Adam : Adaptative moment estimation

Apprentissage Statistique

Alexandre Capel & Anne Bernard

Université de Montpellier

Introduction

Deux points importants de la méthode :

Introduction

Deux points importants de la méthode :

- Gradient stochastique : calcul du gradient avec un sous-ensemble

STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

Deux points importants de la méthode :

- Gradient stochastique : calcul du gradient avec un sous-ensemble
- Taux d'apprentissage adaptatif : variation de α au cours de l'algorithme

Soit $f(., \theta)$ une fonction différentiable en θ.

Considérons des réalisations $f(x_1, \theta), ...f(x_T, \theta)$.

Soit $f(., \theta)$ une fonction différentiable en θ.

Considérons des réalisations $f(x_1, \theta), ...f(x_T, \theta)$.

Objectif : Minimisation en θ de $\mathbb{E}[f(X, \theta)]$

Soit $f(., \theta)$ une fonction différentiable en θ.

Considérons des réalisations $f(x_1, \theta), ...f(x_T, \theta)$.

Objectif : Minimisation en θ de $\mathbb{E}[f(X, \theta)]$

Vocabulaire : f est appelée fonction objective.

Adam est une méthode qui découle de deux autres méthodes :

AdaGrad et **RMSProp**

Adam est une méthode qui découle de deux autres méthodes :

AdaGrad et **RMSProp**

Calcul simultané de deux quantités :

Adam est une méthode qui découle de deux autres méthodes :

AdaGrad et **RMSProp**

Calcul simultané de deux quantités :

- Estimateur du moment d'ordre 1
- Estimateur du moment d'ordre 2

Adam est une méthode qui découle de deux autres méthodes :

AdaGrad et **RMSProp**

Calcul simultané de deux quantités :

- Estimateur du moment d’ordre 1
- Estimateur du moment d’ordre 2

Ne dépend pas de tous les gradients !

Algorithme 1 : ADAM

Entrées : $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$

initialisation : θ_0, $m_0 = 0$, $v_0 = 0$, $t = 0$;

tant que θ_t *ne converge pas* **faire**

$\quad$ $\mathrm{t} = \mathrm{t}+1$;

$\quad$ $g_t = \nabla_\theta f_t(\theta_{t-1})$;

$\quad$ $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$;

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

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

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

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

fin

Figure: Pseudo algorithme de Adam

Avantages

Avantages des méthodes parentes :

Avantages

Avantages des méthodes parentes :

- **AdaGrad** : *sparse* gradients (gradients clairsemés)
- **RMSProp** : environnement non-stationnaire

Avantages

Avantages des méthodes parentes :

- **AdaGrad** : *sparse* gradients (gradients clairsemés)
- **RMSProp** : environnement non-stationnaire

Avantages de **Adam** :

Avantages

Avantages des méthodes parentes :

- **AdaGrad** : *sparse* gradients (gradients clairsemés)
- **RMSProp** : environnement non-stationnaire

Avantages de **Adam** :

- Facilement implémentable
- Peu de mémoire requise
- Fonctionne en grandes dimensions

Exemples Quelle fonction objective ?

Etude de deux exemples différents pour la classification de MNIST:

Etude de deux exemples différents pour la classification de MNIST:

- **Régression multinomiale** :

  fonction objective = - log-vraisemblance

Exemples Quelle fonction objective ?

Etude de deux exemples différents pour la classification de MNIST:

- **Régression multinomiale** :

  fonction objective = - log-vraisemblance

- **Réseau de neurone jouet** :

  fonction objective = entropie croisée

Etude de deux exemples différents pour la classification de MNIST:

- **Régression multinomiale** :

  fonction objective = - log-vraisemblance

- **Réseau de neurone jouet** :

  fonction objective = entropie croisée

Comparons avec les deux méthodes : **RMSProp** et **AdaGrad**.

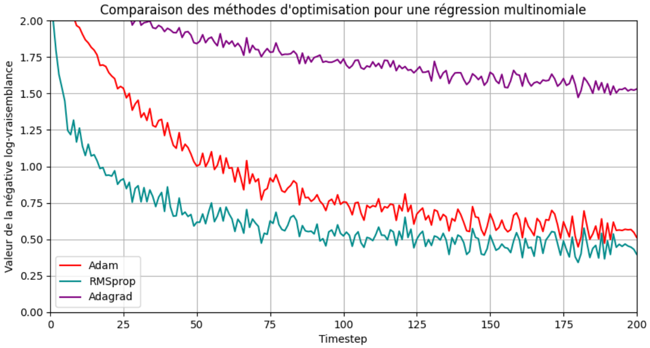

Figure: Comparaison des méthodes dans le cadre d’une régression multinomiale sur les données MNIST

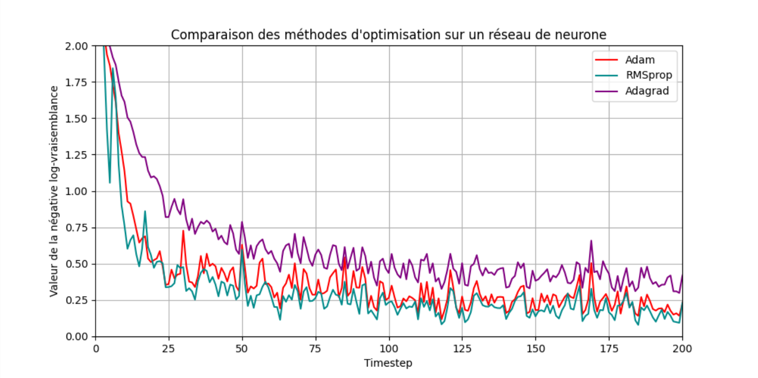

Figure: Comparaison des méthodes dans le cadre de l'optimisation d'un réseau de neurone sur les données MNIST

Résultats

Un seul résultat théorique.

Un seul résultat théorique.

Définition

Soit $f(X_1, \theta), ...f(X_T, \theta)$, un échantillon. On définit le regret par :

$$R(T) = \sum_{t=1}^{T} (f(X_t, \theta_t) - f(X_t, \tilde{\theta}_T))$$

où $\tilde{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} f(X_t, \theta)$

Résultats

Un seul résultat théorique.

Définition

Soit $f(X_1, \theta), ...f(X_T, \theta)$, un échantillon. On définit le regret par :

$$R(T) = \sum_{t=1}^{T} (f(X_t, \theta_t) - f(X_t, \tilde{\theta}_T))$$

où $\tilde{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} f(X_t, \theta)$

Sous de bonnes hypothèses, on a $\frac{R(T)}{T} = O\left(\frac{1}{\sqrt{T}}\right)$.

Limites et perspectives

Pour conclure,

- développement d'une méthode simple et qui ne requiert pas beaucoup de mémoire
- possède les deux avantages de ses prédécesseurs
- extension de la méthode en norme infini : **Adamax**
- on pourrait étudier l'effet des paramètres β_i dans les calculs

Références et fin I

Merci pour votre attention !

[1] D. P. Kingma and J. Ba.
Adam: A method for stochastic optimization.
2015.