

---

# Optimistic Acceleration of AMSGrad for Nonconvex Optimization.

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Nonconvex Analysis

We tackle the following classical optimization problem:

$$\min_{w \in \Theta} f(w) := \mathbb{E}[f(w, \xi)] \quad (1)$$

where  $\xi$  is some random noise and only noisy versions of the objective function are accessible in this work. The objective function  $f(w)$  is (potentially) nonconvex and has Lipschitz gradients.

**Optimistic Algorithm** We present here the algorithm studied in this paper to tackle problem (1). Set the terminating iteration number,  $K \in \{0, \dots, K_{\max} - 1\}$ , as a discrete r.v. with:

$$P(K = k) = \frac{\eta_k}{\sum_{f=0}^{K_{\max}-1} \eta_f}. \quad (2)$$

where  $K_{\max} \leftarrow$  is the maximum number of iteration. The random termination number (2) is inspired by [Ghadimi and Lan, 2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization. Consider constants  $(\beta_1, \beta_2) \in [0, 1]$ , a sequence of decreasing stepsizes  $\{\eta_k\}_{k>0}$ , Algorithm 1 introduces the new optimistic AMSGrad method.

---

### Algorithm 1 OPTIMISTIC-AMSGRAD

---

```
1: Input: Parameters  $\beta_1, \beta_2, \epsilon, \eta_k$ 
2: Init.:  $w_1 = w_{-1/2} \in \mathcal{K} \subseteq \mathbb{R}^d$  and  $v_0 = \epsilon \mathbf{1} \in \mathbb{R}^d$ 
3: for  $k = 0, 1, 2, \dots, K$  do
4:   Get mini-batch stochastic gradient  $g_k$  at  $w_k$ 
5:    $\theta_k = \beta_1 \theta_{k-1} + (1 - \beta_1) g_k$ 
6:    $v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2$ 
7:    $\hat{v}_k = \max(\hat{v}_{k-1}, v_k)$ 
8:    $w_{k+\frac{1}{2}} = \Pi_{\mathcal{K}} \left[ w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}} \right]$ 
9:    $w_{k+1} = \Pi_{\mathcal{K}} \left[ w_{k+\frac{1}{2}} - \eta_k \frac{h_{k+1}}{\sqrt{\hat{v}_k}} \right]$ 
10:   where  $h_{k+1} := \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$ 
11:   and  $m_{k+1}$  is a guess of  $g_{k+1}$ 
12: end for
13: Return:  $w_{K+1}$ .
```

---

The final update at iteration  $k$  can be summarized as:

$$w_{k+1} = w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}} - \eta_k \frac{h_{k+1}}{\sqrt{\hat{v}_k}} \quad (3)$$

We make the following assumptions:

13 **H1.** The loss function  $f(w)$  is nonconvex w.r.t. the parameter  $w$ .

14 **H2.** For any  $k > 0$ , the estimated weight  $w_k$  stays within a  $\ell_\infty$ -ball. There exists a constant  $W > 0$   
15 such that:

$$\|w_k\| \leq W \quad \text{almost surely} \quad (4)$$

16 **H3.** The function  $f(w)$  is  $L$ -smooth w.r.t. the parameter  $w$ . There exist some constant  $L > 0$  such  
17 that for  $(w, \vartheta) \in \Theta^2$ :

$$f(w) - f(\vartheta) - \nabla f(\vartheta)^\top (w - \vartheta) \leq \frac{L}{2} \|w - \vartheta\|^2. \quad (5)$$

**H4.** There exists a constant  $a > 0$  such that for any  $k > 0$ :

$$\|m_{k+1}\| \leq a \|g_{k+1}\|$$

18 Classically (see [Ghadimi and Lan, 2013]) in nonconvex optimization, we make an assumption on  
19 the magnitude of the gradient:

**H5.** There exists a constant  $M > 0$  such that

$$\|\nabla f(w, \xi)\| < M \quad \text{for any } w \text{ and } \xi$$

20 We begin with some auxiliary Lemmas important for the analysis. The first one ensures bounded  
21 norms of various quantities of interests (boiling down from the classical stochastic gradient bound-  
22 edness assumption):

