A Proof of Theorem 1

Theorem. *Suppose the learner incurs a sequence of convex loss functions* $\{\ell_t(\cdot)\}$*. Then,* OPT-AMSGRAD *(Algorithm 2) has regret*

$$\mathcal{R}_T \leq \frac{B_{\psi_1}(w^*, \tilde{w}_1)}{\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|g_t - \tilde{m}_t\|^2_{\psi^*_{t-1}} + \frac{D^2_\infty}{\eta_{\min}} \sum_{i=1}^{d} \hat{v}_T^{1/2}[i] + D^2_\infty \beta_1^2 \sum_{t=1}^{T} \|g_t - \theta_{t-1}\|_{\psi^*_{t-1}} ,$$

where $\tilde{m}_{t+1} = \beta_1 \theta_{t-1} + (1-\beta_1) m_{t+1}$*,* $g_t := \nabla \ell_t(w_t)$*,* $\eta_{\min} := \min_t \eta_t$ *and* D^2_∞ *is the diameter of the bounded set* Θ*. The result holds for any benchmark* $w^* \in \Theta$ *and any step size sequence* $\{\eta_t\}_{t>0}$*.*

Proof Beforehand, we denote:

$$\begin{aligned} \tilde{g}_t &= \beta_1 \theta_{t-1} + (1-\beta_1) g_t , \\ \tilde{m}_{t+1} &= \beta_1 \theta_{t-1} + (1-\beta_1) m_{t+1} , \end{aligned} \tag{7}$$

where we recall that g_t and m_{t+1} are respectively the gradient $\nabla \ell_t(w_t)$ and the predictable guess. By regret decomposition, we have that

$$\begin{aligned} \mathcal{R}_T &:= \sum_{t=1}^{T} \ell_t(w_t) - \min_{w \in \Theta} \sum_{t=1}^{T} \ell_t(w) \\ &\leq \sum_{t=1}^{T} \langle w_t - w^*, \nabla \ell_t(w_t) \rangle \\ &= \sum_{t=1}^{T} \langle w_t - \tilde{w}_{t+1}, g_t - \tilde{m}_t \rangle + \langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle + \langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle + \langle \tilde{w}_{t+1} - w^*, g_t - \tilde{g}_t \rangle . \end{aligned} \tag{8}$$

Recall the notation $\psi_t(x)$ and the Bregman divergence $B_{\psi_t}(u, v)$ defined Section 4. We exploit a useful inequality (which appears in e.g., [35]). For any update of the form $\hat{w} = \arg\min_{w \in \Theta} \langle w, \theta \rangle + B_\psi(w, v)$, it holds that

$$\langle \hat{w} - u, \theta \rangle \leq B_\psi(u, v) - B_\psi(u, \hat{w}) - B_\psi(\hat{w}, v) \quad \text{for any } u \in \Theta . \tag{9}$$

For $\beta_1 = 0$, we can rewrite the update on line 8 of (Algorithm 2) as

$$\tilde{w}_{t+1} = \arg\min_{w \in \Theta} \eta_t \langle w, \tilde{g}_t \rangle + B_{\psi_t}(w, \tilde{w}_t) . \tag{10}$$

By using (9) for (10) with $\hat{w} = \tilde{w}_{t+1}$ (the output of the minimization problem), $u = w^*$ and $v = \tilde{w}_t$, we have

$$\langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle \leq \frac{1}{\eta_t} \left[B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t) \right] . \tag{11}$$

We can also rewrite the update on line 9 of (Algorithm 2) at time t as

$$w_{t+1} = \arg\min\nolimits_{w \in \Theta} \eta_{t+1} \langle w, \tilde{m}_{t+1} \rangle + B_{\psi_t}(w, \tilde{w}_{t+1}) . \tag{12}$$

and, by using (9) for (12) (written at iteration t), with $\hat{w} = w_t$ (the output of the minimization problem), $u = \tilde{w}_{t+1}$ and $v = \tilde{w}_t$, we have

$$\langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle \leq \frac{1}{\eta_t} \left[B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_{t-1}}(\tilde{w}_{t+1}, w_t) - B_{\psi_{t-1}}(w_t, \tilde{w}_t) \right] . \tag{13}$$

By (8), (11), and (13), we obtain

$$\begin{aligned} \mathcal{R}_T &\overset{(8)}{\leq} \sum_{t=1}^{T} \langle w_t - \tilde{w}_{t+1}, g_t - \tilde{m}_t \rangle + \langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle + \langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle + \langle \tilde{w}_{t+1} - w^*, g_t - \tilde{g}_t \rangle \\ &\overset{(11),(13)}{\leq} \sum_{t=1}^{T} \|w_t - \tilde{w}_{t+1}\|_{\psi_{t-1}} \|g_t - \tilde{m}_t\|_{\psi^*_{t-1}} + \|\tilde{w}_{t+1} - w^*\|_{\psi_{t-1}} \|g_t - \tilde{g}_t\|_{\psi^*_{t-1}} \\ &+ \frac{1}{\eta_t} \left[B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_{t-1}}(\tilde{w}_{t+1}, w_t) - B_{\psi_{t-1}}(w_t, \tilde{w}_t) \right. \\ &\left. + B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t) \right] , \end{aligned} \tag{14}$$

which is further bounded by

$$
\begin{aligned}
\mathcal{R}_T \leq \sum_{t=1}^{T} \Big\{ & \frac{1}{2\eta_t}\|w_t - \tilde{w}_{t+1}\|^2_{\psi_{t-1}} + \frac{\eta_t}{2}\|g_t - m_t\|^2_{\psi^*_{t-1}} + \|\tilde{w}_{t+1} - w^*\|_{\psi_{t-1}}\|g_t - \tilde{g}_t\|_{\psi^*_{t-1}} \\
& + \frac{1}{\eta_t}\big(\underbrace{B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t)}_{A_1} - \frac{1}{2}\|\tilde{w}_{t+1} - w_t\|^2_{\psi_{t-1}} \\
& + \underbrace{B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1})}_{A_2}\big)\Big\} ,
\end{aligned} \tag{15}
$$

where the inequality is due to $\|w_t - \tilde{w}_{t+1}\|_{\psi_{t-1}}\|g_t - m_t\|_{\psi^*_{t-1}} = \inf_{\beta>0} \frac{1}{2\beta}\|w_t - \tilde{w}_{t+1}\|^2_{\psi_{t-1}} + \frac{\beta}{2}\|g_t - m_t\|^2_{\psi^*_{t-1}}$ by Young's inequality and the 1-strongly convex of $\psi_{t-1}(\cdot)$ with respect to $\|\cdot\|_{\psi_{t-1}}$ which yields that $B_{\psi_{t-1}}(\tilde{w}_{t+1}, w_t) \geq \frac{1}{2}\|\tilde{w}_{t+1} - w_t\|^2_{\psi_t} \geq 0$.

