Proof for the convergence properties of ARSG in non-convex optimization

1 Main results

In this document, we prove the convergence properties of ARSG in non-convex optimization, which is more applicable for training deep neural works than the convergence proof for convex or strongly convex problems in the paper.

The proof is en extension of Chen et al. [2019] which analyzes the convergence bound of generalized ADAM (Algorithm 1) and AMSGRAD (Algorithm 3) as a special case. It is show that generalized ARSG (Algorithm 2) shares the form of the convergence bound of generalized ADAM, and improves the coefficients for typical hyper-parameters settings. Particularly, ARSG with the preconditioner of AMSGRAD (Algorithm 4) also shares the $O(\log(T)/\sqrt{T})$ convergence rate of AMSGRAD, whilst the coefficients are improved.

Algorithm 1. Generalized ADAM

S0. Initialize $m_0 = 0$ **and** x_1
For $t = 1, \cdots, T$**, do**
  S1. $m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t$
  S2. $v_t = h_t(g_1, g_2, ..., g_t)$
  S3. $x_{t+1} = x_t - \alpha_t m_t / \sqrt{v_t}$
End

Algorithm 2. Generalized ARSG

S0. Initialize $m_0 = 0$ **and** x_1
For $t = 1, \cdots, T$**, do**
  S1. $m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t$
  S2. $v_t = h_t(g_1, g_2, ..., g_t)$
  S3. $x_{t+1} = x_t - \alpha_t((1 - \mu_t) m_t + \mu_t g_t) / \sqrt{v_t}$
End

Algorithm 3. AMSGRAD

(S0). Initialize $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$;
For $t = 1, \cdots, T$, do
  (S1). $m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t$
  (S2). $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
  (S3). $\hat{v}_t = \max\{\hat{v}_{t-1}, v_t\}$
  (S4). $x_{t+1} = x_t - \alpha_t m_t / \sqrt{\hat{v}_t}$
End

Algorithm 4. ARSG

(S0). Initialize $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$;
For $t = 1, \cdots, T$, do
  (S1). $m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t$
  (S2). $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
  (S3). $\hat{v}_t = \max\{\hat{v}_{t-1}, v_t\}$
  (S4). $x_{t+1} = x_t - \alpha_t((1 - \mu_t) m_t + \mu_t g_t) / \sqrt{\hat{v}_t}$
End

In Algorithm 1, 2, 3, and 4, α_t is the step size at time t, $\beta_{1,t} > 0$ is a sequence of problem parameters, $m_t \in \mathbb{R}^d$ denotes some (exponentially weighted) gradient estimate, and $\hat{v}_t = h_t(g_1, g_2, ..., g_t) \in \mathbb{R}^d$ takes all the past gradients as input and returns a vector of dimension d, which is later used to inversely weight the gradient estimate m_t. And note that $m_t/\sqrt{\hat{v}_t} \in \mathbb{R}^d$ represents element-wise division. The vector $\alpha_t/\sqrt{\hat{v}_t}$ is referred to as the effective stepsize.

Notations: We use $z = x/y$ to denote element-wise division if x and y are both vectors of size d; $x \odot y$ is element-wise product, x^2 is element-wise square if x is a vector, $\sqrt{x}$ is element-wise square root if x is a vector, $(x)_j$ denotes jth coordinate of x, $\|x\|$ is $\|x\|_2$ if not otherwise specified. We use $[N]$ to denote the set $\{1, \cdots, N\}$, and use $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$ as standard asymptotic notations.

In the convergence analysis of ARSG, we make the same assumptions as those required for the generalized ADAM methods [Chen et al., 2019].

Assumptions:

A1: f is differentiable and has L-Lipschitz gradient, i.e. $\forall x, y$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. It is also lower bounded, i.e. $f(x^*) > -\infty$ where x^* is an optimal solution.

A2: At time t, the algorithm can access a bounded noisy gradient and the true gradient is bounded, i.e. $\|\nabla f(x_t)\| \leq H, \quad \|g_t\| \leq H, \quad \forall t > 1$.

A3: The noisy gradient is unbiased and the noise is independent, i.e. $g_t = \nabla f(x_t) + \zeta_t$, $E[\zeta_t] = 0$ and ζ_i is independent of ζ_j if $i \neq j$.

The convergence properties of generalized ARSG can be characterized as the following algorithm.

