Optimistic Acceleration of AMSGrad for Nonconvex Optimization.

Anonymous Author(s)
Affiliation
Address
email

1 Nonconvex Analysis

We tackle the following classical optimization problem:

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

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, $K \in \{0, \dots, K_{\mathsf{max}} - 1\}$, as a discrete r.v. with:

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

where $K_{\mathsf{max}} \leftarrow$ 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

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

The final update at iteration k can be summarized as:

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

We make the following assumptions:

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

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

H2. *For any $k > 0$, the estimated weight w_k stays within a ℓ_∞−ball. There exists a constant $W > 0$ such that:*

$$\|w_k\| \leq W \quad \textit{almost surely} \tag{4}$$

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

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

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 $k > 0$:*

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

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

H5. *There exists a constant $\mathsf{M} > 0$ such that*

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

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

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

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

See Proof in Appendix A.1.

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

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

and derive an important Lemma:

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

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

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

See Proof in Appendix A.2

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

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

See Proof in Appendix A.3.

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

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

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

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

See proof in Appendix B.

We remark that the bound for our OPT-AMSGrad method matched the complexity bound of $\mathcal{O}\left(\sqrt{\frac{d}{K_{\mathsf{max}}}} + \frac{1}{K_{\mathsf{max}}}\right)$ 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 $\mathsf{MLN}(w, \xi) : \mathbb{R}^l \to \mathbb{R}$:

$$\mathsf{MLN}(w, \xi) = \sigma\left(w^{(L)}\sigma\left(w^{(L-1)} \ldots \sigma\left(w^{(1)}\xi\right)\right)\right) \tag{11}$$

where $w = [w^{(1)}, w^{(2)}, \cdots, w^{(L)}]$ 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:

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

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 $h^{(\ell)}(w, \xi)$:

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

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:*

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

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

References

A. Défossez, L. Bottou, F. Bach, and N. Usunier. On the convergence of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.

D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

A Proofs of Auxiliary Lemmas

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 $k > 0$:*

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

Proof Assume assumption H 5 we have:

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

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

$$\|\theta_{k+1}\| = \|\beta_1 \theta_k + (1-\beta_1)\, g_{k+1}\| \leq \beta_1 \|\theta_k\| + (1-\beta_1)\|g_{k+1}\| \leq \mathsf{M} \tag{14}$$

Using the same induction reasoning we prove that

$$\|\hat{v}_{k+1}\| = \left\|\beta_2 \hat{v}_k + (1-\beta_2)\, g_{k+1}^2\right\| \leq \beta_2 \|\hat{v}_k\| + (1-\beta_1)\left\|g_{k+1}^2\right\| \leq \mathsf{M}^2 \tag{15}$$

□

A.2 Proof of Lemma 2

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

$$\overline{w}_{k+1} - \overline{w}_k \leq \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{k-1}\left[\eta_{k-1}\hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2}\right] - \eta_k \hat{v}_k^{-1/2}\tilde{g}_k \ , \tag{16}$$

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

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

$$\begin{aligned}
\overline{w}_{k+1} - \overline{w}_k &= \frac{1}{1-\beta_1}(w_{k+1} - w_k) - \frac{\beta_1}{1-\beta_1}(w_k - w_{k-1}) \\
&= -\frac{1}{1-\beta_1}\eta_k \hat{v}_k^{-1/2}(\theta_k + h_{k+1}) + \frac{\beta_1}{1-\beta_1}\eta_{k-1}\hat{v}_{k-1}^{-1/2}(\theta_{k-1} + h_k) \\
&= -\frac{1}{1-\beta_1}\eta_k \hat{v}_k^{-1/2}(\theta_k + \beta_1 \theta_{k-1}) - \frac{1}{1-\beta_1}\eta_k \hat{v}_k^{-1/2}(1-\beta_1) m_{k+1} \\
&\quad + \frac{\beta_1}{1-\beta_1}\eta_{k-1}\hat{v}_{k-1}^{-1/2}(\theta_{k-1} + \beta_1 \theta_{k-2}) + \frac{\beta_1}{1-\beta_1}\eta_{k-1}\hat{v}_{k-1}^{-1/2}(1-\beta_1) m_k
\end{aligned} \tag{17}$$