**Lemma 1.** Assume assumption H 5, then the quantities defined in Algorithm 1 satisfy for any  $w \in \Theta$   
and  $k > 0$ :

$$\|\nabla f(w_k)\| < M, \quad \|\theta_k\| < M^2, \quad \|\hat{v}_k\| < M.$$

**Proof** Assume assumption H 5 we have:

$$\|\nabla f(w)\| = \|\mathbb{E}[\nabla f(w, \xi)]\| \leq \mathbb{E}[\|\nabla f(w, \xi)\|] \leq M$$

23 By induction reasoning, since  $\|\theta_0\| = 0 \leq M$  and suppose that for  $\|\theta_k\| \leq M$  then we have

$$\|\theta_{k+1}\| = \|\beta_1 \theta_k + (1 - \beta_1) g_{k+1}\| \leq \beta_1 \|\theta_k\| + (1 - \beta_1) \|g_{k+1}\| \leq M \quad (6)$$

24 Using the same induction reasoning we prove that

$$\|\hat{v}_{k+1}\| = \|\beta_2 \hat{v}_k + (1 - \beta_2) g_{k+1}^2\| \leq \beta_2 \|\hat{v}_k\| + (1 - \beta_1) \|g_{k+1}^2\| \leq M^2 \quad (7)$$

25 □

26 Then, following [Yan et al., 2018] and their study of the SGD with Momentum (not AMSGrad but  
27 simple momentum) we denote for any  $k > 0$ :

$$\bar{w}_k = w_k + \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) = \frac{1}{1 - \beta_1} w_k - \frac{\beta_1}{1 - \beta_1} w_{k-1}, \quad (8)$$

28 and derive an important Lemma:

29 **Lemma 2.** Assume a strictly positive and non increasing sequence of stepsizes  $\{\eta_k\}_{k>0}$ ,  $\beta \in [0, 1]$ ,  
30 then the following holds:

$$\bar{w}_{k+1} - \bar{w}_k = \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[ \eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k, \quad (9)$$

31 where  $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$  and  $\tilde{g}_k = g_k - \beta_1 g_{k-1}$ .

32 **Proof** By definition (8) and using the Algorithm updates, we have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{1}{1 - \beta_1} (w_{k+1} - w_k) - \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + h_{k+1}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + h_k) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + \beta_1 \theta_{k-1}) - \frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (1 - \beta_1) m_{k+1} \\ &\quad + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + \beta_1 \theta_{k-2}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (1 - \beta_1) m_k \end{aligned} \quad (10)$$

33 Denote  $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$  and  $\tilde{g}_k = g_k - \beta_1 g_{k-1}$ . Notice that  $\tilde{\theta}_k = \beta_1 \tilde{\theta}_{k-1} +$   
 34  $(1 - \beta_1)(g_k + \beta_1 g_{k-1})$ .

$$\bar{w}_{k+1} - \bar{w}_k = \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[ \eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \quad (11)$$

35 □

36 **Lemma 3.** Assume H 5, a strictly positive and non increasing sequence of stepsizes  $\{\eta_k\}_{k>0}$ ,  
 37  $\beta \in [0, 1]$ , then the following holds:

$$\sum_{k=1}^K \eta_k^2 \mathbb{E} \left[ \left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] \leq \frac{\eta^2 d K (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \quad (12)$$

38 **Proof** We denote by index  $p \in [1, d]$  the dimension of each component of vectors of interest. Noting  
 39 that for any  $k > 0$  and dimension  $p$  we have  $\hat{v}_{k,p} \geq v_{k,p}$ , then:

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[ \left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &= \eta_k^2 \mathbb{E} \left[ \sum_{p=1}^d \frac{\theta_{k,p}^2}{\hat{v}_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[ \sum_{i=1}^d \frac{\theta_{k,p}^2}{v_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[ \sum_{i=1}^d \frac{(\sum_{t=1}^k (1 - \beta_1) \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k (1 - \beta_2) \beta_2^{k-t} g_{t,p}^2} \right] \end{aligned} \quad (13)$$

