MODIA **Février 2022**

Durée : 1h.

Les notes de cours sont autorisées.

Dans cet examen, nous allons essayer de mieux comprendre certains aspects algorithmiques des descentes de gradients stochastiques. Il y a de nombreuses questions. N'essayez pas forcément de toutes les traiter, mais celles qui le sont doivent être correctes, claires et concises (sans quoi les points ne seront pas donnés).

Question 1. On considère la suite suivante :

$$\begin{aligned} m_0 &\in \mathbb{R}^n \\ m_k &= \beta m_{k-1} + (1-\beta) g_k \end{aligned} \tag{1}$$

où (g_k) est une suite de vecteurs de $\mathbb{R}^n$ et $\beta \in [0, 1[$ est un paramètre.

1. Écrire les termes m_1, m_2, m_3.
2. Établir par récurrence la formule de m_k en fonction de k et de g_k.
3. Supposons que la suite (g_k) soit constante, i.e. $g_k = g_0 \in \mathbb{R}^n$ pour tout k. Dans ce cas que vaut le terme m_k ?

Je rappelle ci-dessous la structure de l'algorithme Adam :

Algorithm 1: La méthode ADAM.

Input:
$w^{(0)}$: un point de départ. $\alpha > 0$, le pas.
Poser $m_0 = 0$ (moment du premier ordre)
Poser $v_0 = 0$ (moment du second ordre)
for $k = 0, \ldots, K$ **do**
  $g_k = \nabla f(w_{k-1})$ (gradient courant)
  $m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k$ (mise à jour du premier moment)
  $v_k = \beta_2 v_{k-1} + (1-\beta_2) g_k^2$ (mise à jour moment d'ordre 2)
  $\hat{m}_k = m_k / (1-\beta_1^k)$ (estimation non biaisée du moment d'ordre 1)
  $\hat{v}_k = v_k / (1-\beta_2^k)$ (estimation non biaisée du moment d'ordre 2)
  $w_k = w_{k-1} - \alpha \hat{m}_k / \sqrt{(\hat{v}_k + \epsilon)}$
Renvoyer w_K.

6. Pouvez-vous expliquer ce que représente le vecteur m_k ?
7. Expliquez la division par $(1-\beta_1^k)$ dans la définition de $\hat{m}_k$.
8. Que fait l'algorithme quand $\beta_1 = \beta_2 = 0$? Pouvez-vous le relier à un autre algorithme vu en cours (et dans le polycopié).
9. Les valeurs par défaut dans Pytorch sont $\beta_1 = 0.9$ et $\beta_2 = 0.99$. Pouvez-vous commenter ce choix ?

Question 2.

Dans cette question, on va continuer à aborder le thème de la récurrence (1) sous un angle différent. Soit $a \in \mathbb{R}^n$ un vecteur. On considère le problème simple suivant :

$$\min_{w \in \mathbb{R}} f(w) = \frac{1}{2n} \sum_{i=1}^{n} (w - a_i)^2. \tag{2}$$

1. Déterminez le minimiseur $\hat{w}$ de (2).
2. Déterminez un ensemble de fonctions f_i telles que $f = \frac{1}{n} \sum_{i=1}^{n} f_i$.
3. Pour résoudre ce problème numériquement, on considère un algorithme de descente stochastique :

$$\begin{aligned} w_0 &= 0 \\ w_k &= w_{k-1} - \tau \nabla f_{i_k}(w_{k-1}). \end{aligned} \tag{3}$$

   où i_k est un indice aléatoire et $\tau > 0$ un pas de descente.
4. Quelle condition minimale sur τ est nécessaire pour établir des résultats de convergence sur (w_k) ?
5. Déterminez une relation entre les suites (3) et (1).
6. Pouvez-vous en déduire le terme général de la suite w_k ?
7. Dans le cas où on effectue une descente par coordonnées cyclique, i.e. $i_k = 1 + \mathrm{mod}(k, n)$, quelle valeur de τ vous semble pertinente ?