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 $\tilde{\theta}_k = \beta_1 \tilde{\theta}_{k-1} + (1-\beta_1)(g_k + \beta_1 g_{k-1})$.

$$\overline{w}_{k+1} - \overline{w}_k \leq \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{k-1}\left[\eta_{k-1}\hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2}\right] - \eta_k \hat{v}_k^{-1/2}\tilde{g}_k \tag{18}$$

□

A.3 Proof of Lemma 3

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

$$\sum_{k=1}^{K_{\mathsf{max}}} \eta_k^2 \mathbb{E}\left[\left\|\hat{v}_k^{-1/2}\theta_k\right\|_2^2\right] \leq \frac{\eta^2 d K_{\mathsf{max}}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \tag{19}$$

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

$$
\begin{aligned}
\eta_k^2 \mathbb{E}\left[\left\|\hat{v}_k^{-1/2}\theta_k\right\|_2^2\right] &= \eta_k^2 \mathbb{E}\left[\sum_{p=1}^{d} \frac{\theta_{k,p}^2}{\hat{v}_{k,p}}\right] \\
&\leq \eta_k^2 \mathbb{E}\left[\sum_{i=1}^{d} \frac{\theta_{k,p}^2}{v_{k,p}}\right] \\
&\leq \eta_k^2 \mathbb{E}\left[\sum_{i=1}^{d} \frac{(\sum_{t=1}^{k}(1-\beta_1)\beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^{k}(1-\beta_2)\beta_2^{k-t} g_{t,p}^2}\right]
\end{aligned}
\tag{20}
$$

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

$$
\begin{aligned}
\eta_k^2 \mathbb{E}\left[\left\|\hat{v}_k^{-1/2}\theta_k\right\|_2^2\right] &\leq \frac{\eta_k^2(1-\beta_1)^2}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d} \frac{(\sum_{t=1}^{k}\beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^{k}\beta_2^{k-t} g_{t,p}^2}\right] \\
&\overset{(a)}{\leq} \frac{\eta_k^2(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d} \frac{\sum_{t=1}^{k}\beta_1^{k-t} g_{t,p}^2}{\sum_{t=1}^{k}\beta_2^{k-t} g_{t,p}^2}\right] \\
&\leq \frac{\eta_k^2(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d}\sum_{t=1}^{k} \gamma^{k-t}\right] = \frac{\eta_k^2 d(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{t=1}^{k} \gamma^{k-t}\right]
\end{aligned}
\tag{21}
$$

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

$$
\begin{aligned}
\sum_{k=1}^{K_{\mathsf{max}}} \eta_k^2 \mathbb{E}\left[\left\|\hat{v}_k^{-1/2}\theta_k\right\|_2^2\right] &\leq \frac{\eta_k^2 d(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{k=1}^{K_{\mathsf{max}}}\sum_{t=1}^{k} \gamma^{k-t}\right] \\
&\leq \frac{\eta^2 dK(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{t=t}^{k} \gamma^{k-t}\right] \\
&\leq \frac{\eta^2 dK(1-\beta_1)}{(1-\beta_2)(1-\gamma)}
\end{aligned}
\tag{22}
$$

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

B Proofs of Theorem 1

Theorem. *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:*

$$
\mathbb{E}\left[\|\nabla f(w_K)\|^2\right] \leq \tilde{C}_1 \sqrt{\frac{d}{K_{\mathsf{max}}}} + \tilde{C}_2 \frac{1}{K_{\mathsf{max}}}
\tag{23}
$$

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

$$
\begin{aligned}
\tilde{C}_1 &= C_1 + \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\left[\frac{a(1-\beta_1)^2}{1-\beta_2} + 2L\frac{1}{1-\beta_2}\right] \\
C_1 &= \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\Delta f + \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \mathsf{M}}{(1-a\beta_1)+(\beta_1+a)} \frac{(1+\beta_1^2)(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\
\tilde{C}_2 &= \frac{\mathsf{M}}{(1-\beta_1)\left((1-a\beta_1)+(\beta_1+a)\right)} \tilde{\mathsf{M}}^2 \mathbb{E}\left[\left\|\hat{v}_0^{-1/2}\right\|\right]
\end{aligned}
\tag{24}
$$

