
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 ξ 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:

$$P(K = \ell) = \frac{\eta_\ell}{\sum_{j=0}^{K_{\max}-1} \eta_j}. \quad (2)$$

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  $k = 1, 2, \dots, K$  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 ℓ_∞ -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 (has L -Lipschitz gradients) w.r.t. the parameter w . There exist
 17 some constant $L > 0$ such that for $(w, \vartheta) \in \Theta^2$:

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

18 We assume that the optimistic guess m_k at iteration k and the true gradient g_k are correlated:

H4. There exists a constant $a \in \mathbb{R}$ such that for any $k > 0$:

$$\langle m_k | g_k \rangle \leq a \|g_k\|^2$$

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

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

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

21 We begin with some auxiliary Lemmas important for the analysis. The first one ensures bounded
 22 norms of various quantities of interests (resulting from the classical stochastic gradient boundedness
 23 assumption):

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

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

24 See Proof in Appendix A.1.

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

$$\bar{w}_k = w_k + \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) = \frac{1}{1 - \beta_1} w_k - \frac{\beta_1}{1 - \beta_1} w_{k-1}, \quad (6)$$

27 and derive an important Lemma:

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

$$\bar{w}_{k+1} - \bar{w}_k \leq \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k, \quad (7)$$

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

31 See Proof in Appendix A.2

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

$$\sum_{k=1}^{K_{\max}} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] \leq \frac{\eta^2 d K_{\max} (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \quad (8)$$

34 See Proof in Appendix A.3.

35 We now formulate the main result of our paper giving a finite-time upper bound of the quantity
 36 $\mathbb{E} [\|\nabla f(w_K)\|^2]$ where K is a random termination number distributed according to 2, see [Ghadimi
 37 and Lan, 2013].

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

$$\mathbb{E} [\|\nabla f(w_K)\|^2] \leq \tilde{C}_1 \sqrt{\frac{d}{K_{\max}}} + \tilde{C}_2 \frac{1}{K_{\max}} \quad (9)$$

where K is a random termination number distributed according (2) and the constants are defined as follows:

$$\begin{aligned} \tilde{C}_1 &= C_1 + \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \left[\frac{a(1 - \beta_1)^2}{1 - \beta_2} + 2L \frac{1}{1 - \beta_2} \right] \\ C_1 &= \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \Delta f + \frac{4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{(1 + \beta_1^2)(1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \\ \tilde{C}_2 &= \frac{M}{(1 - \beta_1)((1 - a\beta_1) + (\beta_1 + a))} \tilde{M}^2 \mathbb{E} \left[\left\| \hat{v}_0^{-1/2} \right\| \right] \end{aligned} \quad (10)$$

See proof in Appendix B.

We remark that the bound for our OPT-AMSGrad method matched the complexity bound of $\mathcal{O} \left(\sqrt{\frac{d}{K_{\max}}} + \frac{1}{K_{\max}} \right)$ of [Ghadimi and Lan, 2013] for SGD and [Zhou et al., 2018] for AMSGrad method.

2 Checking H 2 for a Deep Neural Network

We show in this section that the weights satisfy assumption H 2 and stay in a bounded set when the model we are fitting, using our method, is a fully connected feed forward neural network. The activation function for this section will be sigmoid function and we add a ℓ_2 regularization.

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

$$\text{MLN}(w, \xi) = \sigma \left(w^{(L)} \sigma \left(w^{(L-1)} \dots \sigma \left(w^{(1)} \xi \right) \right) \right) \quad (11)$$

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

$$f(w, \xi) = \mathcal{L}(\text{MLN}(w, \xi), y) + \frac{\lambda}{2} \|w\|^2 \quad (12)$$

where $\mathcal{L}(\cdot, y)$ is the loss function (can be Huber loss or cross entropy), y are the true labels and $\lambda > 0$ is the regularization parameter. For any layer index $\ell \in [1, L]$ we denote the output of layer ℓ by $h^{(\ell)}(w, \xi)$:

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

The following Lemma verifies that assumption H 2 is satisfied with a fully connected feed forward neural network (11):

Lemma 4. Given the multilayer model (11), assume the boundedness of the input data and of the loss function, i.e., for any $\xi \in \mathbb{R}^l$ and $y \in \mathbb{R}$ there is a constant $T > 0$ such that:

$$\|\xi\| \leq 1 \quad \text{a.s.} \quad \text{and} \quad |\mathcal{L}'(\cdot, y)| \leq T \quad (13)$$

where $\mathcal{L}'(\cdot, y)$ denotes its derivative w.r.t. the parameter. Then for each layer $\ell \in [1, L]$, there exist a constant $A_{(\ell)}$ such that:

$$\left\| w^{(\ell)} \right\| \leq A_{(\ell)}$$

See Proof in Appendix C

60 3 Convex Analysis:

61 **Theorem 2.** Let $\beta_1 = 0$. Suppose the learner incurs a sequence of convex loss functions $\{\ell_t(\cdot)\}$.
 62 OPTIMISTIC-AMSGRAD (Algorithm 1) has regret

$$\text{Regret}_T \leq \frac{1}{\eta_{\min}} D_\infty^2 \sum_{i=1}^d \hat{v}_T^{1/2}[i] + \frac{B_{\psi_1}(w^*, w_{1/2})}{\eta_1} + \sum_{t=1}^T \frac{\eta_t}{2} \|g_t - m_t\|_{\psi_{t-1}^*}^2, \quad (14)$$

63 where $g_t := \nabla \ell_t(w_t)$ and $\eta_{\min} := \min_t \eta_t$. The result holds for any benchmark $w^* \in \Theta$ and any
 64 step size sequence $\{\eta_t\}$.

65 **Proof** By regret decomposition, we have that

$$\begin{aligned} \text{Regret}_T &:= \sum_{t=1}^T \ell_t(w_t) - \min_{w \in \Theta} \sum_{t=1}^T \ell_t(w) \\ &\leq \sum_{t=1}^T \langle w_t - w^*, \nabla \ell_t(w_t) \rangle \\ &= \sum_{t=1}^T \langle w_t - w_{t+\frac{1}{2}}, g_t - m_t \rangle + \langle w_t - w_{t+\frac{1}{2}}, m_t \rangle + \langle w_{t+\frac{1}{2}} - w^*, g_t \rangle, \end{aligned} \quad (15)$$

66 where we denote $g_t := \nabla \ell_t(w_t)$.

67 Recall the notation $\psi_t(x)$ and the Bregman divergence $B_{\psi_t}(u, v)$ we defined in the beginning of this
 68 section. Now we are going to exploit a useful inequality (which appears in e.g., ?); for any update
 69 of the form $\hat{w} = \arg \min_{w \in \Theta} \langle w, \theta \rangle + B_\psi(w, v)$, it holds that

$$\langle \hat{w} - u, \theta \rangle \leq B_\psi(u, v) - B_\psi(u, \hat{w}) - B_\psi(\hat{w}, v), \quad (16)$$

70 for any $u \in \Theta$.

71 For $\beta_1 = 0$, we can rewrite the update on line 8 of (Algorithm 1) as

$$w_{t+\frac{1}{2}} = \arg \min_{w \in \Theta} \eta_t \langle w, g_t \rangle + B_{\psi_t}(w, w_t), \quad (17)$$

72 By using (16) for (17) with $\hat{w} = w_{t+\frac{1}{2}}$ (the output of the minimization problem), $u = w^*$ and
 73 $v = w_t$, we have

$$\langle w_{t+\frac{1}{2}} - w^*, g_t \rangle \leq \frac{1}{\eta_t} [B_{\psi_t}(w^*, w_t) - B_{\psi_t}(w^*, w_{t+\frac{1}{2}}) - B_{\psi_t}(w_{t+\frac{1}{2}}, w_t)]. \quad (18)$$