To proceed, notice that

$$
\begin{aligned}
A_1 &:= B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t) \\
&= \langle \tilde{w}_{t+1} - \tilde{w}_t, \text{diag}(\hat{v}_{t-1}^{1/2} - \hat{v}_t^{1/2})(\tilde{w}_{t+1} - \tilde{w}_t)\rangle \leq 0 ,
\end{aligned} \tag{16}
$$

as the sequence $\{\hat{v}_t\}$ is non-decreasing. And that

$$
\begin{aligned}
A_2 := B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) &= \langle w^* - \tilde{w}_{t+1}, \text{diag}(\hat{v}_{t+1}^{1/2} - \hat{v}_t^{1/2})(w^* - \tilde{w}_{t+1})\rangle \\
&\leq (\max_i (w^*[i] - \tilde{w}_{t+1}[i])^2) \cdot (\sum_{i=1}^{d} \hat{v}_{t+1}^{1/2}[i] - \hat{v}_t^{1/2}[i]) .
\end{aligned} \tag{17}
$$

Therefore, by (15),(17),(16), we have

$$
\mathcal{R}_T \leq \frac{D_\infty^2}{\eta_{\min}} \sum_{i=1}^{d} \hat{v}_T^{1/2}[i] + \frac{B_{\psi_1}(w^*, \tilde{w}_1)}{\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2}\|g_t - \tilde{m}_t\|^2_{\psi^*_{t-1}} + D_\infty^2 \beta_1^2 \sum_{t=1}^{T} \|g_t - \theta_{t-1}\|_{\psi^*_{t-1}} ,
$$

since $\|g_t - \tilde{g}_t\|_{\psi^*_{t-1}} = \|g_t - \beta_1\theta_{t-1} - (1-\beta_1)g_t\|_{\psi^*_{t-1}} = \beta^2\|g_t - \theta_{t-1}\|_{\psi^*_{t-1}}$. This completes the proof.

□

B Proof of Corollary 1

Corollary. *Suppose $\beta_1 = 0$ and $\{v_t\}_{t>0}$ is a monotonically increasing sequence, then we obtain the following regret bound for any $w^* \in \Theta$ and sequence of stepsizes $\{\eta_t = \eta/\sqrt{t}\}_{t>0}$:*

$$
\mathcal{R}_T \leq \frac{B_{\psi_1}}{\eta_1} + \frac{\eta\sqrt{1 + \log T}}{\sqrt{1-\beta_2}} \sum_{i=1}^{d} \|(g-m)_{1:T}[i]\|_2 + \frac{D_\infty^2}{\eta_{\min}} \sum_{i=1}^{d} \left[(1-\beta_2)\sum_{s=1}^{T} \beta_2^{T-s} g_s^2[i]\right]^{1/2} ,
$$

where $B_{\psi_1} := B_{\psi_1}(w^*, \tilde{w}_1)$, $g_t := \nabla \ell_t(w_t)$ and $\eta_{\min} := \min_t \eta_t$.

Proof Recall the bound in Theorem 1:

$$
\mathcal{R}_T \leq \frac{B_{\psi_1}(w^*, \tilde{w}_1)}{\eta_1} + \sum_{t=1}^{T} \frac{\eta_t}{2}\|g_t - \tilde{m}_t\|^2_{\psi^*_{t-1}} + \frac{D_\infty^2}{\eta_{\min}} \sum_{i=1}^{d} \hat{v}_T^{1/2}[i] + D_\infty^2 \beta_1^2 \sum_{t=1}^{T} \|g_t - \theta_{t-1}\|_{\psi^*_{t-1}} .
$$

The second term reads:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\eta_t}{2} \|g_t - m_t\|^2_{\psi^*_{t-1}} \\
=&\sum_{t=1}^{T-1} \frac{\eta_t}{2} \|g_t - m_t\|^2_{\psi^*_{t-1}} + \eta_T \sum_{i=1}^{d} \frac{(g_T[i] - m_T[i])^2}{\sqrt{v_{T-1}[i]}} \\
=&\sum_{t=1}^{T-1} \frac{\eta_t}{2} \|g_t - m_t\|^2_{\psi^*_{t-1}} + \eta \sum_{i=1}^{d} \frac{(g_T[i] - m_T[i])^2}{\sqrt{T\big((1-\beta_2)\sum_{s=1}^{T-1} \beta_2^{T-1-s} (g_s[i] - m_s[i])^2\big)}} \\
\leq&\eta \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{(g_t[i] - m_t[i])^2}{\sqrt{t\big((1-\beta_2)\sum_{s=1}^{t-1} \beta_2^{t-1-s} (g_s[i] - m_s[i])^2\big)}} \ .
\end{aligned}$$

To interpret the bound, let us make a rough approximation such that $\sum_{s=1}^{t-1} \beta_2^{t-1-s} (g_s[i] - m_s[i])^2 \simeq (g_t[i] - m_t[i])^2$. Then, we can further get an upper-bound as

$$\sum_{t=1}^{T} \frac{\eta_t}{2} \|g_t - m_t\|^2_{\psi^*_{t-1}} \leq \frac{\eta}{\sqrt{1-\beta_2}} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{|g_t[i] - m_t[i]|}{\sqrt{t}} \leq \frac{\eta \sqrt{1 + \log T}}{\sqrt{1-\beta_2}} \sum_{i=1}^{d} \|(g-m)_{1:T}[i]\|_2,$$

where the last inequality is due to Cauchy-Schwarz.