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

$$
\begin{aligned}
f(\overline{w}_{k+1}) &\leq f(\overline{w}_k) + \nabla f(\overline{w}_k)^\top (\overline{w}_{k+1} - \overline{w}_k) + \frac{L}{2} \|\overline{w}_{k+1} - \overline{w}_k\|^2 \\
&\leq f(\overline{w}_k) + \underbrace{\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k)}_{A} + \underbrace{(\nabla f(\overline{w}_k) - \nabla f(w_k))^\top (\overline{w}_{k+1} - \overline{w}_k)}_{B} + \frac{L}{2} \|\overline{w}_{k+1} - \overline{w}_k\|
\end{aligned}
\tag{25}
$$

Term A. Using Lemma 2, we have that:

$$
\begin{aligned}
\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k) &\leq \nabla f(w_k)^\top \left[\frac{\beta_1}{1-\beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \right] \\
&\leq \frac{\beta_1}{1-\beta_1} \|\nabla f(w_k)\| \left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right\| \left\| \tilde{\theta}_{k-1} \right\| - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k
\end{aligned}
\tag{26}
$$

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

$$
\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k) \leq \frac{\beta_1(1+\beta_1)}{1-\beta_1} \mathsf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k
\tag{27}
$$

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

$$
\begin{aligned}
-\nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} \tilde{g}_k &= -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k - \nabla f(w_k)^\top \left[\eta_k \hat{v}_k^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] \bar{g}_k \\
&\quad - \nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\beta_1 g_{k-1} + m_{k+1}) \\
&\leq -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k + (1 - a\beta_1) \mathsf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] \\
&\quad - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} (\beta_1 g_{k-1} + m_{k+1})
\end{aligned}
\tag{28}
$$

using Lemma 1 on $\|g_k\|$ and where that $\tilde{g}_k = \bar{g}_k + \beta_1 g_{k-1} + m_{k+1} = g_k - \beta_1 m_k + \beta_1 g_{k-1} + m_{k+1}$. Plugging (28) into (27) yields:

$$
\begin{aligned}
&\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k) \\
&\leq -\nabla f(w_k)^\top \eta_{k-1} \hat{v}_{k-1}^{-1/2} \bar{g}_k + \frac{1}{1-\beta_1} (a\beta_1^2 - 2a\beta_1 + \beta 1) \mathsf{M}^2 \left[\left\| \eta_{k-1} \hat{v}_{k-1}^{-1/2} \right\| - \left\| \eta_k \hat{v}_k^{-1/2} \right\| \right] \\
&- \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2} (\beta_1 g_{k-1} + m_{k+1})
\end{aligned}
\tag{29}
$$

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

$$
(\nabla f(\overline{w}_k) - \nabla f(w_k))^\top (\overline{w}_{k+1} - \overline{w}_k) \leq \|\nabla f(\overline{w}_k) - \nabla f(w_k)\| \|\overline{w}_{k+1} - \overline{w}_k\|
\tag{30}
$$

Using smoothness assumption H 3:

$$
\begin{aligned}
\|\nabla f(\overline{w}_k) - \nabla f(w_k)\| &\leq L \|\overline{w}_k - w_k\| \\
&\leq L \frac{\beta_1}{1-\beta_1} \|w_k - w_{k-1}\|
\end{aligned}
\tag{31}
$$

By Lemma 2 we also have:

$$
\begin{aligned}
\overline{w}_{k+1} - \overline{w}_k &= \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \\
&= \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{k-1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} \left[I - (\eta_k \hat{v}_k^{-1/2})(\eta_{k-1} \hat{v}_{k-1}^{-1/2})^{-1} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \\
&= \frac{\beta_1}{1-\beta_1} \left[I - (\eta_k \hat{v}_k^{-1/2})(\eta_{k-1} \hat{v}_{k-1}^{-1/2})^{-1} \right] (w_{k-1} - w_k) - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k
\end{aligned}
\tag{32}
$$