40 where the last inequality is due to initializations. Denote  $\gamma = \frac{\beta_1}{\beta_2}$ . Then,

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[ \left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 (1 - \beta_1)^2}{1 - \beta_2} \mathbb{E} \left[ \sum_{i=1}^d \frac{(\sum_{t=1}^k \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\stackrel{(a)}{\leq} \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[ \sum_{i=1}^d \frac{\sum_{t=1}^k \beta_1^{k-t} g_{t,p}^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\leq \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[ \sum_{i=1}^d \sum_{t=1}^k \gamma^{k-t} \right] = \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[ \sum_{t=1}^k \gamma^{k-t} \right] \end{aligned} \quad (14)$$

41 where (a) is due to  $\sum_{t=1}^k \beta_1^{k-t} \leq \frac{1}{1 - \beta_1}$ . Summing from  $k = 1$  to  $k = K$  on both sides yields:

$$\begin{aligned} \sum_{k=1}^K \eta_k^2 \mathbb{E} \left[ \left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[ \sum_{k=1}^K \sum_{t=1}^k \gamma^{k-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[ \sum_{t=1}^K \gamma^{K-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \end{aligned} \quad (15)$$

42 where the last inequality is due to  $\sum_{t=1}^k \gamma^{k-t} \leq \frac{1}{1 - \gamma}$  as a consequence of the definition of  $\gamma$ . □

43 We now formulate the main result of our paper giving a finite-time upper bound of the quantity  
 44  $\mathbb{E} [\|\nabla f(w_K)\|^2]$  where  $K$  is a random termination number distributed according to 2, see [Ghadimi  
 45 and Lan, 2013].

46 **Theorem 1.** Assume H 3-H 5,  $(\beta_1, \beta_2) \in [0, 1]$  and a sequence of decreasing stepsizes  $\{\eta_k\}_{k>0}$ ,  
 47 then the following result holds:

$$\mathbb{E} [\|\nabla f(w_K)\|^2] \leq \text{tocomplete} \quad (16)$$

48 **Proof** Using H 3 and the iterate  $\bar{w}_k$  we have:

$$\begin{aligned} f(\bar{w}_{k+1}) &\leq f(\bar{w}_k) + \nabla f(\bar{w}_k)^\top (\bar{w}_{k+1} - \bar{w}_k) + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 \\ &\leq f(\bar{w}_k) + \underbrace{\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k)}_A + \underbrace{(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k)}_B + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\| \end{aligned} \quad (17)$$

49 **Term A.** Using Lemma 2, we have that:

$$\begin{aligned} \nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) &= \nabla f(w_k)^\top \left[ \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[ \eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \right] \\ &\leq \frac{\beta_1}{1 - \beta_1} \|\nabla f(w_k)\| \left\| \eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right\| \left\| \tilde{\theta}_{k-1} \right\| - \nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k \end{aligned} \quad (18)$$

50 where the inequality is due to trivial inequality for positive diagonal matrix. Using Lemma 1 and  
51 assumption H4 we obtain:

$$\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \frac{\beta_1(1 + \beta_1 + a)}{1 - \beta_1} M^2 \left[ \left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] - \nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k \quad (19)$$

52 where we have used the fact that  $\eta_k v_k^{-1/2}$  is a diagonal matrix such that  $\eta_{k-1} v_{k-1}^{-1/2} \succcurlyeq \eta_k v_k^{-1/2} \succcurlyeq 0$   
53 (decreasing stepsize and max operator). Also note that:

$$\begin{aligned} -\nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k &= -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k - \nabla f(w_k)^\top \left[ \eta_k v_k^{-1/2} - \eta_{k-1} v_{k-1}^{-1/2} \right] \tilde{g}_k \\ &\leq -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k + (1 - \beta_1) M^2 \left[ \left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] \end{aligned} \quad (20)$$