74 We can also rewrite the update on line 9 of (Algorithm 1) at time t as

$$w_{t+1} = \arg \min_{w \in \Theta} \eta_{t+1} \langle w, m_{t+1} \rangle + B_{\psi_t}(w, w_{t+\frac{1}{2}}). \quad (19)$$

75 and, by using (16) for (19), with $\hat{w} = w_t$ (the output of the minimization problem), $u = w_{t+\frac{1}{2}}$ and
 76 $v = w_{t-\frac{1}{2}}$, we have

$$\langle w_t - w_{t+\frac{1}{2}}, m_t \rangle \leq \frac{1}{\eta_t} [B_{\psi_{t-1}}(w_{t+\frac{1}{2}}, w_{t-\frac{1}{2}}) - B_{\psi_{t-1}}(w_{t+\frac{1}{2}}, w_t) - B_{\psi_{t-1}}(w_t, w_{t-\frac{1}{2}})], \quad (20)$$

77 By (15), (18), and (20), we obtain

$$\begin{aligned} \text{Regret}_T &\stackrel{(15)}{\leq} \sum_{t=1}^T \langle w_t - w_{t+\frac{1}{2}}, g_t - m_t \rangle + \langle w_t - w_{t+\frac{1}{2}}, m_t \rangle + \langle w_{t+\frac{1}{2}} - w^*, g_t \rangle \\ &\stackrel{(18), (20)}{\leq} \sum_{t=1}^T \|w_t - w_{t+\frac{1}{2}}\|_{\psi_{t-1}} \|g_t - m_t\|_{\psi_{t-1}^*} + \frac{1}{\eta_t} [B_{\psi_{t-1}}(w_{t+\frac{1}{2}}, w_{t-\frac{1}{2}}) - B_{\psi_{t-1}}(w_{t+\frac{1}{2}}, w_t) \\ &\quad - B_{\psi_{t-1}}(w_t, w_{t-\frac{1}{2}}) + B_{\psi_t}(w^*, w_t) - B_{\psi_t}(w^*, w_{t+\frac{1}{2}}) - B_{\psi_t}(w_{t+\frac{1}{2}}, w_t)], \end{aligned} \quad (21)$$

78 which is further bounded by

$$\begin{aligned}
\text{Regret}_T &\stackrel{(a)}{\leq} \sum_{t=1}^T \left\{ \frac{1}{2\eta_t} \|w_t - w_{t+\frac{1}{2}}\|_{\psi_{t-1}}^2 + \frac{\eta_t}{2} \|g_t - m_t\|_{\psi_{t-1}^*}^2 + \frac{1}{\eta_t} (B_{\psi_{t-1}}(w_{t+\frac{1}{2}}, w_{t-\frac{1}{2}}) - \frac{1}{2} \|w_{t+\frac{1}{2}} - w_t\|_{\psi_{t-1}}^2 \right. \\
&\quad \left. - B_{\psi_{t-1}}(w_t, w_{t-\frac{1}{2}}) + B_{\psi_t}(w^*, w_t) - B_{\psi_t}(w^*, w_{t+\frac{1}{2}}) - B_{\psi_t}(w_{t+\frac{1}{2}}, w_t) \right\} \\
&\leq \sum_{t=1}^T \left\{ \frac{\eta_t}{2} \|g_t - m_t\|_{\psi_{t-1}^*}^2 \right. \\
&\quad \left. + dd \right\}, \tag{22}
\end{aligned}$$

79 where (a) is because $\|w_t - w_{t+\frac{1}{2}}\|_{\psi_{t-1}} \|g_t - m_t\|_{\psi_{t-1}^*} = \inf_{\beta>0} \frac{1}{2\beta} \|w_t - w_{t+\frac{1}{2}}\|_{\psi_{t-1}}^2 + \frac{\beta}{2} \|g_t -$
80 $m_t\|_{\psi_{t-1}^*}^2$ by Young's inequality and the 1-strongly convex of $\psi_{t-1}(\cdot)$ with respect to $\|\cdot\|_{\psi_{t-1}}$.

