

# ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

## Résumé d'article

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

5 octobre 2024

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Algorithme Adam</b>	<b>2</b>
2.1	Objectif	2
2.2	Description de l'algorithme Adam	2
<b>3</b>	<b>Points forts de la méthode</b>	<b>3</b>
3.1	Contraction du biais	3
3.2	Majoration du pas	3
3.3	Convergence	3
<b>4</b>	<b>Comparaisons expérimentales</b>	<b>4</b>
<b>5</b>	<b>Annexes</b>	<b>4</b>

## 1 Introduction

L'article « ADAM : A method for Stochastic Optimization [1] » a été publié en 2014 (et corrigé jusqu'en 2017) par Diederik P. Kingma (Université d'Amsterdam, OpenAI) et Jimmy Lei Ba (Université de Toronto) dans le cadre de l'International Conference on Learning Representations (ICLR) de 2015.

Cet article présente l'algorithme Adam, algorithme d'optimisation stochastique, basé 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 semble offrir 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.

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

## 2 Algorithme Adam

### 2.1 Objectif

On considère une fonction-objectif stochastique  $f(\theta)$  de paramètres  $\theta$ , qu'on suppose différentiable. L'algorithme Adam est une méthode d'ordre 1 (c'est-à-dire qui repose sur des évaluations de la fonctionnelle  $f$  et du gradient  $\nabla_\theta f$ ), qui a pour objectif d'optimiser les paramètres  $\theta$  afin de minimiser l'espérance  $\mathbb{E}[f(\theta)]$ .

L'aspect stochastique peut venir d'une fonction-objectif intrinsèquement bruitée ou bien d'un échantillonnage réalisé à chaque pas de l'algorithme. Typiquement,  $f$  peut être une fonction perte de la forme

$$f(\theta) = \sum_{i=1}^n \ell(x_i|\theta), \text{ où le calcul de gradient } \nabla_\theta f(\theta) = \sum_{i=1}^n \nabla_\theta \ell(x_i|\theta) \text{ est trop coûteux en nombre d'évaluations}$$

de gradients. Il est alors remplacé à chaque étape  $t$  de l'algorithme par le calcul de  $\nabla_\theta f_t(\theta) = \sum_{i \in I_t} \nabla_\theta \ell(x_i|\theta)$

pour  $I_t$  un sous-échantillon et  $f_t$  définie par  $f_t(\theta) = \sum_{i \in I_t} \ell(x_i|\theta)$ .

### 2.2 Description de l'algorithme Adam

Outre la fonction  $f(\theta)$ , l'algorithme nécessite la donnée d'un pas  $\alpha$ , de taux  $\beta_1, \beta_2 \in [0, 1[$ , d'une constante de stabilisation numérique  $\varepsilon > 0$  et de paramètres initiaux  $\theta_0$  (valeurs par défaut :  $\alpha = 0.001$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\varepsilon = 10^{-8}$ ). Il exécute alors le schéma suivant :

---

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

---

Pour comprendre cet algorithme et identifier les apports de la méthode Adam, observons les différentes étapes et comparons-les avec celles d'autres algorithmes classiques de descente de gradient stochastique (présentés dans l'annexe).

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

Fournit une estimation de  $\mathbb{E}[g_t]$  par moyenne mobile à décroissance exponentielle :  $m_t = (1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \cdot g_i$ .

Permet de garder mémoire des directions de descente précédentes afin d'atténuer les variations liées au bruit de la fonction.

On trouve une idée similaire dans l'algorithme SGD avec moment.

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

Fournit une estimation de  $\mathbb{E}[g_t^2]$ , là aussi par moyenne mobile à décroissance exponentielle.

Permettra, lors de la mise à jour des paramètres, une mise à l'échelle du gradient, c'est-à-dire qu'on ne va pas utiliser le même pas pour tous les paramètres. On ralentit le pas en cas de forte variation liée à un paramètre, et on l'accélère en cas de faible variation. On trouve une idée similaire dans les algorithmes AdaGrad (mais sans moyenne mobile) et RMSProp.

- $\hat{m}_t \leftarrow m_t/(1 - \beta_1^t)$  et  $\hat{v}_t \leftarrow v_t/(1 - \beta_2^t)$   
 Permet de réduire le biais vers 0 de  $m_t$  et  $v_t$  venant de l'initialisation de  $m_0$  et  $v_0$  à 0.  
 C'est l'innovation fournie par l'algorithme Adam.

On peut donc observer que l'algorithme Adam combine les idées fournies par les précédents algorithmes de descente de gradient, tels que la personnalisation des pas à chaque paramètre et la mémorisation du passé par le biais des moyennes mobiles, tout en apportant un élément supplémentaire : la correction des biais des moments d'ordre 1 et 2.