Theorem 1. Suppose that Assumptions A1-A3 are satisfied, β_1 is chosen such that $\beta_1 \geq \beta_{1,t}$, $\beta_{1,t} \in [0, 1)$ is non-increasing, $0 \leq \mu_t = \mu < 1$, and for some constant $G > 0$, $\left\|\alpha_t m_t/\sqrt{\hat{v}_t}\right\| \leq G, \left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\| \leq G, \forall\, t$. Then Algorithm2 yields

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T} \alpha_t \langle \nabla f(x_t), \nabla f(x_t)/\sqrt{\hat{v}_t}\rangle\right] \\
\leq &E\left[C_1 \sum_{t=1}^{T} \left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\|^2 + C_2 \sum_{t=2}^{T} \left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1 + C_3 \sum_{t=2}^{T-1} \left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|^2\right] + C_4,
\end{aligned} \tag{1}
$$

where C_1, C_2, C_3 are constants independent of d and T, C_4 is a constant independent of T, the expectation is taken with respect to all the randomness corresponding to $\{g_t\}$.

Further, let $\gamma_t := \min_{j\in[d]} \min_{\{g_i\}_{i=1}^t} \alpha_t/(\sqrt{\hat{v}_t})_j$, denote the minimum possible value of effective stepsize at time t over all possible coordinate and past gradients $\{g_i\}_{i=1}^t$. Then the convergence rate of Algorithm 2 is given by

$$
\min_{t\in[T]} E\left[\|\nabla f(x_t)\|^2\right] = O\left(\frac{s_1(T)}{s_2(T)}\right), \tag{2}
$$

where $s_1(T)$ is defined through the upper bound of RHS of equation (1), namely, $O(s_1(T))$, and $\sum_{t=1}^{T} \gamma_t = \Omega(s_2(T))$.

Proof. See Section 2.

Theorem 1 inherits the form of Theorem 3.1 in Chen et al. [2019], that shows convergence properties of generalized ADAM. However, the coefficients are different.

The coefficients in Theorem 1 are

$$
\begin{aligned}
C_1 \leq & L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+L\left(\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)+\frac{1}{2}\frac{\beta_1\mu}{1-\beta_1}\right)+\frac{1}{2}\\
C_2 \leq & H^2\left(\frac{\beta_1(1-\mu)}{1-\beta_1}+2\right)\\
C_3 \leq & L^2H^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}+\frac{3}{2}LH^2\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\frac{\beta_1^2}{(1-\beta_1)^2}\\
C_4 \leq & \frac{\beta_1(1-\mu)}{1-\beta_1}\left(H^2+G^2\right)+\frac{3}{2}LG^2\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2\\
& +\frac{\beta_1\mu}{1-\beta_1}HG+2H^2E\left[\|\alpha_1/\sqrt{\hat{v}_1}\|_1\right]+E\left[f(z_1)-f(z^*)\right],
\end{aligned}
\tag{3}
$$

where z^* is an optimal of f, i.e. $z^* \in \arg\min_z f(z)$.

The coefficients in Theorem 3.1 in Chen et al. [2019] are [1]

$$
\begin{aligned}
C_1 = & L^2\frac{\beta_1^2}{(1-\beta_1)^4}+\frac{3}{2}L+\frac{1}{2}\\
C_2 = & H^2\left(\frac{\beta_1}{1-\beta_1}+2\right)\\
C_3 = & L^2H^2\frac{\beta_1^4}{(1-\beta_1)^6}+\frac{3}{2}LH^2\frac{\beta_1^2}{(1-\beta_1)^2}\\
C_4 = & \left(\frac{\beta_1}{1-\beta_1}\right)\left(H^2+G^2\right)+\frac{3}{2}LG^2\left(\frac{\beta_1}{1-\beta_1}\right)^2+2H^2E\left[\|\alpha_1/\sqrt{\hat{v}_1}\|_1\right]+E\left[f(z_1)-f(z^*)\right],
\end{aligned}
\tag{4}
$$

where z^* is an optimal of f, i.e. $z^* \in \arg\min_z f(z)$.

It should be noted that although the effect of a large β_1 is negative in these coefficients in the bounds for the worst cases in general non-convex settings, β_1 close to 1 is required to gain fast convergence in the local quadratic approximation problems (as shown by Theorem 1 in the ARSG paper). Compared with generalized ADAM, when $1-\beta_1 \ll 1, L \gg 1, G \gg 1, H \gg 1$ in the typical situations, generalized ARSG has lower coefficients on C_1, C_2, and C_3. Although the constant term C_4 of generalized ARSG is possible to be slightly larger than that of generalized ADAM, it is irrelevant to T. Consequently, the effect of the possible increment in C_4 is weak when T is large.