81

□

82 **References**

- 83 A. Défossez, L. Bottou, F. Bach, and N. Usunier. On the convergence of adam and adagrad. *arXiv*
84 *preprint arXiv:2003.02395*, 2020.
- 85 S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic pro-
86 gramming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- 87 Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods
88 for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.
- 89 D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods
90 for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

91 A Proofs of Auxiliary Lemmas

92 A.1 Proof of Lemma 1

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

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

Proof Assume assumption H 5 we have:

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

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

$$\|\theta_{k+1}\| = \|\beta_1 \theta_k + (1 - \beta_1) g_{k+1}\| \leq \beta_1 \|\theta_k\| + (1 - \beta_1) \|g_{k+1}\| \leq M \quad (23)$$

94 Using the same induction reasoning we prove that

$$\|\hat{v}_{k+1}\| = \|\beta_2 \hat{v}_k + (1 - \beta_2) g_{k+1}^2\| \leq \beta_2 \|\hat{v}_k\| + (1 - \beta_1) \|g_{k+1}^2\| \leq M^2 \quad (24)$$

95

□

96 A.2 Proof of Lemma 2

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

$$\bar{w}_{k+1} - \bar{w}_k \leq \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k, \quad (25)$$

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

100 **Proof** By definition (6) and using the Algorithm updates, we have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{1}{1 - \beta_1} (w_{k+1} - w_k) - \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + h_{k+1}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + h_k) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + \beta_1 \theta_{k-1}) - \frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (1 - \beta_1) m_{k+1} \\ &\quad + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + \beta_1 \theta_{k-2}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (1 - \beta_1) m_k \end{aligned} \quad (26)$$

101 Denote $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1}$ and $\tilde{g}_k = g_k - \beta_1 m_k + \beta_1 g_{k-1} + m_{k+1}$. Notice that $\tilde{\theta}_k = \beta_1 \tilde{\theta}_{k-1} +$
102 $(1 - \beta_1)(g_k + \beta_1 g_{k-1})$.

$$\bar{w}_{k+1} - \bar{w}_k \leq \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \quad (27)$$

103

□

104 A.3 Proof of Lemma 3

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

$$\sum_{k=1}^{K_{\max}} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] \leq \frac{\eta^2 d K_{\max} (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \quad (28)$$

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

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &= \eta_k^2 \mathbb{E} \left[\sum_{p=1}^d \frac{\theta_{k,p}^2}{\hat{v}_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[\sum_{i=1}^d \frac{\theta_{k,p}^2}{v_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[\sum_{i=1}^d \frac{(\sum_{t=1}^k (1 - \beta_1) \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k (1 - \beta_2) \beta_2^{k-t} g_{t,p}^2} \right] \end{aligned} \quad (29)$$

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

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 (1 - \beta_1)^2}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \frac{(\sum_{t=1}^k \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\stackrel{(a)}{\leq} \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \frac{\sum_{t=1}^k \beta_1^{k-t} g_{t,p}^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\leq \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \sum_{t=1}^k \gamma^{k-t} \right] = \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=1}^k \gamma^{k-t} \right] \end{aligned} \quad (30)$$

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

$$\begin{aligned} \sum_{k=1}^{K_{\max}} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{k=1}^{K_{\max}} \sum_{t=1}^k \gamma^{k-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=1}^k \gamma^{k-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \end{aligned} \quad (31)$$

111 where the last inequality is due to $\sum_{t=1}^k \gamma^{k-t} \leq \frac{1}{1 - \gamma}$ by definition of γ . □

112 B Proofs of Theorem 1

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

$$\mathbb{E} [\|\nabla f(w_K)\|^2] \leq \tilde{C}_1 \sqrt{\frac{d}{K_{\max}}} + \tilde{C}_2 \frac{1}{K_{\max}} \quad (32)$$

