

Supplementary Material: An Optimistic Acceleration of AMSGrad for Nonconvex Optimization

Appendix 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

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

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:

$$\begin{aligned} \tilde{g}_t &= \beta_1 \theta_{t-1} + (1 - \beta_1) g_t, \\ \tilde{m}_{t+1} &= \beta_1 \theta_{t-1} + (1 - \beta_1) m_{t+1}, \end{aligned} \tag{10}$$

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

$$\begin{aligned} \mathcal{R}_T &:= \sum_{t=1}^T \ell_t(w_t) - \min_{w \in \Theta} \sum_{t=1}^T \ell_t(w) \\ &\leq \sum_{t=1}^T \langle w_t - w^*, \nabla \ell_t(w_t) \rangle \\ &= \sum_{t=1}^T \langle w_t - \tilde{w}_{t+1}, g_t - \tilde{m}_t \rangle + \langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle + \langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle + \langle \tilde{w}_{t+1} - w^*, g_t - \tilde{g}_t \rangle. \end{aligned} \tag{11}$$

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., (Tseng, 2008)). For any update of the form $\hat{w} = \arg \min_{w \in \Theta} \langle w, \theta \rangle + B_{\psi}(w, v)$, it holds that

$$\langle \hat{w} - u, \theta \rangle \leq B_{\psi}(u, v) - B_{\psi}(u, \hat{w}) - B_{\psi}(\hat{w}, v) \quad \text{for any } u \in \Theta. \tag{12}$$

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

$$\tilde{w}_{t+1} = \arg \min_{w \in \Theta} \eta_t \langle w, \tilde{g}_t \rangle + B_{\psi_t}(w, \tilde{w}_t). \tag{13}$$

By using (12) for (13) with $\hat{w} = \tilde{w}_{t+1}$ (the output of the minimization problem), $u = w^*$ and $v = \tilde{w}_t$, we have

$$\langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle \leq \frac{1}{\eta_t} [B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t)]. \tag{14}$$

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

$$w_{t+1} = \arg \min_{w \in \Theta} \eta_{t+1} \langle w, \tilde{m}_{t+1} \rangle + B_{\psi_t}(w, \tilde{w}_{t+1}) . \quad (15)$$

and, by using (12) for (15) (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

$$\langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle \leq \frac{1}{\eta_t} [B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_{t-1}}(\tilde{w}_{t+1}, w_t) - B_{\psi_{t-1}}(w_t, \tilde{w}_t)] . \quad (16)$$

By (11), (14), and (16), we obtain

$$\begin{aligned} \mathcal{R}_T &\stackrel{(11)}{\leq} \sum_{t=1}^T \langle w_t - \tilde{w}_{t+1}, g_t - \tilde{m}_t \rangle + \langle w_t - \tilde{w}_{t+1}, \tilde{m}_t \rangle + \langle \tilde{w}_{t+1} - w^*, \tilde{g}_t \rangle + \langle \tilde{w}_{t+1} - w^*, g_t - \tilde{g}_t \rangle \\ &\stackrel{(14), (16)}{\leq} \sum_{t=1}^T \|w_t - \tilde{w}_{t+1}\|_{\psi_{t-1}} \|g_t - \tilde{m}_t\|_{\psi_{t-1}^*} + \|\tilde{w}_{t+1} - w^*\|_{\psi_{t-1}} \|g_t - \tilde{g}_t\|_{\psi_{t-1}^*} \\ &\quad + \frac{1}{\eta_t} [B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_{t-1}}(\tilde{w}_{t+1}, w_t) - B_{\psi_{t-1}}(w_t, \tilde{w}_t) \\ &\quad + B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t)] , \end{aligned} \quad (17)$$

which is further bounded by

$$\begin{aligned} \mathcal{R}_T &\leq \sum_{t=1}^T \left\{ \frac{1}{2\eta_t} \|w_t - \tilde{w}_{t+1}\|_{\psi_{t-1}}^2 + \frac{\eta_t}{2} \|g_t - m_t\|_{\psi_{t-1}^*}^2 + \|\tilde{w}_{t+1} - w^*\|_{\psi_{t-1}} \|g_t - \tilde{g}_t\|_{\psi_{t-1}^*} \right. \\ &\quad + \frac{1}{\eta_t} \underbrace{(B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t))}_{A_1} - \frac{1}{2} \|\tilde{w}_{t+1} - w_t\|_{\psi_{t-1}}^2 \\ &\quad \left. + \underbrace{(B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}))}_{A_2} \right\} , \end{aligned} \quad (18)$$