54 using Lemma 1 on  $\|g_k\|$  and recalling that  $\tilde{g}_k = g_k - \beta_1 g_{k-1}$ . Plugging (20) into (19) yields:

$$\begin{aligned} \nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k + \frac{1}{1 - \beta_1} (\beta_1^2 + a\beta_1 + 1) M^2 \left[ \left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] \end{aligned} \quad (21)$$

55 **Term B.** By Cauchy-Schwarz (CS) inequality we have:

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| \|\bar{w}_{k+1} - \bar{w}_k\| \quad (22)$$

56 Using smoothness assumption H 3:

$$\begin{aligned} \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| &\leq L \|\bar{w}_k - w_k\| \\ &\leq L \frac{\beta_1}{1 - \beta_1} \|w_k - w_{k-1}\| \end{aligned} \quad (23)$$

57 By Lemma 2 we also have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[ \eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \eta_{k-1} v_{k-1}^{-1/2} \left[ I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \left[ I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right] (w_{k-1} - w_k) - \eta_k v_k^{-1/2} \tilde{g}_k \end{aligned} \quad (24)$$

58 where the last equality is due to  $\tilde{\theta}_{k-1} \eta_{k-1} v_{k-1}^{-1/2} = w_{k-1} - w_k$  by construction of  $\tilde{\theta}_k$ . Taking the  
59 norms on both sides, observing  $\left\| I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right\| \leq 1$  due to the decreasing stepsize  
60 and the construction of  $\hat{v}_k$  and using CS inequality yield:

$$\|\bar{w}_{k+1} - \bar{w}_k\| \leq \frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \quad (25)$$

61 Plugging (23) and (25) into (22) returns:

$$\begin{aligned} (\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq L \frac{\beta_1}{1 - \beta_1} \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \|w_k - w_{k-1}\| \\ &\quad + L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2 \end{aligned} \quad (26)$$

We recall Young's inequality with a constant  $\delta \in (0, 1)$  as follows:

$$\langle X | Y \rangle \leq \frac{1}{\delta} \|X\|^2 + \delta \|Y\|^2$$

62 Applying Young's inequality with  $\delta \rightarrow \frac{\beta_1}{1 - \beta_1}$  on the product  $\left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \|w_k - w_{k-1}\|$  yields:

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq L \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\|^2 + 2L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2 \quad (27)$$

63 The last term  $\frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2$  can be upper bounded using (25):

$$\begin{aligned} \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 &\leq \frac{L}{2} \left[ \frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \right]^2 \\ &\leq L \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\|^2 + 2L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2 \end{aligned} \quad (28)$$

64

□

65 Plugging (21), (27) and (28) into (17) and taking the expectations on both sides give:

$$\begin{aligned} &\mathbb{E} \left[ f(\bar{w}_{k+1}) + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\| \eta_k v_k^{-1/2} \right\| - \left( f(\bar{w}_k) - \frac{1}{1 - \beta_1} \tilde{M}^2 \left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| \right) \right] \\ &\leq \mathbb{E} \left[ -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k + 2L \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\|^2 + 4L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2 \right] \end{aligned} \quad (29)$$

66 where  $\tilde{M}^2 = (\beta_1^2 + a\beta_1 + 1)M^2$ . Note that  $w_{k-1} - w_k = -\eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + h_k)$  with  $h_k =$   
67  $\beta_1 \theta_{k-2} + (1 - \beta_1) m_k$  and that the expectation of  $\tilde{g}_k$  conditioned on the filtration  $\mathcal{F}_k$  reads as follows  
68

$$\begin{aligned} \mathbb{E}[\tilde{g}_k] &= \mathbb{E}[g_k - \beta_1 g_{k-1}] \\ &= \nabla f(w_k) - \beta_1 \nabla f(w_{k-1}) \end{aligned} \quad (30)$$

## 69 2 Containment of the iterates for a Deep Neural Network

70 We show in this section that the weights satisfy assumption H 2 and stay in a bounded set when  
71 the model we are fitting, using our method, is a fully connected feed forward neural network. The  
72 activation function for this section will be sigmoid function and we add a  $\ell_2$  regularization.