115 where K is a random termination number distributed according (2) and the constants are defined
 116 as follows:

$$\begin{aligned} \tilde{C}_1 &= C_1 + \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \left[\frac{a(1 - \beta_1)^2}{1 - \beta_2} + 2L \frac{1}{1 - \beta_2} \right] \\ C_1 &= \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \Delta f + \frac{4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{(1 + \beta_1^2)(1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \\ \tilde{C}_2 &= \frac{M}{(1 - \beta_1)((1 - a\beta_1) + (\beta_1 + a))} \tilde{M}^2 \mathbb{E} \left[\left\| \hat{v}_0^{-1/2} \right\| \right] \end{aligned} \quad (33)$$

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

$$\begin{aligned} f(\bar{w}_{k+1}) &\leq f(\bar{w}_k) + \nabla f(\bar{w}_k)^\top (\bar{w}_{k+1} - \bar{w}_k) + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 \\ &\leq f(\bar{w}_k) + \underbrace{\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k)}_A + \underbrace{(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k)}_B + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\| \end{aligned} \quad (34)$$

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

$$\begin{aligned} \nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq \nabla f(w_k)^\top \left[\frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \right] \\ &\leq \frac{\beta_1}{1 - \beta_1} \|\nabla f(w_k)\| \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right\| \left\| \tilde{\theta}_{k-1} \right\| - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \end{aligned} \quad (35)$$

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

$$\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \frac{\beta_1(1 + \beta_1)}{1 - \beta_1} \mathbf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \quad (36)$$

121 where we have used the fact that $\eta_k \hat{v}_k^{-1/2}$ is a diagonal matrix such that $\eta_{k-1} \hat{v}_{k-1}^{-1/2} \succcurlyeq \eta_k \hat{v}_k^{-1/2} \succcurlyeq 0$
122 (decreasing stepsize and max operator). Also note that:

$$\begin{aligned} -\nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k &= -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k - \nabla f(w_k)^\top \left[\eta_k \hat{v}_k^{-1/2} - \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right] \bar{g}_k \\ &\quad - \nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\beta_1 g_{k-1} + m_{k+1}) \\ &\leq -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k + (1 - a\beta_1) \mathbf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] \\ &\quad - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} (\beta_1 g_{k-1} + m_{k+1}) \end{aligned} \quad (37)$$

123 using Lemma 1 on $\|g_k\|$ and where that $\tilde{g}_k = \bar{g}_k + \beta_1 g_{k-1} + m_{k+1} = g_k - \beta_1 m_k + \beta_1 g_{k-1} + m_{k+1}$.
124 Plugging (37) into (36) yields:

$$\begin{aligned} &\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) \\ &\leq -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k + \frac{1}{1 - \beta_1} (a\beta_1^2 - 2a\beta_1 + \beta_1) \mathbf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] \\ &\quad - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} (\beta_1 g_{k-1} + m_{k+1}) \end{aligned} \quad (38)$$

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

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| \|\bar{w}_{k+1} - \bar{w}_k\| \quad (39)$$

126 Using smoothness assumption H 3:

$$\begin{aligned} \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| &\leq L \|\bar{w}_k - w_k\| \\ &\leq L \frac{\beta_1}{1 - \beta_1} \|w_k - w_{k-1}\| \end{aligned} \quad (40)$$

127 By Lemma 2 we also have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} \left[I - (\eta_k \hat{v}_k^{-1/2})(\eta_{k-1} \hat{v}_{k-1}^{-1/2})^{-1} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \left[I - (\eta_k \hat{v}_k^{-1/2})(\eta_{k-1} \hat{v}_{k-1}^{-1/2})^{-1} \right] (w_{k-1} - w_k) - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \end{aligned} \quad (41)$$

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

$$\|\bar{w}_{k+1} - \bar{w}_k\| \leq \frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\| \quad (42)$$