From Theorem 1 we can obtain Corollary 1, that shows ARSG (Algorithm 4) yields $O(\log(T)/\sqrt{T})$ convergence rate in non-convex settings.

Corollary 1. Assume $\exists c > 0$ such that $|(g_1)_i| \geq c, \forall i \in [d]$, for ARSG (Algorithm 4) with $\beta_{1,t} \leq \beta_1 \in [0,1)$ and $\beta_{1,t}$ is non-increasing, $\alpha_t = 1/\sqrt{t}$, $0 \leq \mu_t = \mu < 1$, we have for any T,

$$
\min_{t\in[T]} E\left[\|f(x_t)\|^2\right] \leq \frac{1}{\sqrt{T}}\left(Q_1+Q_2\log T\right), \tag{5}
$$

where Q_1 and Q_2 are two constants independent of T, as

$$
\begin{aligned}
Q_1 &= H\left(C_1H^2/c^2+C_2d/c+C_3d/c^2+C_4\right)\\
Q_2 &= H(C_1H^2/c^2).
\end{aligned}
\tag{6}
$$

The coefficients C_1, C_2, C_3, C_4 are defined in equation (3).

Proof. See Section 3.

[1] In Chen et al. [2019], there are several typos when merging the similar terms in equation (39) in their paper. They miswrote $\left(\frac{\beta_1}{1-\beta_1}\right)^2\left(\frac{1}{1-\beta_1}\right)^2$ as $\frac{\beta_1}{1-\beta_1}\left(\frac{1}{1-\beta_1}\right)^2$. They also missed the coefficient $2L/3$ for the terms T_4 and T_5. Consequently, they obtained incorrect coefficients. We corrected the typos in the coefficients listed here.

The bound presented by Corollary 3.1 in [Chen et al., 2019] for AMSGRAD (Algorithm 4) is the same as equation (5) and (6), except for the coefficients C_1, C_2, C_3, C_4 are defined in equation (4). In typical cases where $1-\beta_1 \ll 1, L \gg 1, G \gg 1, H \gg 1$, ARSG improves the coefficients in the bound [2].

It should be noted that Algorithm 4 is slightly different from the original definition of ARSG in the paper, where we set $\hat{v}_0 = \epsilon > 0$. Since the gradients may be sparse in many problems especially when ReLU [Nair and Hinton, 2010] is selected as the activate function, the assumption $\exists\, c > 0$ such that $|(g_1)_i| \geq c, \forall i \in [d]$ may be violated. Then, ϵ in the original definition of ARSG serves as c, and the bound (5) and (6) still holds.

2 Proof of Theorem 1

2.1 Proof of auxiliary lemmas

Lemma 1. Let $x_0 = x_1$ in Algorithm 2, consider the sequence

$$z_t = x_t + \frac{\beta_{1,t}}{1-\beta_{1,t}}(x_t - x_{t-1}),\ \forall t \geq 1. \tag{7}$$

Then the following holds true

$$\begin{aligned} z_{t+1} - z_t =& -\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)m_t/\sqrt{\hat{v}_t} \\ &- \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right)\odot m_{t-1} \\ &+ \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1 + \frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right) g_t/\sqrt{\hat{v}_t}, \quad \forall t > 1 \end{aligned} \tag{8}$$

and

$$z_2 - z_1 = -\left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \alpha_1\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)m_1/\sqrt{\hat{v}_1} - \alpha_1\left(1 + \frac{\beta_{1,2}\mu}{1-\beta_{1,2}}\right) g_1/\sqrt{\hat{v}_1}. \tag{9}$$

[2] In the equation (5) and (6), C_1 and C_3 are divided by c^2, C_2 is divided by c, where c is a positive small number close to 0 in typical cases. Consequently, the negative effect of the possible increment in C_4 is negligible compared to the improvement brought in by the decrease of C_1, C_2, and C_3.