73 For the sake of notation, we assume  $\beta_1 = 0$ . We consider a fully connected feed forward neural  
74 network with  $L$  layers modeled by the function  $\text{MLN}(w, \xi) : \mathbb{R}^l \rightarrow \mathbb{R}$ :

$$\text{MLN}(w, \xi) = \sigma \left( w^{(L)} \sigma \left( w^{(L-1)} \dots \sigma \left( w^{(1)} \xi \right) \right) \right) \quad (31)$$

75 where  $w = [w^{(1)}, w^{(2)}, \dots, w^{(L)}]$  is the vector of parameters,  $\xi \in \mathbb{R}^l$  is the input data and  $\sigma$  is the  
76 sigmoid activation function. We assume a  $l$  dimension input data and a scalar output for simplicity.  
77 The stochastic objective function (1) reads:

$$f(w, \xi) = \mathcal{L}(\text{MLN}(w, \xi), y) + \frac{\lambda}{2} \|w\|^2 \quad (32)$$

78 where  $\mathcal{L}(\cdot, y)$  is the loss function (can be Huber loss or cross entropy),  $y$  are the true labels and  $\lambda > 0$   
 79 is the regularization parameter. Beforehand, two following mild conditions on the boundedness of  
 80 the input data and of the loss function should be verified. For any  $\xi \in \mathbb{R}^l$  and  $y \in \mathbb{R}$  there is a constant  
 81  $T > 0$  such that:

$$\|\xi\| \leq 1 \quad \text{a.s.} \quad \text{and} \quad |\mathcal{L}'(\cdot, y)| \leq T \quad (33)$$

where  $\mathcal{L}'(\cdot, y)$  denotes its derivative *w.r.t.* the paramer. For any layer index  $\ell \in [1, L]$  we denote the output of layer  $\ell$  by  $h^{(\ell)}(w, \xi)$ :

$$h^{(\ell)}(w, \xi) = \sigma \left( w^{(\ell)} \sigma \left( w^{(\ell-1)} \dots \sigma \left( w^{(1)} \xi \right) \right) \right)$$

82 Given the sigmoid assumption we have  $\|h^{(\ell)}(w, \xi)\| \leq 1$  for any  $\ell \in [1, L]$  and any  $(w, \xi) \in$   
 83  $\mathbb{R}^d \times \mathbb{R}^l$ .

84 Observe that at the last layer  $L$ :

$$\begin{aligned} \|\nabla_{w^{(L)}} \mathcal{L}(\text{MLN}(w, \xi), y)\| &= \|\mathcal{L}'(\text{MLN}(w, \xi), y) \nabla_{w^{(L)}} \text{MLN}(w, \xi)\| \\ &= \left\| \mathcal{L}'(\text{MLN}(w, \xi), y) \sigma'(w^{(L)} h^{(L-1)}(w, \xi)) h^{(L-1)}(w, \xi) \right\| \\ &\leq \frac{T}{4} \end{aligned} \quad (34)$$

85 where the last equality is due to mild assumptions (33) and to the fact that the norm of the derivative  
 86 of the sigmoid function is upperbounded by 1/4.

87 From Algorithm 1, with  $\beta_1 = 0$  we have for iteration index  $k > 0$ :

$$\begin{aligned} \|w_k - w_{k-1}\| &= \left\| -\eta_k \hat{v}_k^{-1/2} (\theta_k + h_{k+1}) \right\| \\ &= \left\| \eta_k \hat{v}_k^{-1/2} (g_k + m_{k+1}) \right\| \\ &\leq \hat{\eta} \left\| \hat{v}_k^{-1/2} g_k \right\| + \hat{\eta} a \left\| \hat{v}_k^{-1/2} g_{k+1} \right\| \end{aligned} \quad (35)$$

where  $\hat{\eta} = \max_{k>0} \eta_k$ . For any dimension  $p \in [1, d]$ , using assumption H 4, we note that

$$\sqrt{\hat{v}_{k,p}} \geq \sqrt{1 - \beta_2} g_{k,p} \quad \text{and} \quad m_{k+1} \leq a \|g_{k+1}\|$$