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

To proceed, notice that

$$\begin{aligned} A_1 &:= B_{\psi_{t-1}}(\tilde{w}_{t+1}, \tilde{w}_t) - B_{\psi_t}(\tilde{w}_{t+1}, \tilde{w}_t) \\ &= \langle \tilde{w}_{t+1} - \tilde{w}_t, \text{diag}(\hat{v}_{t-1}^{1/2} - \hat{v}_t^{1/2})(\tilde{w}_{t+1} - \tilde{w}_t) \rangle \leq 0 , \end{aligned} \quad (19)$$

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

$$\begin{aligned} A_2 &:= B_{\psi_t}(w^*, \tilde{w}_t) - B_{\psi_t}(w^*, \tilde{w}_{t+1}) = \langle w^* - \tilde{w}_{t+1}, \text{diag}(\hat{v}_{t+1}^{1/2} - \hat{v}_t^{1/2})(w^* - \tilde{w}_{t+1}) \rangle \\ &\leq (\max_i (w^*[i] - \tilde{w}_{t+1}[i])^2) \cdot \left(\sum_{i=1}^d \hat{v}_{t+1}^{1/2}[i] - \hat{v}_t^{1/2}[i] \right) . \end{aligned} \quad (20)$$

Therefore, by (18),(20),(19), we have

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

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. \square

Appendix 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:

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

The second term reads:

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

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

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

where the last inequality is due to Cauchy-Schwarz. \square

Appendix C. Proofs of Auxiliary Lemmas

Following (Yan et al., 2018) and their study of the SGD with Momentum we denote for any $t > 0$:

$$\bar{w}_t = w_t + \frac{\beta_1}{1 - \beta_1}(w_t - \tilde{w}_{t-1}) = \frac{1}{1 - \beta_1}w_t - \frac{\beta_1}{1 - \beta_1}\tilde{w}_{t-1}. \quad (21)$$

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:

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

where $\tilde{\theta}_t = \theta_t + \beta_1\theta_{t-1}$ and $\tilde{g}_t = g_t - \beta_1m_t + \beta_1g_{t-1} + m_{t+1}$.

Proof By definition (21) and using the Algorithm updates, we have:

$$\begin{aligned} \bar{w}_{t+1} - \bar{w}_t &= \frac{1}{1 - \beta_1}(w_{t+1} - \tilde{w}_t) - \frac{\beta_1}{1 - \beta_1}(w_t - \tilde{w}_{t-1}) \\ &= -\frac{1}{1 - \beta_1}\eta_t\hat{v}_t^{-1/2}(\theta_t + h_{t+1}) + \frac{\beta_1}{1 - \beta_1}\eta_{t-1}\hat{v}_{t-1}^{-1/2}(\theta_{t-1} + h_t) \\ &= -\frac{1}{1 - \beta_1}\eta_t\hat{v}_t^{-1/2}(\theta_t + \beta_1\theta_{t-1}) - \frac{1}{1 - \beta_1}\eta_t\hat{v}_t^{-1/2}(1 - \beta_1)m_{t+1} \\ &\quad + \frac{\beta_1}{1 - \beta_1}\eta_{t-1}\hat{v}_{t-1}^{-1/2}(\theta_{t-1} + \beta_1\theta_{t-2}) + \frac{\beta_1}{1 - \beta_1}\eta_{t-1}\hat{v}_{t-1}^{-1/2}(1 - \beta_1)m_t. \end{aligned} \quad (22)$$

Denote $\tilde{\theta}_t = \theta_t + \beta_1\theta_{t-1}$ and $\tilde{g}_t = g_t - \beta_1m_t + \beta_1g_{t-1} + m_{t+1}$. Notice that $\tilde{\theta}_t = \beta_1\tilde{\theta}_{t-1} + (1 - \beta_1)(g_t + \beta_1g_{t-1})$.

$$\bar{w}_{t+1} - \bar{w}_t \leq \frac{\beta_1}{1 - \beta_1}\tilde{\theta}_{t-1} \left[\eta_{t-1}\hat{v}_{t-1}^{-1/2} - \eta_t\hat{v}_t^{-1/2} \right] - \eta_t\hat{v}_t^{-1/2}\tilde{g}_t. \quad (23)$$