□

C Proofs of Auxiliary Lemmas

Following [38] and their study of the SGD with Momentum we denote for any $t > 0$:

$$\overline{w}_t = w_t + \frac{\beta_1}{1-\beta_1}(w_t - \tilde{w}_{t-1}) = \frac{1}{1-\beta_1} w_t - \frac{\beta_1}{1-\beta_1} \tilde{w}_{t-1} \ . \tag{18}$$

Lemma 3. *Assume a strictly positive and non increasing sequence of stepsizes $\{\eta_t\}_{t>0}$, $\beta_1 < \beta_2 \in [0, 1)$, then the following holds:*

$$\overline{w}_{t+1} - \overline{w}_t \leq \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \left[\eta_{t-1} \hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2} \right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \ ,$$

where $\tilde{\theta}_t = \theta_t + \beta_1 \theta_{t-1}$ and $\tilde{g}_t = g_t - \beta_1 m_t + \beta_1 g_{t-1} + m_{t+1}$.

Proof By definition (18) and using the Algorithm updates, we have:

$$\begin{aligned}
\overline{w}_{t+1} - \overline{w}_t &= \frac{1}{1-\beta_1}(w_{t+1} - \tilde{w}_t) - \frac{\beta_1}{1-\beta_1}(w_t - \tilde{w}_{t-1}) \\
&= -\frac{1}{1-\beta_1} \eta_t \hat{v}_t^{-1/2} (\theta_t + h_{t+1}) + \frac{\beta_1}{1-\beta_1} \eta_{t-1} \hat{v}_{t-1}^{-1/2} (\theta_{t-1} + h_t) \\
&= -\frac{1}{1-\beta_1} \eta_t \hat{v}_t^{-1/2} (\theta_t + \beta_1 \theta_{t-1}) - \frac{1}{1-\beta_1} \eta_t \hat{v}_t^{-1/2} (1-\beta_1) m_{t+1} \\
&\quad + \frac{\beta_1}{1-\beta_1} \eta_{t-1} \hat{v}_{t-1}^{-1/2} (\theta_{t-1} + \beta_1 \theta_{t-2}) + \frac{\beta_1}{1-\beta_1} \eta_{t-1} \hat{v}_{t-1}^{-1/2} (1-\beta_1) m_t \ .
\end{aligned} \tag{19}$$

Denote $\tilde{\theta}_t = \theta_t + \beta_1 \theta_{t-1}$ and $\tilde{g}_t = g_t - \beta_1 m_t + \beta_1 g_{t-1} + m_{t+1}$. Notice that $\tilde{\theta}_t = \beta_1 \tilde{\theta}_{t-1} + (1 - \beta_1)(g_t + \beta_1 g_{t-1})$.

$$\overline{w}_{t+1} - \overline{w}_t \leq \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \left[\eta_{t-1} \hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2} \right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \ . \tag{20}$$

□

Lemma 4. *Assume H4, a strictly positive and a sequence of constant stepsizes $\{\eta_t\}_{t>0}$, $\beta \in [0,1]$, then the following holds:*

$$\sum_{t=1}^{T_{\mathsf{M}}} \eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2}\theta_t\right\|_2^2\right] \leq \frac{\eta^2 d T_{\mathsf{M}}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} . \tag{21}$$

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

$$\begin{aligned}
\eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2}\theta_t\right\|_2^2\right] &= \eta_t^2 \mathbb{E}\left[\sum_{p=1}^{d} \frac{\theta_{t,p}^2}{\hat{v}_{t,p}}\right] \\
&\leq \eta_t^2 \mathbb{E}\left[\sum_{i=1}^{d} \frac{\theta_{t,p}^2}{v_{t,p}}\right] \\
&\leq \eta_t^2 \mathbb{E}\left[\sum_{i=1}^{d} \frac{(\sum_{r=1}^{t}(1-\beta_1)\beta_1^{t-r} g_{r,p})^2}{\sum_{r=1}^{t}(1-\beta_2)\beta_2^{t-r} g_{r,p}^2}\right] ,
\end{aligned} \tag{22}$$

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

$$\begin{aligned}
\eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2}\theta_t\right\|_2^2\right] &\leq \frac{\eta_t^2(1-\beta_1)^2}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d} \frac{(\sum_{r=1}^{t}\beta_1^{t-r} g_{r,p})^2}{\sum_{r=1}^{t}\beta_2^{t-r} g_{r,p}^2}\right] \\
&\overset{(a)}{\leq} \frac{\eta_t^2(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d} \frac{\sum_{r=1}^{t}\beta_1^{t-r} g_{r,p}^2}{\sum_{r=1}^{t}\beta_2^{t-r} g_{r,p}^2}\right] \\
&\leq \frac{\eta_t^2(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{i=1}^{d}\sum_{r=1}^{t} \gamma^{t-r}\right] = \frac{\eta_t^2 d(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{r=1}^{t} \gamma^{t-r}\right] ,
\end{aligned} \tag{23}$$

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

$$\begin{aligned}
\sum_{t=1}^{T_{\mathsf{M}}} \eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2}\theta_t\right\|_2^2\right] &\leq \frac{\eta_t^2 d(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{t=1}^{T_{\mathsf{M}}}\sum_{r=1}^{t} \gamma^{t-r}\right] \\
&\leq \frac{\eta^2 d T(1-\beta_1)}{1-\beta_2} \mathbb{E}\left[\sum_{t=t}^{t} \gamma^{t-r}\right] \\
&\leq \frac{\eta^2 d T(1-\beta_1)}{(1-\beta_2)(1-\gamma)} ,
\end{aligned} \tag{24}$$

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

C.1 Proof of Lemma 1

Lemma. *Assume assumption H4, then the quantities defined in Algorithm 2 satisfy for any $w \in \Theta$ and $t > 0$:*

$$\|\nabla f(w_t)\| < \mathsf{M}, \quad \|\theta_t\| < \mathsf{M}, \quad \|\hat{v}_t\| < \mathsf{M}^2 .$$