Proof. By the update rules S1-S3 in Algorithm 2, we have when $t > 1$,

$$
\begin{aligned}
& x_{t+1} - x_t \\
=& -\alpha_t \left((1-\mu)m_t + \mu g_t\right)/\sqrt{\hat{v}_t} \\
\overset{\text{S1}}{=}& -\alpha_t \left(\beta_{1,t}(1-\mu)m_{t-1} + ((1-\beta_{1,t})(1-\mu)+\mu)g_t\right)/\sqrt{\hat{v}_t} \\
\overset{\text{S3}}{=}& \beta_{1,t}\frac{\alpha_t}{\sqrt{\hat{v}_t}} \odot \left(\frac{\sqrt{\hat{v}_{t-1}}}{\alpha_{t-1}} \odot (x_t - x_{t-1}) + \mu g_{t-1}\right) - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t} \\
=& \beta_{1,t}(x_t - x_{t-1}) + \beta_{1,t}\left(\frac{\alpha_t}{\alpha_{t-1}}\frac{\sqrt{\hat{v}_{t-1}}}{\sqrt{\hat{v}_t}} - 1\right) \odot (x_t - x_{t-1}) \\
& + \alpha_t\beta_{1,t}\mu g_{t-1}/\sqrt{\hat{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t} \\
\overset{\text{S3}}{=}& \beta_{1,t}(x_t - x_{t-1}) - \beta_{1,t}\left(\frac{\alpha_t}{\alpha_{t-1}}\frac{\sqrt{\hat{v}_{t-1}}}{\sqrt{\hat{v}_t}} - 1\right) \odot \alpha_{t-1}\left((1-\mu)m_{t-1} + \mu g_{t-1}\right)/\sqrt{\hat{v}_{t-1}} \\
& + \alpha_t\beta_{1,t}\mu g_{t-1}/\sqrt{\hat{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t} \\
=& \beta_{1,t}(x_t - x_{t-1}) - \beta_{1,t}\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot \left((1-\mu)m_{t-1} + \mu g_{t-1}\right) \\
& + \alpha_t\beta_{1,t}\mu g_{t-1}/\sqrt{\hat{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t} \\
=& \beta_{1,t}(x_t - x_{t-1}) - \beta_{1,t}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \alpha_{t-1}\beta_{1,t}\mu g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t}.
\end{aligned} \tag{10}
$$

Since $x_{t+1} - x_t = (1-\beta_{1,t})x_{t+1} + \beta_{1,t}(x_{t+1} - x_t) - (1-\beta_{1,t})x_t$, based on equation (10) we have

$$
\begin{aligned}
& (1-\beta_{1,t})x_{t+1} + \beta_{1,t}(x_{t+1} - x_t) \\
=& (1-\beta_{1,t})x_t + \beta_{1,t}(x_t - x_{t-1}) - \beta_{1,t}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \alpha_{t-1}\beta_{1,t}\mu g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)g_t/\sqrt{\hat{v}_t}.
\end{aligned} \tag{11}
$$

Divide both sides by $1-\beta_{1,t}$, we have

$$
\begin{aligned}
& x_{t+1} + \frac{\beta_{1,t}}{1-\beta_{1,t}}(x_{t+1} - x_t) \\
=& x_t + \frac{\beta_{1,t}}{1-\beta_{1,t}}(x_t - x_{t-1}) - \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1 - \mu + \frac{\mu}{1-\beta_{1,t}}\right) g_t/\sqrt{\hat{v}_t}.
\end{aligned} \tag{12}
$$

According to the definition (7), Then equation (12) can be written as

$$
\begin{aligned}
& z_{t+1} \\
=& z_t + \left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)(x_{t+1} - x_t) - \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1-\mu+\frac{\mu}{1-\beta_{1,t}}\right) g_t/\sqrt{\hat{v}_t} \\
\overset{\text{S3}}{=}& z_t - \left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\left(\alpha_t\left((1-\mu)m_t + \mu g_t\right)/\sqrt{\hat{v}_t}\right) \\
& - \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1-\mu+\frac{\mu}{1-\beta_{1,t}}\right) g_t/\sqrt{\hat{v}_t} \\
=& z_t - \left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)m_t/\sqrt{\hat{v}_t} \\
& - \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left((1-\mu)+\frac{\mu}{1-\beta_{1,t+1}}\right) g_t/\sqrt{\hat{v}_t} \\
=& z_t - \left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)m_t/\sqrt{\hat{v}_t} \\
& - \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
& + \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}} g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1+\frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right) g_t/\sqrt{\hat{v}_t}, \quad \forall t > 1.
\end{aligned}
\tag{13}
$$

