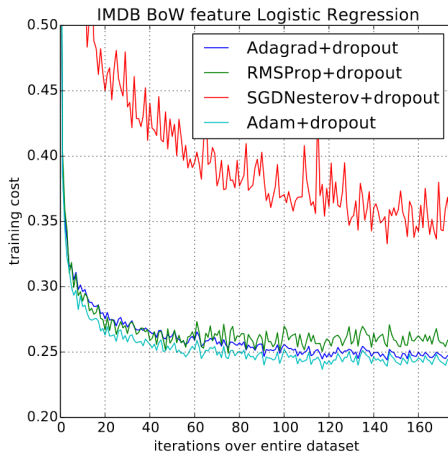
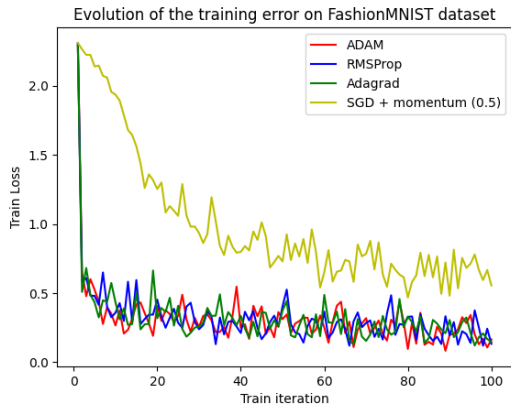
$$R(T) = \sum_{t=1}^T f_t(\theta_t) - \min_{\theta \in \mathcal{X}} \sum_{t=1}^T f_t(\theta)$$

## Théorème

*Sous certaines hypothèses de majoration du gradient de  $f_t$ , et de l'écart entre les valeurs de  $\theta_n$  on a :*

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

# Simulations numériques



# 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