Proof Assume assumption H4 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 $\|\theta_t\| \leq \mathsf{M}$ then we have

$$\|\theta_{t+1}\| = \|\beta_1\theta_t + (1-\beta_1)\, g_{t+1}\| \leq \beta_1 \|\theta_t\| + (1-\beta_1)\|g_{t+1}\| \leq \mathsf{M} . \tag{25}$$

Using the same induction reasoning we prove that

$$\|\hat{v}_{t+1}\| = \left\|\beta_2\hat{v}_t + (1-\beta_2)\, g_{t+1}^2\right\| \leq \beta_2 \|\hat{v}_t\| + (1-\beta_1)\left\|g_{t+1}^2\right\| \leq \mathsf{M}^2 . \tag{26}$$

□

D Proof of Theorem 2

Theorem. *Assume H2-H4, $(\beta_1, \beta_2) \in [0, 1]$ and a sequence of decreasing stepsizes $\{\eta_t\}_{t>0}$, then the following result holds:*

$$\mathbb{E}[\|\nabla f(w_T)\|^2] \leq \tilde{C}_1 \sqrt{\frac{d}{T_{\mathsf{M}}}} + \tilde{C}_2 \frac{1}{T_{\mathsf{M}}} \ , \tag{27}$$

where T is a random termination number distributed according to (4) and the constants are defined as follows:

$$\begin{aligned}
\tilde{C}_1 &= C_1 + \frac{\mathsf{M}}{(1 - a\beta_1) + (\beta_1 + a)} \left[\frac{a(1-\beta_1)^2}{1-\beta_2} + 2L \frac{1}{1-\beta_2} \right] \\
C_1 &= \frac{\mathsf{M}}{(1 - a\beta_1) + (\beta_1 + a)} \Delta f + \frac{4L \left(\frac{\beta_1}{1-\beta_1} \right)^2 \mathsf{M}}{(1 - a\beta_1) + (\beta_1 + a)} \frac{(1+\beta_1^2)(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\
\tilde{C}_2 &= \frac{\mathsf{M}}{(1-\beta_1)\left((1 - a\beta_1) + (\beta_1 + a)\right)} \tilde{\mathsf{M}}^2 \mathbb{E}[\|\hat{v}_0^{-1/2}\|]
\end{aligned} \tag{28}$$

Proof Using H2 and the iterate $\overline{w}_t$ we have:

$$\begin{aligned}
f(\overline{w}_{t+1}) \leq & f(\overline{w}_t) + \nabla f(\overline{w}_t)^\top (\overline{w}_{t+1} - \overline{w}_t) + \frac{L}{2} \|\overline{w}_{t+1} - \overline{w}_t\|^2 \\
\leq & f(\overline{w}_t) + \underbrace{\nabla f(w_t)^\top (\overline{w}_{t+1} - \overline{w}_t)}_{A} \\
& + \underbrace{\left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top (\overline{w}_{t+1} - \overline{w}_t)}_{B} + \frac{L}{2} \|\overline{w}_{t+1} - \overline{w}_t\| \ .
\end{aligned} \tag{29}$$

Term A. Using Lemma 3, we have that:

$$\begin{aligned}
\nabla f(w_t)^\top (\overline{w}_{t+1} - \overline{w}_t) &\leq \nabla f(w_t)^\top \left[\frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \left[\eta_{t-1} \hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2} \right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \right] \\
&\leq \frac{\beta_1}{1-\beta_1} \|\nabla f(w_t)\| \|\eta_{t-1} \hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2}\| \|\tilde{\theta}_{t-1}\| - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \ ,
\end{aligned}$$

where the inequality is due to trivial inequality for positive diagonal matrix. Using Lemma 1 and assumption H3 we obtain:

$$\nabla f(w_t)^\top (\overline{w}_{t+1} - \overline{w}_t) \leq \frac{\beta_1 (1+\beta_1)}{1-\beta_1} \mathsf{M}^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \ , \tag{30}$$

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

$$\begin{aligned}
-\nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} \tilde{g}_t = & -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t - \nabla f(w_t)^\top \left[\eta_t \hat{v}_t^{-1/2} - \eta_t \hat{v}_t^{-1/2} \right] \bar{g}_t \\
& - \nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} (\beta_1 g_{t-1} + m_{t+1}) \\
\leq & -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t + (1 - a\beta_1) \mathsf{M}^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] \\
& - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} (\beta_1 g_{t-1} + m_{t+1}) \ ,
\end{aligned} \tag{31}$$

where we have used Lemma 1 on $\|g_t\|$ and where that $\tilde{g}_t = \bar{g}_t + \beta_1 g_{t-1} + m_{t+1} = g_t - \beta_1 m_t + \beta_1 g_{t-1} + m_{t+1}$. Plugging (31) into (30) yields:

$$\begin{aligned}
& \nabla f(w_t)^\top (\overline{w}_{t+1} - \overline{w}_t) \\
& \leq -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t + \frac{1}{1-\beta_1} (a\beta_1^2 - 2a\beta_1 + \beta 1) \mathsf{M}^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] \\
& - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} (\beta_1 g_{t-1} + m_{t+1}) \ .
\end{aligned} \tag{32}$$

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

$$\left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top (\overline{w}_{t+1} - \overline{w}_t) \leq \|\nabla f(\overline{w}_t) - \nabla f(w_t)\| \|\overline{w}_{t+1} - \overline{w}_t\| \,. \tag{33}$$

Using smoothness assumption H2:

$$\begin{aligned} \|\nabla f(\overline{w}_t) - \nabla f(w_t)\| &\leq L\|\overline{w}_t - w_t\| \\ &\leq L\frac{\beta_1}{1-\beta_1}\|w_t - \tilde{w}_{t-1}\| \,. \end{aligned} \tag{34}$$