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 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 and the construction of $\hat{v}_k$ and using CS inequality yield:

$$\|\overline{w}_{k+1} - \overline{w}_k\| \leq \frac{\beta_1}{1-\beta_1}\|w_{k-1} - w_k\| + \left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\| \tag{33}$$

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

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

Plugging (31) and (33) into (30) returns:

$$\begin{aligned}(\nabla f(\overline{w}_k) - \nabla f(w_k))^\top (\overline{w}_{k+1} - \overline{w}_k) \leq & L\frac{\beta_1}{1-\beta_1}\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\| \|w_k - w_{k-1}\| \\ & + L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \|w_{k-1} - w_k\|^2\end{aligned} \tag{34}$$

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

$$(\nabla f(\overline{w}_k) - \nabla f(w_k))^\top (\overline{w}_{k+1} - \overline{w}_k) \leq L\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \|w_{k-1} - w_k\|^2 \tag{35}$$

The last term $\frac{L}{2}\|\overline{w}_{k+1} - \overline{w}_k\|$ can be upper bounded using (33):

$$\begin{aligned}\frac{L}{2}\|\overline{w}_{k+1} - \overline{w}_k\|^2 & \leq \frac{L}{2}\left[\frac{\beta_1}{1-\beta_1}\|w_{k-1} - w_k\| + \left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|\right] \\ & \leq L\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \|w_{k-1} - w_k\|^2\end{aligned} \tag{36}$$

Plugging (29), (35) and (36) into (25) and taking the expectations on both sides give:

$$\begin{aligned}& \mathbb{E}\left[f(\overline{w}_{k+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_k \hat{v}_k^{-1/2}\right\| - \left(f(\overline{w}_k) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}\right\|\right)\right] \\ & \leq \mathbb{E}\left[-\nabla f(w_k)^\top \eta_{k-1}\hat{v}_{k-1}^{-1/2}\bar{g}_k - \nabla f(w_k)^\top \eta_k \hat{v}_k^{-1/2}(\beta_1 g_{k-1} + m_{k+1})\right] \\ & + \mathbb{E}\left[2L\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|^2 + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \|w_{k-1} - w_k\|^2\right]\end{aligned} \tag{37}$$

where $\tilde{\mathsf{M}}^2 = (a\beta_1^2 - 2a\beta_1 + \beta 1)\mathsf{M}^2$. Note that the expectation of $\tilde{g}_k$ conditioned on the filtration $\mathcal{F}_k$ reads as follows

$$\begin{aligned}\mathbb{E}\left[\nabla f(w_k)^\top \bar{g}_k\right] & = \mathbb{E}\left[\nabla f(w_k)^\top (g_k - \beta_1 m_k)\right] \\ & = (1 - a\beta_1)\|\nabla f(w_k)\|^2\end{aligned} \tag{38}$$

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

$$\begin{aligned}& \frac{1}{\mathsf{M}}\sum_{k=1}^{K_{\mathsf{max}}}\left((1 - a\beta_1)\eta_{k-1} + (\beta_1 + a)\eta_k\right)\|\nabla f(w_k)\|^2 \leq \\ & \mathbb{E}\left[f(\overline{w}_1) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_0 \hat{v}_0^{-1/2}\right\| - \left(f(\overline{w}_{K_{\mathsf{max}}+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_{K_{\mathsf{max}}}\hat{v}_{K_{\mathsf{max}}}^{-1/2}\right\|\right)\right] \\ & + 2L\sum_{k=1}^{K_{\mathsf{max}}}\mathbb{E}\left[\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|^2\right] + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{k=1}^{K_{\mathsf{max}}}\mathbb{E}\left[\|w_{k-1} - w_k\|^2\right] \\ & \leq \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_0 \hat{v}_0^{-1/2}\right\|\right] + 2L\sum_{k=1}^{K_{\mathsf{max}}}\mathbb{E}\left[\left\|\eta_k \hat{v}_k^{-1/2}\tilde{g}_k\right\|^2\right] + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{k=1}^{K_{\mathsf{max}}}\mathbb{E}\left[\|w_{k-1} - w_k\|^2\right]\end{aligned} \tag{39}$$

