# 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*

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

*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* $D^2_\infty$ *is the diameter of the bounded set* $\Theta$. *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

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

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

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

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

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

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

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

which is further bounded by

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

where the inequality is due to $\|w_t - \tilde{w}_{t+1}\|_{\psi_{t-1}} \|g_t - \tilde{m}_t\|_{\psi^*_{t-1}} = \inf_{\beta>0} \frac{1}{2\beta}\|w_t - \tilde{w}_{t+1}\|^2_{\psi_{t-1}} + \frac{\beta}{2}\|g_t - \tilde{m}_t\|^2_{\psi^*_{t-1}}$ 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

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

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

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

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

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

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.

□

## 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 $\{\eta_t = \eta/\sqrt{t}\}_{t>0}$:*

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

*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:

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

The second term reads:

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

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

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

where the last inequality is due to Cauchy-Schwarz.

□

## Appendix C. Proofs of Auxiliary Lemmas

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

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

**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:*

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

*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 (21) and using the Algorithm updates, we have:

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

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 $\tilde{\theta}_t = \beta_1\tilde{\theta}_{t-1} + (1 - \beta_1)(g_t + \beta_1 g_{t-1})$.

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

□

**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:*

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

**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:

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

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

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

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

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

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

### 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

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

Using the same induction reasoning we prove that

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

□

## 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:*

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

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

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

*where $\Delta f = f(\overline{w}_1) - f(\overline{w}_{T_{\mathsf{M}}+1})$ and $a_m = \min_{t=1,\ldots,T} a_t$.*

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

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

**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:

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

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:

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

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:

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

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

$$\left(\nabla f(\overline{w}_t) - \nabla f(w_t)\right)^\top \left(\overline{w}_{t+1} - \overline{w}_t\right) \leq \|\nabla f(\overline{w}_t) - \nabla f(w_t)\|\|\overline{w}_{t+1} - \overline{w}_t\| \ . \tag{34}$$

Using smoothness assumption H2:

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

By Lemma 3 we also have:

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

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:

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

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

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

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:

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

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

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

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

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

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

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

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

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

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 $\eta_t$, 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

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

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

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(w_T)\|^2] &= \frac{1}{\sum_{j=1}^{T_\mathsf{M}} \eta_j} \sum_{t=1}^{T_\mathsf{M}} \eta_t \|\nabla f(w_t)\|^2 = \frac{\sum_1^{T_M} \|\nabla f(w_t)\|^2}{T_M} \\
&\leq \frac{\mathsf{M}}{T_M \eta((1-a_m\beta_1)+(\beta_1+a_m))} \mathbb{E}\left[\Delta f + \frac{1}{1-\beta_1}\tilde{\mathsf{M}}_m^2 \|\eta_0 \hat{v}_0^{-1/2}\|\right] \\
&+ \frac{4L\left(\frac{\beta_1}{1-\beta_1}\right)^2 \mathsf{M}}{T_M \eta((1-a_m\beta_1)+(\beta_1+a_m))}(1+\beta_1^2)\frac{\eta^2 d T_\mathsf{M}(1-\beta_1)}{(1-\beta_2)(1-\gamma)} \\
&+ \frac{\mathsf{M}}{T_M \eta((1-a_m\beta_1)+(\beta_1+a_m))}(1-\beta_1)^2 \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\eta_{t-1}\hat{v}_{t-1}^{-1/2} m_t\|] \\
&+ \frac{2L\mathsf{M}}{T_M \eta((1-a_m\beta_1)+(\beta_1+a_m))} \sum_{t=1}^{T_\mathsf{M}} \mathbb{E}[\|\eta_t \hat{v}_t^{-1/2} \tilde{g}_t\|^2] \ ,
\end{aligned}
$$

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

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

where

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

**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:

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

which leads to the final bound:

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

where

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

□