By Lemma 3 we also have:

$$\begin{aligned} \overline{w}_{t+1} - \overline{w}_t &= \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{t-1}\left[\eta_{t-1}\hat{v}_{t-1}^{-1/2} - \eta_t\hat{v}_t^{-1/2}\right] - \eta_t\hat{v}_t^{-1/2}\tilde{g}_t \\ &= \frac{\beta_1}{1-\beta_1}\tilde{\theta}_{t-1}\eta_{t-1}\hat{v}_{t-1}^{-1/2}\left[I - (\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\right] - \eta_t\hat{v}_t^{-1/2}\tilde{g}_t \\ &= \frac{\beta_1}{1-\beta_1}\left[I - (\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\right](\tilde{w}_{t-1} - w_t) - \eta_t\hat{v}_t^{-1/2}\tilde{g}_t \,, \end{aligned} \tag{35}$$

where the last equality is due to $\tilde{\theta}_{t-1}\eta_{t-1}\hat{v}_{t-1}^{-1/2} = \tilde{w}_{t-1} - w_t$ by construction of $\tilde{\theta}_t$. Taking the norms on both sides, observing $\|I - (\eta_t\hat{v}_t^{-1/2})(\eta_{t-1}\hat{v}_{t-1}^{-1/2})^{-1}\| \leq 1$ due to the decreasing stepsize and the construction of $\hat{v}_t$ and using CS inequality yield:

$$\|\overline{w}_{t+1} - \overline{w}_t\| \leq \frac{\beta_1}{1-\beta_1}\|\tilde{w}_{t-1} - w_t\| + \|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\| \,. \tag{36}$$

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

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

Plugging (34) and (36) into (33) returns:

$$\begin{aligned} \left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top (\overline{w}_{t+1} - \overline{w}_t) \leq & L\frac{\beta_1}{1-\beta_1}\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\|w_t - \tilde{w}_{t-1}\| \\ &+ L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1} - w_t\|^2 \,. \end{aligned}$$

Applying Young's inequality with $\delta \to \frac{\beta_1}{1-\beta_1}$ on the product $\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\|w_t - \tilde{w}_{t-1}\|$ yields:

$$\left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top (\overline{w}_{t+1} - \overline{w}_t) \leq L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1} - w_t\|^2 \,. \tag{37}$$

The last term $\frac{L}{2}\|\overline{w}_{t+1} - \overline{w}_t\|$ can be upper bounded using (36):

$$\begin{aligned} \frac{L}{2}\|\overline{w}_{t+1} - \overline{w}_t\|^2 &\leq \frac{L}{2}\left[\frac{\beta_1}{1-\beta_1}\|\tilde{w}_{t-1} - w_t\| + \|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|\right] \\ &\leq L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 2L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1} - w_t\|^2 \,. \end{aligned} \tag{38}$$

Plugging (32), (37) and (38) into (29) and taking the expectations on both sides give:

$$\begin{aligned} &\mathbb{E}\left[f(\overline{w}_{t+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\|\eta_t\hat{v}_t^{-1/2}\| - \left(f(\overline{w}_t) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2\|\eta_{t-1}\hat{v}_{t-1}^{-1/2}\|\right)\right] \\ &\leq \mathbb{E}\left[-\nabla f(w_t)^\top\eta_{t-1}\hat{v}_{t-1}^{-1/2}\bar{g}_t - \nabla f(w_t)^\top\eta_t\hat{v}_t^{-1/2}(\beta_1 g_{t-1} + m_{t+1})\right] \\ &+ \mathbb{E}\left[2L\|\eta_t\hat{v}_t^{-1/2}\tilde{g}_t\|^2 + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2\|\tilde{w}_{t-1} - w_t\|^2\right] \,, \end{aligned}$$

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

$$\mathbb{E}\left[\nabla f(w_t)^\top \bar{g}_t\right] = \mathbb{E}\left[\nabla f(w_t)^\top (g_t - \beta_1 m_t)\right] = (1 - a\beta_1)\|\nabla f(w_t)\|^2 \, . \tag{39}$$

Summing from $t = 1$ to $t = T$ leads to

$$\begin{aligned}
&\frac{1}{\mathsf{M}} \sum_{t=1}^{T_\mathsf{M}} \left((1 - a\beta_1)\eta_{t-1} + (\beta_1 + a)\eta_t\right) \|\nabla f(w_t)\|^2 \leq \\
&\mathbb{E}\left[f(\overline{w}_1) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2 \|\eta_0 \hat{v}_0^{-1/2}\| - \left(f(\overline{w}_{T_\mathsf{M}+1}) + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2 \|\eta_{T_\mathsf{M}} \hat{v}_{T_\mathsf{M}}^{-1/2}\|\right)\right] \\
&+ 2L \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}\left[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2\right] + 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}\left[\|\tilde{w}_{t-1} - w_t\|^2\right] \\
&\leq \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2 \|\eta_0 \hat{v}_0^{-1/2}\|\right] + 2L \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}\left[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2\right] \\
&+ 4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}\left[\|\tilde{w}_{t-1} - w_t\|^2\right] \, ,
\end{aligned} \tag{40}$$

where we denote $\Delta f := f(\overline{w}_1) - f(\overline{w}_{T_\mathsf{M}+1})$. We note that by definition of $\hat{v}_t$, and a constant learning rate η_t, we have

$$\begin{aligned}
\|\tilde{w}_{t-1} - w_t\|^2 &= \|\eta_{t-1} \hat{v}_{t-1}^{-1/2} (\theta_{t-1} + h_t)\|^2 \\
&= \|\eta_{t-1} \hat{v}_{t-1}^{-1/2} (\theta_{t-1} + \beta_1 \theta_{t-2} + (1-\beta_1) m_t)\|^2 \\
&\leq \|\eta_{t-1} \hat{v}_{t-1}^{-1/2} \theta_{t-1}\|^2 + \|\eta_{t-2} \hat{v}_{t-2}^{-1/2} \beta_1 \theta_{t-2}\|^2 + (1-\beta_1)^2 \|\eta_{t-1} \hat{v}_{t-1}^{-1/2} m_t\|^2 \, .
\end{aligned}$$

Using Lemma 4 we have

$$\begin{aligned}
&\sum_{t=1}^{T_\mathsf{M}} \mathbb{E}\left[\|\tilde{w}_{t-1} - w_t\|^2\right] \\
&\leq (1 + \beta_1^2) \frac{\eta^2 d T_\mathsf{M} (1-\beta_1)}{(1-\beta_2)(1-\gamma)} + (1-\beta_1)^2 \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\eta_{t-1} \hat{v}_{t-1}^{-1/2} m_t\|] \, .
\end{aligned}$$