□

Lemma 4. Assume H4, a strictly positive and a sequence of constant stepsizes $\{\eta_t\}_{t>0}$, $(\beta_1, \beta_2) \in [0, 1]$, then the following holds:

$$\sum_{t=1}^{T_M} \eta_t^2 \mathbb{E} \left[\left\| \hat{v}_t^{-1/2} \theta_t \right\|_2^2 \right] \leq \frac{\eta^2 d T_M (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)}. \quad (24)$$

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:

$$\begin{aligned} \eta_t^2 \mathbb{E} \left[\left\| \hat{v}_t^{-1/2} \theta_t \right\|_2^2 \right] &= \eta_t^2 \mathbb{E} \left[\sum_{p=1}^d \frac{\theta_{t,p}^2}{\hat{v}_{t,p}} \right] \\ &\leq \eta_t^2 \mathbb{E} \left[\sum_{p=1}^d \frac{\theta_{t,p}^2}{v_{t,p}} \right] \\ &\leq \eta_t^2 \mathbb{E} \left[\sum_{p=1}^d \frac{(\sum_{r=1}^t (1 - \beta_1)\beta_1^{t-r} g_{r,p})^2}{\sum_{r=1}^t (1 - \beta_2)\beta_2^{t-r} g_{r,p}^2} \right], \end{aligned} \quad (25)$$

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

$$\begin{aligned}
 \eta_t^2 \mathbb{E} \left[\left\| \hat{v}_t^{-1/2} \theta_t \right\|_2^2 \right] &\leq \frac{\eta_t^2 (1 - \beta_1)^2}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \frac{(\sum_{r=1}^t \beta_1^{t-r} g_{r,p})^2}{\sum_{r=1}^t \beta_2^{t-r} g_{r,p}^2} \right] \\
 &\stackrel{(a)}{\leq} \frac{\eta_t^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \frac{\sum_{r=1}^t \beta_1^{t-r} g_{r,p}^2}{\sum_{r=1}^t \beta_2^{t-r} g_{r,p}^2} \right] \\
 &\leq \frac{\eta_t^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{i=1}^d \sum_{r=1}^t \gamma^{t-r} \right] = \frac{\eta_t^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{r=1}^t \gamma^{t-r} \right],
 \end{aligned} \tag{26}$$

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

$$\begin{aligned}
 \sum_{t=1}^{T_M} \eta_t^2 \mathbb{E} \left[\left\| \hat{v}_t^{-1/2} \theta_t \right\|_2^2 \right] &\leq \frac{\eta_t^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=1}^{T_M} \sum_{r=1}^t \gamma^{t-r} \right] \\
 &\leq \frac{\eta^2 d T (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=t}^t \gamma^{t-r} \right] \\
 &\leq \frac{\eta^2 d T (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)},
 \end{aligned} \tag{27}$$

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

C.1. Proof of Lemma 1

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

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

Proof Assume assumption H4 we have:

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

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

$$\|\theta_{t+1}\| = \|\beta_1 \theta_t + (1 - \beta_1) g_{t+1}\| \leq \beta_1 \|\theta_t\| + (1 - \beta_1) \|g_{t+1}\| \leq M. \tag{28}$$

Using the same induction reasoning we prove that

$$\|\hat{v}_{t+1}\| = \|\beta_2 \hat{v}_t + (1 - \beta_2) g_{t+1}^2\| \leq \beta_2 \|\hat{v}_t\| + (1 - \beta_1) \|g_{t+1}^2\| \leq M^2. \tag{29}$$

\square

Appendix D. Proof of Theorem 2

Theorem. Assume H1-H4, $\beta_1 < \beta_2 \in [0, 1)$ and a sequence of decreasing stepsizes $\{\eta_t\}_{t>0}$, then the following result holds:

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

where T is a random termination number distributed according (4). The constants are defined as:

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

where $\Delta f = f(\bar{w}_1) - f(\bar{w}_{T_M+1})$ and $a_m = \min_{t=1, \dots, T} a_t$.