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

$$\begin{aligned}\|w_{k-1} - w_k\|^2 &= \left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}(\theta_{k-1} + h_k)\right\|^2 \\ &= \left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}(\theta_{k-1} + \beta_1\theta_{k-2} + (1-\beta_1)m_k)\right\|^2 \\ &\leq \left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}\theta_{k-1}\right\|^2 + \left\|\eta_{k-2}\hat{v}_{k-2}^{-1/2}\beta_1\theta_{k-2}\right\|^2 + (1-\beta_1)^2\left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}m_k\right\|^2\end{aligned} \tag{40}$$

Using Lemma 3 we have

$$\begin{aligned}&\sum_{k=1}^{K_{\max}} \mathbb{E}\left[\|w_{k-1} - w_k\|^2\right] \\ &\leq (1+\beta_1^2)\frac{\eta^2 d K_{\max}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} + (1-\beta_1)^2\sum_{k=1}^{K_{\max}}\mathbb{E}\left[\left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}m_k\right\|^2\right]\end{aligned} \tag{41}$$

And thus, setting the learning rate to a constant value η and injecting in (39) yields:

$$\begin{aligned}&\mathbb{E}\left[\|\nabla f(w_K)\|^2\right] = \frac{1}{\sum_{j=1}^{K_{\max}}\eta_j}\sum_{k=1}^{K_{\max}}\eta_k\|\nabla f(w_k)\|^2 \\ &\leq \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\frac{1}{\sum_{j=1}^{K_{\max}}\eta_j}\mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\left\|\eta_0\hat{v}_0^{-1/2}\right\|\right] \\ &+ \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\frac{1}{\sum_{j=1}^{K_{\max}}\eta_j}(1+\beta_1^2)\frac{\eta^2 d K_{\max}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\ &+ \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\frac{1}{\sum_{j=1}^{K_{\max}}\eta_j}(1-\beta_1)^2\sum_{k=1}^{K_{\max}}\mathbb{E}\left[\left\|\eta_{k-1}\hat{v}_{k-1}^{-1/2}m_k\right\|^2\right] \\ &+ \frac{2L\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\frac{1}{\sum_{j=1}^{K_{\max}}\eta_j}\sum_{k=1}^{K_{\max}}\mathbb{E}\left[\left\|\eta_k\hat{v}_k^{-1/2}\tilde{g}_k\right\|^2\right]\end{aligned} \tag{42}$$

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

$$\begin{aligned}&\mathbb{E}\left[\|\nabla f(w_K)\|^2\right] \\ &\leq C_1\sqrt{\frac{d}{K_{\max}}} + C_2\frac{1}{K_{\max}} \\ &+ D_1\frac{\eta}{K_{\max}}\sum_{k=1}^{K_{\max}}\mathbb{E}\left[\left\|\hat{v}_{k-1}^{-1/2}m_k\right\|^2\right] + D_2\frac{\eta}{K_{\max}}\sum_{k=1}^{K_{\max}}\mathbb{E}\left[\left\|\hat{v}_{k-1}^{-1/2}\tilde{g}_k\right\|^2\right]\end{aligned} \tag{43}$$

where

$$\begin{aligned}C_1 &= \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\Delta f + \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\frac{(1+\beta_1^2)(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\ C_2 &= \frac{\mathsf{M}}{(1-\beta_1)\left((1-a\beta_1)+(\beta_1+a)\right)}\tilde{\mathsf{M}}^2\mathbb{E}\left[\left\|\hat{v}_0^{-1/2}\right\|\right]\end{aligned} \tag{44}$$