And thus, setting the learning rate to a constant value η and injecting in (40) yields:

$$\begin{aligned}
\mathbb{E}[\|\nabla f(w_T)\|^2] &= \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} \sum_{t=1}^{T_\mathsf{M}} \eta_t \|\nabla f(w_t)\|^2 \\
&\leq \frac{\mathsf{M}}{(1-a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}^2 \|\eta_0 \hat{v}_0^{-1/2}\|\right] \\
&+ \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \mathsf{M}}{(1-a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} (1 + \beta_1^2) \frac{\eta^2 d T_\mathsf{M} (1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\
&+ \frac{\mathsf{M}}{(1-a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} (1-\beta_1)^2 \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\eta_{t-1} \hat{v}_{t-1}^{-1/2} m_t\|] \\
&+ \frac{2L\mathsf{M}}{(1-a\beta_1) + (\beta_1 + a)} \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2] \, ,
\end{aligned}$$

where T is a random termination number distributed according (4). Setting the stepsize to $\eta = \frac{1}{\sqrt{d T_\mathsf{M}}}$ yields :

$$\mathbb{E}[\|\nabla f(w_T)\|^2] \leq C_1 \sqrt{\frac{d}{T_\mathsf{M}}} + C_2 \frac{1}{T_\mathsf{M}} + D_1 \frac{\eta}{T_\mathsf{M}} \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\hat{v}_{t-1}^{-1/2} m_t\|] + D_2 \frac{\eta}{T_\mathsf{M}} \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\hat{v}_{t-1}^{-1/2} \tilde{g}_t\|] \, ,$$

where

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

Simple case as in [41]: if $\beta_1 = 0$ then $\tilde{g}_t = g_t + m_{t+1}$ and $g_t = \theta_t$. Also using Lemma 4 we have that:

$$\sum_{t=1}^{T_{\mathsf{M}}} \eta_t^2 \mathbb{E}\left[\left\|\hat{v}_t^{-1/2} g_t\right\|_2^2\right] \le \frac{\eta^2 d T_{\mathsf{M}}}{(1-\beta_2)} \,;$$

which leads to the final bound:

$$\mathbb{E}[\|\nabla f(w_T)\|^2] \le \tilde{C}_1\sqrt{\frac{d}{T_{\mathsf{M}}}} + \tilde{C}_2\frac{1}{T_{\mathsf{M}}} \,,$$

where

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

□

E Proof of Lemma 2 (Boundedness of the iterates)

Lemma. *Given the multilayer model (5), assume the boundedness of the input data and of the loss function,* i.e., *for any $\xi \in \mathbb{R}^p$ and $y \in \mathbb{R}$ there is a constant $T > 0$ such that:*

$$\|\xi\| \le 1 \quad a.s. \quad and |\mathcal{L}'(\cdot, y)| \le T \,, \tag{41}$$

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:*

$$\|w^{(\ell)}\| \le A_{(\ell)} \,.$$

Proof For any index $\ell \in [1, L]$ we denote the output of layer ℓ by

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

Given the sigmoid assumption we have $\|h^{(\ell)}(w, \xi)\| \le 1$ for any $\ell \in [1, L]$ and any $(w, \xi) \in \mathbb{R}^d \times \mathbb{R}^p$. We also recall that $\mathcal{L}(\cdot, y)$ is the loss function, which can be Huber loss or cross entropy. Observe that at the last layer L:

$$\begin{aligned}\|\nabla_{w^{(L)}\|\mathcal{L}(\mathsf{MLN}(w,\xi),y)} &= \|\mathcal{L}'(\mathsf{MLN}(w,\xi),y)\nabla_{w^{(L)}}\mathsf{MLN}(w,\xi)\| \\ &= \|\mathcal{L}'(\mathsf{MLN}(w,\xi),y)\sigma'(w^{(L)}h^{(L-1)}(w,\xi))h^{(L-1)}(w,\xi)\| \\ &\le \frac{T}{4} \,,\end{aligned} \tag{42}$$

where the last equality is due to mild assumptions (41) and to the fact that the norm of the derivative of the sigmoid function is upperbounded by $1/4$.

From Algorithm 2, and with $\beta_1 = 0$ for the sake of notation, we have for iteration index $t > 0$:

$$\begin{aligned}\|w_t - \tilde{w}_{t-1}\| = \|-\eta_t\hat{v}_t^{-1/2}(\theta_t + h_{t+1})\| &= \|\eta_t\hat{v}_t^{-1/2}(g_t + m_{t+1})\| \\ &\le \hat{\eta}\|\hat{v}_t^{-1/2}g_t\| + \hat{\eta}a\|\hat{v}_t^{-1/2}g_{t+1}\| \,,\end{aligned}$$

where $\hat{\eta} = \max_{t>0} \eta_t$. For any dimension $p \in [1, d]$, using assumption H3, we note that

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

Thus:

$$\|w_t - \tilde{w}_{t-1}\| \leq \hat{\eta} \left(\|\hat{v}_t^{-1/2} g_t\| + a \|\hat{v}_t^{-1/2} g_{t+1}\| \right) \leq \hat{\eta} \frac{a+1}{\sqrt{1-\beta_2}} \ .$$

In short there exist a constant B such that $\|w_t - \tilde{w}_{t-1}\| \leq B$.

Proof by induction: As in [9], we will prove the containment of the weights by induction. Suppose an iteration index T and a coordinate i of the last layer L such that $w_{T,i}^{(L)} \geq \frac{T}{4\lambda} + B$. Using (42), we have

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

where $f(w, \xi) = \mathcal{L}(\mathsf{MLN}(w, \xi), y) + \frac{\lambda}{2} \left\| w \right\|^2$ and is the loss of our MLN. This last equation yields $\theta_{T,i}^{(L)} \geq 0$ (given the algorithm and $\beta_1 = 0$) and using the fact that $\|w_t - \tilde{w}_{t-1}\| \leq B$ we have

$$0 \leq w_{T-1,i}^{(L)} - B \leq w_{T,i}^{(L)} \leq w_{T-1,i}^{(L)} \ , \tag{43}$$

which means that $|w_{T,i}^{(L)}| \leq w_{T-1,i}^{(L)}$. So if the first assumption of that induction reasoning holds, *i.e.,* $w_{T-1,i}^{(L)} \geq \frac{T}{4\lambda} + B$, then the next iterates $w_{T,i}^{(L)}$ decreases, see (43) and go below $\frac{T}{4\lambda} + B$. This yields that for any iteration index $t > 0$ we have

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

since B is the biggest jump an iterate can do since $\|w_t - \tilde{w}_{t-1}\| \leq B$. Likewise we can end up showing that

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

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 H1 by induction.