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

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

131 Plugging (40) and (42) into (39) returns:

$$\begin{aligned} (\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq L \frac{\beta_1}{1 - \beta_1} \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\| \|w_k - w_{k-1}\| \\ &\quad + L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \|w_{k-1} - w_k\|^2 \end{aligned} \quad (43)$$

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

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq L \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 2L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \|w_{k-1} - w_k\|^2 \quad (44)$$

133 The last term $\frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2$ can be upper bounded using (42):

$$\begin{aligned} \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 &\leq \frac{L}{2} \left[\frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\| \right]^2 \\ &\leq L \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 2L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \|w_{k-1} - w_k\|^2 \end{aligned} \quad (45)$$

134 Plugging (38), (44) and (45) into (34) and taking the expectations on both sides give:

$$\begin{aligned} &\mathbb{E} \left[f(\bar{w}_{k+1}) + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\|\eta_k\hat{v}_k^{-1/2}\right\| - \left(f(\bar{w}_k) + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}\right\| \right) \right] \\ &\leq \mathbb{E} \left[-\nabla f(w_k)^\top \eta_{k-1}\hat{v}_{k-1}^{-1/2}\tilde{g}_k - \nabla f(w_k)^\top \eta_k\hat{v}_k^{-1/2}(\beta_1 g_{k-1} + m_{k+1}) \right] \\ &\quad + \mathbb{E} \left[2L \left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 4L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \|w_{k-1} - w_k\|^2 \right] \end{aligned} \quad (46)$$

135 where $\tilde{M}^2 = (a\beta_1^2 - 2a\beta_1 + \beta_1)M^2$. Note that the expectation of \tilde{g}_k conditioned on the filtration
 136 \mathcal{F}_k reads as follows

$$\begin{aligned} \mathbb{E} [\nabla f(w_k)^\top \tilde{g}_k] &= \mathbb{E} [\nabla f(w_k)^\top (g_k - \beta_1 m_k)] \\ &= (1 - a\beta_1) \|\nabla f(w_k)\|^2 \end{aligned} \quad (47)$$

137 Summing from $k = 1$ to $k = K$ leads to

$$\begin{aligned} &\frac{1}{M} \sum_{k=1}^{K_{\max}} ((1 - a\beta_1)\eta_{k-1} + (\beta_1 + a)\eta_k) \|\nabla f(w_k)\|^2 \leq \\ &\mathbb{E} \left[f(\bar{w}_1) + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\|\eta_0\hat{v}_0^{-1/2}\right\| - \left(f(\bar{w}_{K_{\max}+1}) + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\|\eta_{K_{\max}}\hat{v}_{K_{\max}}^{-1/2}\right\| \right) \right] \\ &\quad + 2L \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 \right] + 4L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \sum_{k=1}^{K_{\max}} \mathbb{E} [\|w_{k-1} - w_k\|^2] \\ &\leq \mathbb{E} \left[\Delta f + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\|\eta_0\hat{v}_0^{-1/2}\right\| \right] + 2L \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 \right] + 4L \left(\frac{\beta_1}{1 - \beta_1}\right)^2 \sum_{k=1}^{K_{\max}} \mathbb{E} [\|w_{k-1} - w_k\|^2] \end{aligned} \quad (48)$$

138 where $\Delta f = f(\bar{w}_1) - f(\bar{w}_{K_{\max}+1})$. We note that by definition of \hat{v}_k , and a constant learning rate
 139 η_k , we have

$$\begin{aligned}\|w_{k-1} - w_k\|^2 &= \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + h_k) \right\|^2 \\ &= \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + \beta_1 \theta_{k-2} + (1 - \beta_1) m_k) \right\|^2 \\ &\leq \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \theta_{k-1} \right\|^2 + \left\| \eta_{k-2} \hat{v}_{k-2}^{-1/2} \beta_1 \theta_{k-2} \right\|^2 + (1 - \beta_1)^2 \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} m_k \right\|^2\end{aligned}\quad (49)$$