### 3 Points forts de la méthode

#### 3.1 Contraction du biais

La grande différence de l'algorithme Adam par rapport aux autres algorithmes de descente de gradients est fournie par les étapes déterminant  $\hat{m}_t$  et  $\hat{v}_t$ , qui ont pour effet de contracter le biais des estimateurs des moments  $m_t$  de  $\mathbb{E}[g_t]$  et  $v_t$  de  $\mathbb{E}[g_t^2]$ . En effet, nous avons :

$$\mathbb{E}[m_t] = (1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} \mathbb{E}[g_i] = \underbrace{\mathbb{E}[g_t](1 - \beta_1^t)}_{<0} + \underbrace{(1 - \beta_1) \sum_{i=1}^t \beta_1^{t-i} (\mathbb{E}[g_i] - \mathbb{E}[g_t])}_{\zeta}$$

avec  $\zeta$  qui pourra être rendu petit par un bon choix de  $\beta_1$ . La division par  $(1 - \beta_1^t)$  permet donc de réduire le biais de  $m_t$ . Il en est de même pour  $v_t$ .

En particulier, en cas de gradients parcimonieux (sparse gradient) qui nécessitent une bonne mémoire du passé, donc des taux  $\beta_1$  et  $\beta_2$  grands, le biais obtenu peut être conséquent. Le calcul de  $\hat{m}_t$  et  $\hat{v}_t$  peut alors s'avérer pertinent.

#### 3.2 Majoration du pas

Dans l'algorithme Adam, le pas à chaque étape est donné par :  $\Delta_t = \alpha \cdot \hat{m}_t / \sqrt{\hat{v}_t + \varepsilon}$ , où  $\varepsilon > 0$  évite les divisions par des nombres trop petits. En supposant  $\varepsilon = 0$ , l'article donne la borne suivante du pas :

$$|\Delta_t| \leq \begin{cases} \alpha \cdot (1 - \beta_1) / \sqrt{1 - \beta_2} & \text{si } (1 - \beta_1) > \sqrt{1 - \beta_2} \\ \alpha & \text{sinon} \end{cases}$$

La première majoration apparaît en fait dans le cadre de gradients parcimonieux. Dans les autres cas, on a une borne de l'ordre de  $\alpha$  :  $|\Delta_t| \lesssim \alpha$ .

Le choix de  $\alpha$  permet donc d'avoir un contrôle sur la taille des pas effectués, et donc de définir une « zone de confiance » autour des paramètres  $\theta_{t-1}$  dans laquelle on peut se déplacer à partir du calcul du gradient  $g_t$ .

#### 3.3 Convergence

Nous considérons ici le regret de la méthode défini par :

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

pour  $f_t$  des fonctions de pertes convexes arbitraires. Le regret fournit une sorte de mesure de performance de la trajectoire suivie par l'algorithme dans la recherche de l'optimum. Cette quantité croît avec le temps  $T$ , mais l'article fournit, sous certaines hypothèses de majoration des gradients et de l'écart entre les valeurs de  $\theta_t$ , la relation suivante, qui garantit une croissance de  $R(T)$  contrôlée :

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

On pourra se référer au résultat récent fourni par [2] pour un résultat de convergence presque sûr et en norme  $L^p$  de  $\theta_t$ .

## 4 Comparaisons expérimentales

### Sources

- [1] Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization. <https://arxiv.org/abs/1412.6980>, 2017.
- [2] Steffen Dereich and Arnulf Jentzen. Convergence rates for the adam optimizer. <https://arxiv.org/abs/2407.21078>, 2024.
- [3] Josh Starmer. Optimization for deep learning (momentum, rmsprop, adagrad, adam). <https://www.youtube.com/watch?v=NE88eqLngkg>, 2023.

## 5 Annexes

---

**Algorithme 2 : SGD**

---

**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 :**  $\theta_t$

---

---

**Algorithme 3 : SGD avec moment (1964)**

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\rho$ ,  $\theta_0$   
 $t \leftarrow 0$   
**tant que**  $\theta_t$  ne converge pas **faire**  
     $t \leftarrow t + 1$   
     $g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$   
     $m_t \leftarrow \rho \cdot m_{t-1} - \alpha \cdot g_t$   
     $\theta_t \leftarrow \theta_{t-1} + m_t$   
**fin**  
**Sorties :**  $\theta_t$

---

---

**Algorithme 4 : AdaGrad (2011)**

---

**Entrées :**  $f(\theta)$ ,  $\alpha$ ,  $\varepsilon$ ,  $\theta_0$   
 $v_0 \leftarrow 0$   
 $t \leftarrow 0$   
**tant que**  $\theta_t$  ne converge pas **faire**  
     $t \leftarrow t + 1$   
     $g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$   
     $v_t \leftarrow v_{t-1} + g_t^2$   
     $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot g_t / (\sqrt{v_t} + \varepsilon)$   
**fin**  
**Sorties :**  $\theta_t$

---

---

**Algorithme 5 : RMSProp (2012)**

---

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

$v_0 \leftarrow 0$

$t \leftarrow 0$

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

$t \leftarrow t + 1$

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

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

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

**fin**

**Sorties :**  $\theta_t$

---