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

$$\nabla_{w^{(\ell)}} \mathcal{L}(\mathsf{MLN}(w, \xi), y) = \mathcal{L}'(\mathsf{MLN}(w, \xi), y) \left(\prod_{j=1}^{\ell+1} \sigma' \left(w^{(j)} h^{(j-1)}(w, \xi) \right) \right) h^{(\ell-1)}(w, \xi) \ . \tag{44}$$

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 $\left\| w^{(r)} \right\|_F \leq F_r$ for any layer $r > \ell$. Then having this gradient (44) 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

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

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. $\square$

F Comparison to some related methods

Comparison to nonconvex optimization works. Recently, [39, 5, 37, 41, 42, 22] provide some theoretical analysis of ADAM-type algorithms when applying them to smooth nonconvex optimization problems. For example, [5] provides a bound, which is $\min_{t\in[T]} \mathbb{E}[\|\nabla f(w_t)\|^2] = \mathcal{O}(\log T/\sqrt{T})$. Yet, this data independent bound does not show any advantage over standard stochastic gradient descent. Similar concerns appear in other papers.

To get some adaptive data dependent bound that are in terms of the gradient norms observed along the trajectory) when applying OPT-AMSGRAD to nonconvex optimization, one can follow the approach of [2] or [6]. They provide ways to convert algorithms with adaptive data dependent regret bound for convex loss functions (e.g. ADAGRAD) to the ones that can find an approximate stationary point of nonconvex loss functions. Their approaches are modular so that simply using OPT-AMSGRAD as the base algorithm in their methods will immediately lead to a variant of OPT-AMSGRAD that enjoys some guarantee on nonconvex optimization. The variant can outperform the ones instantiated by other ADAM-type algorithms when the gradient prediction m_t is close to g_t. The details are omitted since this is a straightforward application.

Comparison to AO-FTRL [26]. In [26], the authors propose AO-FTRL, which has the update of the form $w_{t+1} = \arg\min_{w\in\Theta} (\sum_{s=1}^{t} g_s)^\top w + m_{t+1}^\top w + r_{0:t}(w)$, where $r_{0:t}(\cdot)$ is a 1-strongly convex loss function with respect to some norm $\|\cdot\|_{(t)}$ that may be different for different iteration t. Data dependent regret bound was provided in the paper, which is $r_{0:T}(w^*) + \sum_{t=1}^{T} \|g_t - m_t\|_{(t)^*}$ for any benchmark $w^* \in \Theta$. We see that if one selects $r_{0:t}(w) := \langle w, \mathrm{diag}\{\hat{v}_t\}^{1/2} w\rangle$ and $\|\cdot\|_{(t)} := \sqrt{\langle \cdot, \mathrm{diag}\{\hat{v}_t\}^{1/2} \cdot \rangle}$, then the update might be viewed as an optimistic variant of ADAGRAD. However, no experiments was provided in [26].

Comparison to OPTIMISTIC-ADAM [8]. We are aware that [8] proposed one version of optimistic algorithm for ADAM, which is called OPTIMISTIC-ADAM in their paper. A slightly modified version is summarized in Algorithm 4. Here, OPTIMISTIC-ADAM+$\hat{v}_t$ is OPTIMISTIC-ADAM in [8] with the additional max operation $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ to guarantee that the weighted second moment is monotone increasing.

Algorithm 4 OPTIMISTIC-ADAM [8]+$\hat{v}_t$.

```
Required: parameter β1, β2, and ηt.
Init: w1 ∈ Θ and v̂0 = v0 = ε1 ∈ R^d.
for t = 1 to T do
    Get mini-batch stochastic gradient vector gt ∈ R^d at wt.
    θt = β1 θt−1 + (1 − β1) gt.
    vt = β2 vt−1 + (1 − β2) gt^2.
    v̂t = max(v̂t−1, vt).
    wt+1 = Πk[wt − 2ηt θt/√v̂t + ηt θt−1/√v̂t−1].
end for
```

1. Required: parameter β_1, β_2, and η_t.
2. Init: $w_1 \in \Theta$ and $\hat{v}_0 = v_0 = \epsilon 1 \in \mathbb{R}^d$.
3. **for** $t = 1$ to T **do**
4. Get mini-batch stochastic gradient vector $g_t \in \mathbb{R}^d$ at w_t.
5. $\theta_t = \beta_1 \theta_{t-1} + (1-\beta_1) g_t$.
6. $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$.
7. $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$.
8. $w_{t+1} = \Pi_k[w_t - 2\eta_t \frac{\theta_t}{\sqrt{\hat{v}_t}} + \eta_t \frac{\theta_{t-1}}{\sqrt{\hat{v}_{t-1}}}]$.
9. **end for**

We want to emphasize that the motivations are different. OPTIMISTIC-ADAM in their paper is designed to optimize two-player games (e.g. GANs [15]), while the proposed algorithm in this paper is designed to accelerate optimization (e.g. solving empirical risk minimization quickly). [8] focuses on training GANs [15]. GANs is a two-player zero-sum game. There have been some related works in OPTIMISTIC ONLINE LEARNING like [7, 29, 33]) showing that if both players use some kinds of OPTIMISTIC-update, then accelerating the convergence to the equilibrium of the game is possible. [8] was inspired by these related works and showed that OPTIMISTIC-MIRROR-DESCENT can avoid the cycle behavior in a bilinear zero-sum game, which accelerates the convergence. Furthermore, [8] did not provide theoretical analysis of OPTIMISTIC-ADAM.