140 Using Lemma 3 we have

$$\begin{aligned}\sum_{k=1}^{K_{\max}} \mathbb{E} \left[\|w_{k-1} - w_k\|^2 \right] \\ \leq (1 + \beta_1^2) \frac{\eta^2 d K_{\max} (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} + (1 - \beta_1)^2 \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} m_k \right\|^2 \right]\end{aligned}\quad (50)$$

141 And thus, setting the learning rate to a constant value η and injecting in (48) yields:

$$\begin{aligned}\mathbb{E} [\|\nabla f(w_K)\|^2] &= \frac{1}{\sum_{j=1}^{K_{\max}} \eta_j} \sum_{k=1}^{K_{\max}} \eta_k \|\nabla f(w_k)\|^2 \\ &\leq \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{K_{\max}} \eta_j} \mathbb{E} \left[\Delta f + \frac{1}{1 - \beta_1} \tilde{M}^2 \left\| \eta_0 \hat{v}_0^{-1/2} \right\| \right] \\ &\quad + \frac{4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{K_{\max}} \eta_j} (1 + \beta_1^2) \frac{\eta^2 d K_{\max} (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \\ &\quad + \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{K_{\max}} \eta_j} (1 - \beta_1)^2 \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} m_k \right\|^2 \right] \\ &\quad + \frac{2LM}{(1 - a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{K_{\max}} \eta_j} \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\| \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \right\|^2 \right]\end{aligned}\quad (51)$$

142 where K is a random termination number distributed according (2). Setting the stepsize to $\eta =$
 143 $\frac{1}{\sqrt{d K_{\max}}}$ yields :

$$\begin{aligned}\mathbb{E} [\|\nabla f(w_K)\|^2] \\ \leq C_1 \sqrt{\frac{d}{K_{\max}}} + C_2 \frac{1}{K_{\max}} \\ + D_1 \frac{\eta}{K_{\max}} \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\| \hat{v}_{k-1}^{-1/2} m_k \right\|^2 \right] + D_2 \frac{\eta}{K_{\max}} \sum_{k=1}^{K_{\max}} \mathbb{E} \left[\left\| \hat{v}_{k-1}^{-1/2} \tilde{g}_k \right\|^2 \right]\end{aligned}\quad (52)$$

144 where

$$\begin{aligned}C_1 &= \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \Delta f + \frac{4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 M}{(1 - a\beta_1) + (\beta_1 + a)} \frac{(1 + \beta_1^2)(1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \\ C_2 &= \frac{M}{(1 - \beta_1)((1 - a\beta_1) + (\beta_1 + a))} \tilde{M}^2 \mathbb{E} \left[\left\| \hat{v}_0^{-1/2} \right\| \right]\end{aligned}\quad (53)$$

145 **Simple case as in [Zhou et al., 2018]:** if $\beta_1 = 0$ then $\tilde{g}_k = g_k + m_{k+1}$ and $g_k = \theta_k$. Also using
 146 Lemma 3 we have that:

$$\sum_{k=1}^{K_{\max}} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} g_k \right\|_2^2 \right] \leq \frac{\eta^2 d K_{\max}}{(1 - \beta_2)} \quad (54)$$

147 which leads to the final bound:

$$\begin{aligned} & \mathbb{E} [\|\nabla f(w_K)\|^2] \\ & \leq \tilde{C}_1 \sqrt{\frac{d}{K_{\max}}} + \tilde{C}_2 \frac{1}{K_{\max}} \end{aligned} \quad (55)$$

148 where

$$\begin{aligned} \tilde{C}_1 &= C_1 + \frac{M}{(1 - a\beta_1) + (\beta_1 + a)} \left[\frac{a(1 - \beta_1)^2}{1 - \beta_2} + 2L \frac{1}{1 - \beta_2} \right] \\ \tilde{C}_2 &= C_2 = \frac{M}{(1 - \beta_1)((1 - a\beta_1) + (\beta_1 + a))} \tilde{M}^2 \mathbb{E} [\|\hat{v}_0^{-1/2}\|] \end{aligned} \quad (56)$$