88 . Thus:

$$\begin{aligned} \|w_k - w_{k-1}\| &\leq \hat{\eta} \left( \left\| \hat{v}_k^{-1/2} g_k \right\| + a \left\| \hat{v}_k^{-1/2} g_{k+1} \right\| \right) \\ &\leq \hat{\eta} \frac{a + 1}{\sqrt{1 - \beta_2}} \end{aligned} \quad (36)$$

89 In short there exist a constant  $B$  such that  $\|w_k - w_{k-1}\| \leq B$ .

**Proof by induction:** As in [Défossez et al., 2020], we will prove the containment of the weights by induction. Suppose an iteration index  $K$  and a coordinate  $i$  of the last layer  $L$  such that  $w_{K,i}^{(L)} \geq \frac{T}{4\lambda} + B$ . Using (34), we have

$$\nabla_i f(w_K^{(L)}) \geq -\frac{T}{4} + \lambda \frac{T}{4\lambda} \geq 0$$

90 where  $f(\cdot)$  is defined by (32) and is the loss of our MLN. This last equation yields  $\theta_{K,i}^{(L)} \geq 0$  (given  
 91 the algorithm and  $\beta_1 = 0$ ) and using the fact that  $\|w_k - w_{k-1}\| \leq B$  we have

$$0 \leq w_{K-1,i}^{(L)} - B \leq w_{K,i}^{(L)} \leq w_{K-1,i}^{(L)} \quad (37)$$

which means that  $|w_{K,i}^{(L)}| \leq w_{K-1,i}^{(L)}$ . So if the first assumption of that induction reasoning holds, *i.e.*,  $w_{K-1,i}^{(L)} \geq \frac{T}{4\lambda} + B$ , then the next iterates  $w_{K,i}^{(L)}$  decreases, see (37) and go below  $\frac{T}{4\lambda} + B$ . This yields that for any iteration index  $k > 0$  we have

$$w_{K,i}^{(L)} \leq \frac{T}{4\lambda} + 2B$$

since  $B$  is the biggest jump an iterate can do since  $\|w_k - w_{k-1}\| \leq B$ . Likewise we can end up showing that

$$|w_{K,i}^{(L)}| \leq \frac{T}{4\lambda} + 2B$$

92 meaning that the weights of the last layer at any iteration is bounded in some matrix norm.

93 Now that we have shown this boundedness property for the last layer  $L$ , we will do the same for the  
94 previous layers and conclude the verification of assumption H 2 by induction.

95 For any layer  $\ell \in [1, L - 1]$ , we have:

$$\nabla_{w^{(\ell)}} \mathcal{L}(\text{MLN}(w, \xi), y) = \mathcal{L}'(\text{MLN}(w, \xi), y) \left( \prod_{j=1}^{\ell+1} \sigma' \left( w^{(j)} h^{(j-1)}(w, \xi) \right) \right) h^{(\ell-1)}(w, \xi) \quad (38)$$

This last quantity is bounded as long as we can prove that for any layer  $\ell$  the weights  $w^{(\ell)}$  are bounded in some matrix norm as  $\|w^{(\ell)}\|_F \leq F_\ell$  with the Frobenius norm. Suppose we have shown  $\|w^{(r)}\|_F \leq F_r$  for any layer  $r > \ell$ . Then having this gradient (38) bounded we can use the same lines of proof for the last layer  $L$  and show that the norm of the weights at the selected layer  $\ell$  satisfy

$$\|w^{(\ell)}\| \leq \frac{T \prod_{k>\ell} F_k}{4^{L-\ell+1}} + 2B$$

96 Showing that the weights of the previous layers  $\ell \in [1, L - 1]$  as well as for the last layer  $L$  of our  
97 fully connected feed forward neural network are bounded at each iteration, leads by induction, to  
98 the boundedness (at each iteration) assumption we want to check.

99 **References**

- 100 A. Défossez, L. Bottou, F. Bach, and N. Usunier. On the convergence of adam and adagrad. *arXiv*  
101 *preprint arXiv:2003.02395*, 2020.
- 102 S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic pro-  
103 gramming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- 104 Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods  
105 for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.