G Additional Remarks and Runs on the Gradient Prediction Process

Two illustrative examples. We provide two toy examples to demonstrate how OPT-AMSGRAD works with the chosen extrapolation method. First, consider minimizing a quadratic function $H(w) := \frac{b}{2}w^2$ with vanilla gradient descent method $w_{t+1} = w_t - \eta_t \nabla H(w_t)$. The gradient $g_t := \nabla H(w_t)$ has a recursive description as $g_{t+1} = bw_{t+1} = b(w_t - \eta_t g_t) = g_t - b\eta_t g_t$. So, the update can be written in the form of $g_t = Ag_{t-1} + \mathcal{O}(\|g_{t-1}\|_2^2)u_{t-1}$, with $A = (1 - b\eta)$ and $u_{t-1} = 0$ by setting $\eta_t = \eta$ (constant step size). Therefore, the extrapolation method should predict well.

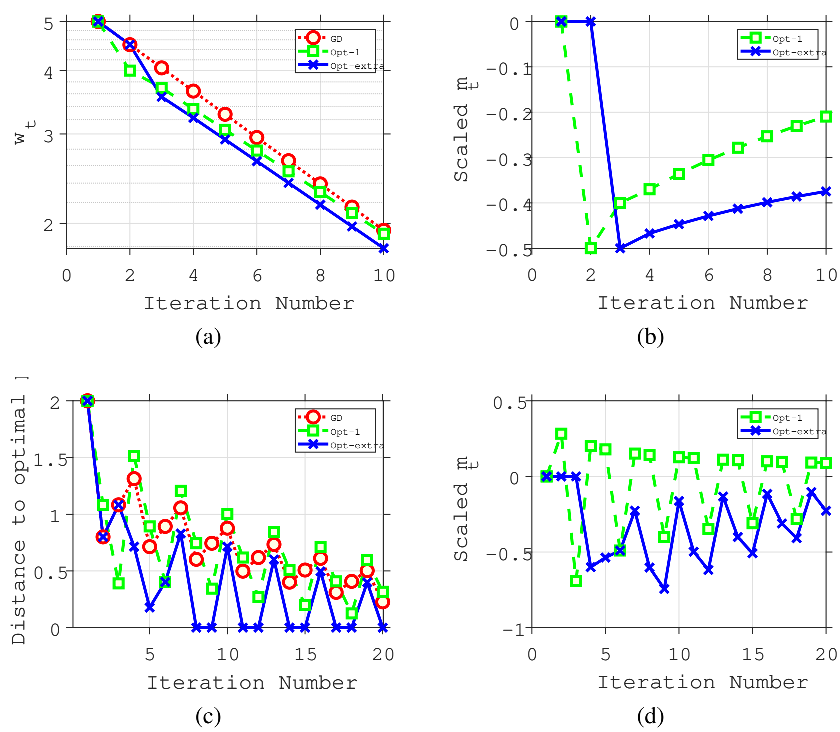

Figure 5: (a): The iterate w_t; the closer to the optimal point 0 the better. (b): A scaled and clipped version of m_t: $w_t - w_{t-1/2}$, which measures how the prediction of m_t drives the update towards the optimal point. In this scenario, the more negative the better. (c): Distance to the optimal point -1. The smaller the better. (d): A scaled and clipped version of m_t: $w_t - w_{t-1/2}$, which measures how the prediction of m_t drives the update towards the optimal point. In this scenario, the more negative the better.

Specifically, consider optimizing $H(w) := w^2/2$ by the following three algorithms with the same step size. One is Gradient Descent (GD): $w_{t+1} = w_t - \eta_t g_t$, while the other two are OPT-AMSGRAD with $\beta_1 = 0$ and the second moment term $\hat{v}_t$ being dropped: $w_{t+\frac{1}{2}} = \Pi_\Theta\left[w_{t-\frac{1}{2}} - \eta_t g_t\right]$, $w_{t+1} = \Pi_\Theta\left[w_{t+\frac{1}{2}} - \eta_{t+1} m_{t+1}\right]$. We denote the algorithm that sets $m_{t+1} = g_t$ as Opt-1, and denote the algorithm that uses the extrapolation method to get m_{t+1} as Opt-extra. We let $\eta_t = 0.1$ and the initial point $w_0 = 5$ for all the three methods. The simulation results are on Figure 5 (a) and (b). Sub-figure (a) plots update w_t over iteration, where the updates should go towards the optimal point 0. Sub-figure (b) is about a scaled and clipped version of m_t, defined as $w_t - w_{t-1/2}$, which can be viewed as $-\eta_t m_t$ if the projection (if exists) is lifted. Sub-figure (a) shows that Opt-extra converges faster than the other methods. Furthermore, sub-figure (b) shows that the prediction by the extrapolation method is better than the prediction by simply using the previous gradient. The sub-figure shows that $-m_t$ from both methods all point to 0 in all iterations and the magnitude is larger for the one produced by the extrapolation method after iteration 2. [2]

[2]The extrapolation needs at least two gradients for prediction. Thus, in the first two iterations, $m_t = 0$.

Now let us consider another problem: an online learning problem proposed in [30] [3]. Assume the learner's decision space is $\Theta = [-1, 1]$, and the loss function is $\ell_t(w) = 3w$ if t mod $3 = 1$, and $\ell_t(w) = -w$ otherwise. The optimal point to minimize the cumulative loss is $w^* = -1$. We let $\eta_t = 0.1/\sqrt{t}$ and the initial point $w_0 = 1$ for all the three methods. The parameter λ of the extrapolation method is set to $\lambda = 10^{-3} > 0$. The results are on Figure 5 (c) and (d). Sub-figure (c) shows that Opt-extra converges faster than the other methods while Opt-1 is not better than GD. The reason is that the gradient changes from -1 to 3 at t mod $3 = 1$ and it changes from 3 to -1 at t mod $3 = 2$. Consequently, using the current gradient as the guess for the next clearly is not a good choice, since the next gradient is in the opposite direction of the current one. Sub-figure (d) shows that $-m_t$ by the extrapolation method always points to $w^* = -1$, while the one by using the previous negative direction points to the opposite direction in two thirds of rounds. It shows that the extrapolation method is much less affected by the gradient oscillation and always makes the prediction in the right direction, which suggests that the method can capture the aggregate effect.

[3] [30] uses this example to show that ADAM [19] fails to converge.