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 Lemma 3 we have that:

$$\sum_{k=1}^{K_{\max}}\eta_k^2\mathbb{E}\left[\left\|\hat{v}_k^{-1/2}g_k\right\|_2^2\right] \leq \frac{\eta^2 d K_{\max}}{(1-\beta_2)} \tag{45}$$

which leads to the final bound:

$$\begin{aligned} &\mathbb{E}\left[\|\nabla f(w_K)\|^2\right] \\ &\leq \tilde{C}_1 \sqrt{\frac{d}{K_{\mathsf{max}}}} + \tilde{C}_2 \frac{1}{K_{\mathsf{max}}} \end{aligned} \tag{46}$$

where

$$\begin{aligned} \tilde{C}_1 &= C_1 + \frac{\mathsf{M}}{(1-a\beta_1)+(\beta_1+a)}\left[\frac{a(1-\beta_1)^2}{1-\beta_2} + 2L\frac{1}{1-\beta_2}\right] \\ \tilde{C}_2 &= C_2 = \frac{\mathsf{M}}{(1-\beta_1)\left((1-a\beta_1)+(\beta_1+a)\right)}\tilde{\mathsf{M}}^2 \mathbb{E}\left[\left\|\hat{v}_0^{-1/2}\right\|\right] \end{aligned} \tag{47}$$

$\square$

C Proof of Lemma 4 (Boundedness of the iterates)

Lemma. *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:*

$$\|\xi\| \leq 1 \quad a.s. \quad and |\mathcal{L}'(\cdot, y)| \leq T \tag{48}$$

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:*

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

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

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

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

$$\begin{aligned} \|\nabla_{w^{(L)}}\mathcal{L}(\mathsf{MLN}(w,\xi),y)\| &= \|\mathcal{L}'(\mathsf{MLN}(w,\xi),y)\nabla_{w^{(L)}}\mathsf{MLN}(w,\xi)\| \\ &= \left\|\mathcal{L}'(\mathsf{MLN}(w,\xi),y)\sigma'(w^{(L)}h^{(L-1)}(w,\xi))h^{(L-1)}(w,\xi)\right\| \\ &\leq \frac{T}{4} \end{aligned} \tag{49}$$

where the last equality is due to mild assumptions (48) and to the fact that the norm of the derivative of the sigmoid function is upperbounded by $1/4$.

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

$$\begin{aligned} \|w_k - w_{k-1}\| &= \left\|-\eta_k \hat{v}_k^{-1/2}(\theta_k + h_{k+1})\right\| \\ &= \left\|\eta_k \hat{v}_k^{-1/2}(g_k + m_{k+1})\right\| \\ &\leq \hat{\eta}\left\|\hat{v}_k^{-1/2} g_k\right\| + \hat{\eta} a \left\|\hat{v}_k^{-1/2} g_{k+1}\right\| \end{aligned} \tag{50}$$

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

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

. Thus:

$$\begin{aligned} \|w_k - w_{k-1}\| &\leq \hat{\eta}\left(\left\|\hat{v}_k^{-1/2} g_k\right\| + a\left\|\hat{v}_k^{-1/2} g_{k+1}\right\|\right) \\ &\leq \hat{\eta}\frac{a+1}{\sqrt{1-\beta_2}} \end{aligned} \tag{51}$$

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 (49), we have

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

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 the algorithm and $\beta_1 = 0$) and using the fact that $\|w_k - w_{k-1}\| \leq B$ we have

$$0 \leq w_{K-1,i}^{(L)} - B \leq w_{K,i}^{(L)} \leq w_{K-1,i}^{(L)} \tag{52}$$

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 (52) 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

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

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

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

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

$$\nabla_{w^{(\ell)}} \mathcal{L}(\mathsf{MLN}(w, \xi), y) = \mathcal{L}'(\mathsf{MLN}(w, \xi), y) \left(\prod_{j=1}^{\ell+1} \sigma' \left(w^{(j)} h^{(j-1)}(w, \xi) \right) \right) h^{(\ell-1)}(w, \xi) \tag{53}$$

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 $\left\| w^{(\ell)} \right\|_F \leq F_\ell$ with the Frobenius norm. Suppose we have shown $\left\| w^{(r)} \right\|_F \leq F_r$ for any layer $r > \ell$. Then having this gradient (53) 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

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

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