149

□

150 C Proof of Lemma 4 (Boundedness of the iterates)

151 **Lemma.** *Given the multilayer model (11), assume the boundedness of the input data and of the loss*
 152 *function, i.e., for any $\xi \in \mathbb{R}^l$ and $y \in \mathbb{R}$ there is a constant $T > 0$ such that:*

$$\|\xi\| \leq 1 \quad \text{a.s.} \quad \text{and} \quad |\mathcal{L}'(\cdot, y)| \leq T \quad (57)$$

where $\mathcal{L}'(\cdot, y)$ denotes its derivative w.r.t. the parameter. Then for each layer $\ell \in [1, L]$, there exist a constant $A_{(\ell)}$ such that:

$$\|w^{(\ell)}\| \leq A_{(\ell)}$$

Proof Recall that for any layer index $\ell \in [1, L]$ we denote the output of layer ℓ by $h^{(\ell)}(w, \xi)$:

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

153 Given the sigmoid assumption we have $\|h^{(\ell)}(w, \xi)\| \leq 1$ for any $\ell \in [1, L]$ and any $(w, \xi) \in$
 154 $\mathbb{R}^d \times \mathbb{R}^l$. Observe that at the last layer L :

$$\begin{aligned} \|\nabla_{w^{(L)}} \mathcal{L}(\text{MLN}(w, \xi), y)\| &= \|\mathcal{L}'(\text{MLN}(w, \xi), y) \nabla_{w^{(L)}} \text{MLN}(w, \xi)\| \\ &= \left\| \mathcal{L}'(\text{MLN}(w, \xi), y) \sigma'(w^{(L)} h^{(L-1)}(w, \xi)) h^{(L-1)}(w, \xi) \right\| \\ &\leq \frac{T}{4} \end{aligned} \quad (58)$$

155 where the last equality is due to mild assumptions (57) and to the fact that the norm of the derivative
 156 of the sigmoid function is upperbounded by 1/4.

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

$$\begin{aligned} \|w_k - w_{k-1}\| &= \left\| -\eta_k \hat{v}_k^{-1/2} (\theta_k + h_{k+1}) \right\| \\ &= \left\| \eta_k \hat{v}_k^{-1/2} (g_k + m_{k+1}) \right\| \\ &\leq \hat{\eta} \left\| \hat{v}_k^{-1/2} g_k \right\| + \hat{\eta} a \left\| \hat{v}_k^{-1/2} g_{k+1} \right\| \end{aligned} \quad (59)$$

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

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

158 . Thus:

$$\begin{aligned} \|w_k - w_{k-1}\| &\leq \hat{\eta} \left(\left\| \hat{v}_k^{-1/2} g_k \right\| + a \left\| \hat{v}_k^{-1/2} g_{k+1} \right\| \right) \\ &\leq \hat{\eta} \frac{a + 1}{\sqrt{1 - \beta_2}} \end{aligned} \quad (60)$$

159 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 (58), we have

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

160 where $f(\cdot)$ is defined by (12) and is the loss of our MLN. This last equation yields $\theta_{K,i}^{(L)} \geq 0$ (given
161 the algorithm and $\beta_1 = 0$) and using the fact that $\|w_k - w_{k-1}\| \leq B$ we have

$$0 \leq w_{K-1,i}^{(L)} - B \leq w_{K,i}^{(L)} \leq w_{K-1,i}^{(L)} \quad (61)$$

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 (61) 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

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

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

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

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

$$\nabla_{w^{(\ell)}} \mathcal{L}(\text{MLN}(w, \xi), y) = \mathcal{L}'(\text{MLN}(w, \xi), y) \left(\prod_{j=1}^{\ell+1} \sigma' \left(w^{(j)} h^{(j-1)}(w, \xi) \right) \right) h^{(\ell-1)}(w, \xi) \quad (62)$$

This last quantity is bounded as long as we can prove that for any layer ℓ 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 (62) 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 ℓ satisfy

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

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