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

$$\begin{aligned} f(\bar{w}_{t+1}) &\leq f(\bar{w}_t) + \nabla f(\bar{w}_t)^\top (\bar{w}_{t+1} - \bar{w}_t) + \frac{L}{2} \|\bar{w}_{t+1} - \bar{w}_t\|^2 \\ &\leq f(\bar{w}_t) + \underbrace{\nabla f(w_t)^\top (\bar{w}_{t+1} - \bar{w}_t)}_A \\ &\quad + \underbrace{(\nabla f(\bar{w}_t) - \nabla f(w_t))^\top (\bar{w}_{t+1} - \bar{w}_t)}_B + \frac{L}{2} \|\bar{w}_{t+1} - \bar{w}_t\|^2. \end{aligned} \tag{30}$$

Term A. Using Lemma 3, we have that:

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

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

$$\nabla f(w_t)^\top (\bar{w}_{t+1} - \bar{w}_t) \leq \frac{\beta_1(1 + \beta_1)}{1 - \beta_1} M^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} \tilde{g}_t, \tag{31}$$

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:

$$\begin{aligned} -\nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} \tilde{g}_t &= -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t - \nabla f(w_t)^\top [\eta_t \hat{v}_t^{-1/2} - \eta_{t-1} \hat{v}_{t-1}^{-1/2}] \bar{g}_t \\ &\quad - \nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} (\beta_1 g_{t-1} + m_{t+1}) \\ &\leq -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t + (1 - a_t \beta_1) M^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] \\ &\quad - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} (\beta_1 g_{t-1} + m_{t+1}), \end{aligned} \tag{32}$$

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 (32) into (31) yields:

$$\begin{aligned} & \nabla f(w_t)^\top (\bar{w}_{t+1} - \bar{w}_t) \\ & \leq -\nabla f(w_t)^\top \eta_{t-1} \hat{v}_{t-1}^{-1/2} \bar{g}_t + \frac{1}{1-\beta_1} (a_t \beta_1^2 - 2a_t \beta_1 + \beta_1) M^2 [\|\eta_{t-1} \hat{v}_{t-1}^{-1/2}\| - \|\eta_t \hat{v}_t^{-1/2}\|] \\ & \quad - \nabla f(w_t)^\top \eta_t \hat{v}_t^{-1/2} (\beta_1 g_{t-1} + m_{t+1}) . \end{aligned} \quad (33)$$

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

$$(\nabla f(\bar{w}_t) - \nabla f(w_t))^\top (\bar{w}_{t+1} - \bar{w}_t) \leq \|\nabla f(\bar{w}_t) - \nabla f(w_t)\| \|\bar{w}_{t+1} - \bar{w}_t\| . \quad (34)$$

Using smoothness assumption H2:

$$\begin{aligned} \|\nabla f(\bar{w}_t) - \nabla f(w_t)\| & \leq L \|\bar{w}_t - w_t\| \\ & \leq L \frac{\beta_1}{1-\beta_1} \|w_t - \tilde{w}_{t-1}\| . \end{aligned} \quad (35)$$

By Lemma 3 we also have:

$$\begin{aligned} \bar{w}_{t+1} - \bar{w}_t & = \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \left[\eta_{t-1} \hat{v}_{t-1}^{-1/2} - \eta_t \hat{v}_t^{-1/2} \right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \\ & = \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{t-1} \eta_{t-1} \hat{v}_{t-1}^{-1/2} \left[I - (\eta_t \hat{v}_t^{-1/2})(\eta_{t-1} \hat{v}_{t-1}^{-1/2})^{-1} \right] - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t \\ & = \frac{\beta_1}{1-\beta_1} \left[I - (\eta_t \hat{v}_t^{-1/2})(\eta_{t-1} \hat{v}_{t-1}^{-1/2})^{-1} \right] (\tilde{w}_{t-1} - w_t) - \eta_t \hat{v}_t^{-1/2} \tilde{g}_t , \end{aligned} \quad (36)$$

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:

$$\|\bar{w}_{t+1} - \bar{w}_t\| \leq \frac{\beta_1}{1-\beta_1} \|\tilde{w}_{t-1} - w_t\| + \|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\| . \quad (37)$$

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 (35) and (37) into (34) returns:

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

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

$$(\nabla f(\bar{w}_t) - \nabla f(w_t))^\top (\bar{w}_{t+1} - \bar{w}_t) \leq L \|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2 + 2L \left(\frac{\beta_1}{1-\beta_1} \right)^2 \|\tilde{w}_{t-1} - w_t\|^2 . \quad (38)$$

The last term $\frac{L}{2} \|\bar{w}_{t+1} - \bar{w}_t\|$ can be upper bounded using (37):

$$\begin{aligned} \frac{L}{2} \|\bar{w}_{t+1} - \bar{w}_t\|^2 &\leq \frac{L}{2} \left[\frac{\beta_1}{1 - \beta_1} \|\tilde{w}_{t-1} - w_t\| + \|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\| \right] \\ &\leq L \|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2 + 2L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \|\tilde{w}_{t-1} - w_t\|^2. \end{aligned} \quad (39)$$

Plugging (33), (38) and (39) into (30) and taking the expectations on both sides give:

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

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

$$\mathbb{E} \left[\nabla f(w_t)^\top \bar{g}_t \right] = \mathbb{E} \left[\nabla f(w_t)^\top (g_t - \beta_1 m_t) \right] = (1 - a_t \beta_1) \|\nabla f(w_t)\|^2. \quad (40)$$

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

$$\begin{aligned} &\frac{1}{M} \sum_{t=1}^{T_M} ((1 - a_t \beta_1) \eta_{t-1} + (\beta_1 + a_t) \eta_t) \|\nabla f(w_t)\|^2 \leq \\ &\mathbb{E} \left[f(\bar{w}_1) + \frac{1}{1 - \beta_1} \tilde{M}_t^2 \|\eta_0 \hat{v}_0^{-1/2}\| - \left(f(\bar{w}_{T_M+1}) + \frac{1}{1 - \beta_1} \tilde{M}_t^2 \|\eta_{T_M} \hat{v}_{T_M}^{-1/2}\| \right) \right] \\ &+ 2L \sum_{t=1}^{T_M} \mathbb{E} \left[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2 \right] + 4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \sum_{t=1}^{T_M} \mathbb{E} [\|\tilde{w}_{t-1} - w_t\|^2] \\ &\leq \mathbb{E} \left[\Delta f + \frac{1}{1 - \beta_1} \tilde{M}_t^2 \|\eta_0 \hat{v}_0^{-1/2}\| \right] + 2L \sum_{t=1}^{T_M} \mathbb{E} [\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2] \\ &+ 4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \sum_{t=1}^{T_M} \mathbb{E} [\|\tilde{w}_{t-1} - w_t\|^2], \end{aligned} \quad (41)$$

where we denote $\Delta f := f(\bar{w}_1) - f(\bar{w}_{T_M+1})$. 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

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

Assume $a_m = \min_{1, \dots, T_M} a_t$ and denote $\tilde{M}_m^2 = (a_m \beta_1^2 + \beta_1) M^2$. Setting a constant learning rate $\eta_t = \eta$ and plugging in (41) yields:

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

where T is a random termination number distributed according (4) and T_M is the maximum number of iteration. Setting the stepsize to $\eta = \frac{1}{\sqrt{dT_M}}$ yields :

$$\mathbb{E} [\|\nabla f(w_T)\|^2] \leq C_{1,m} \sqrt{\frac{d}{T_M}} + C_{2,m} \frac{1}{T_M} + \frac{\eta}{T_M} D_{1,m} \mathbb{E} [\|\hat{v}_{t-1}^{-1/2} m_t\|] + \frac{\eta}{T_M} D_{2,m} \mathbb{E} [\|\hat{v}_{t-1}^{-1/2} \tilde{g}_t\|] ,$$

where

$$\begin{aligned} C_{1,m} &= \frac{M}{(1 - a_m \beta_1) + (\beta_1 + a_m)} \Delta f + \frac{4L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 M}{(1 - a_m \beta_1) + (\beta_1 + a_m)} \frac{(1 + \beta_1^2)(1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} , \\ C_{2,m} &= \frac{M}{(1 - \beta_1) ((1 - a_m \beta_1) + (\beta_1 + a_m))} (a_m \beta_1^2 + \beta_1) M^2 \mathbb{E} [\|\hat{v}_0^{-1/2}\|] . \end{aligned}$$

Simple case as in Zhou et al. (2018): 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:

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

which leads to the final bound:

$$\mathbb{E} [\|\nabla f(w_T)\|^2] \leq \sqrt{\frac{d}{T_M}} \tilde{C}_{1,m} + \frac{1}{T_M} \tilde{C}_{2,m} ,$$

where

$$\begin{aligned}\tilde{C}_{1,m} &= C_{1,m} + \frac{\mathbf{M}}{(1 - a_m\beta_1) + (\beta_1 + a_m)} \left[\frac{a_m(1 - \beta_1)^2}{1 - \beta_2} + 2L \frac{1}{1 - \beta_2} \right] , \\ \tilde{C}_{2,m} &= C_{2,m} = \frac{\mathbf{M}}{(1 - \beta_1) ((1 - a_m\beta_1) + (\beta_1 + a_m))} \tilde{\mathbf{M}}_m^2 \mathbb{E}[\|\hat{v}_0^{-1/2}\|] .\end{aligned}$$

□