For $t = 1$, we have $z_1 = x_1$ (due to $x_1 = x_0$), and

$$
\begin{aligned}
& z_2 - z_1 \\
=& x_2 + \frac{\beta_{1,2}}{1-\beta_{1,2}}(x_2 - x_1) - x_1 \\
=& x_2 + \left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(x_2 - x_1) + \frac{\beta_{1,1}}{1-\beta_{1,1}}(x_2 - x_1) - x_1 \\
\overset{\text{S3}}{=}& \left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(-\alpha_1((1-\mu)m_1 + \mu g_1)/\sqrt{\hat{v}_1}) + \left(\frac{\beta_{1,1}}{1-\beta_{1,1}} + 1\right)(x_2 - x_1) \\
\overset{\text{S3}}{=}& \left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(-\alpha_1((1-\mu)m_1 + \mu g_1)/\sqrt{\hat{v}_1}) \\
& + \frac{1}{1-\beta_{1,1}}(-\alpha_1((1-\beta_{1,1})(1-\mu) + \mu)g_1/\sqrt{\hat{v}_1}) \\
=& -\left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)(\alpha_1 m_1/\sqrt{\hat{v}_1}) - \alpha_1\left((1-\mu) + \frac{1}{1-\beta_{1,2}}\mu\right) g_1/\sqrt{\hat{v}_1} \\
=& -\left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)(\alpha_1 m_1/\sqrt{\hat{v}_1}) - \alpha_1\left(1 + \frac{\beta_{1,2}\mu}{1-\beta_{1,2}}\right) g_1/\sqrt{\hat{v}_1}.
\end{aligned}
\tag{14}
$$

The proof is complete.

Without loss of generality, we initialize Algorithm 2 as below to simplify our analysis,

$$\left(\frac{\alpha_1}{\sqrt{\hat{v}_1}} - \frac{\alpha_0}{\sqrt{\hat{v}_0}}\right) \odot m_0 = 0, \quad \left(\frac{\alpha_1}{\sqrt{\hat{v}_1}} - \frac{\alpha_0}{\sqrt{\hat{v}_0}}\right) \odot g_0 = 0. \tag{15}$$

Lemma 2. Suppose that the conditions in Theorem 1 hold, then

$$E\left[f(z_{t+1}) - f(z_1)\right] \leq \sum_{i=1}^{6} T_i + T_{10}, \tag{16}$$

where

$$\begin{aligned}
T_1 &= -(1-\mu)E\left[\sum_{i=1}^{t}\langle \nabla f(z_i), \frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right) \odot m_{i-1}\rangle\right], \\
T_2 &= -E\left[\sum_{i=1}^{t}\alpha_i\langle \nabla f(z_i), g_i/\sqrt{\hat{v}_i}\rangle\right], \\
T_3 &= -(1-\mu)E\left[\sum_{i=1}^{t}\langle \nabla f(z_i), \left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}} - \frac{\beta_{1,i}}{1-\beta_{1,i}}\right)\alpha_i m_i/\sqrt{\hat{v}_i}\rangle\right], \\
T_4 &= E\left[\sum_{i=1}^{t}\frac{3}{2}L\left\|\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t m_t/\sqrt{v_t}\right\|^2\right], \\
T_5 &= E\left[\sum_{i=1}^{t}\frac{3}{2}L\left\|\frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right) \odot m_{i-1}\right\|^2\right], \\
T_6 &= \left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2} + 6\right)LE\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right] \\
T_{10} &= \mu E\left[\sum_{i=1}^{t-1}\left\langle \nabla f(z_{i+1}) - \nabla f(z_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i}\right\rangle\right] + \frac{\beta_1\mu}{1-\beta_1}HG.
\end{aligned} \tag{17}$$

Proof. By the Lipschitz smoothness of ∇f, we obtain

$$f(z_{t+1}) \leq f(z_t) + \langle \nabla f(z_t), d_t\rangle + \frac{L}{2}\left\|d_t\right\|^2, \tag{18}$$

where $d_t = z_{t+1} - z_t$, and Lemma 1 together with equation (18) yield

$$\begin{aligned}
d_t = &-\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)m_t/\sqrt{\hat{v}_t} \\
&- \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\hat{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right) \odot m_{t-1} \\
&+ \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1 + \frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)g_t/\sqrt{\hat{v}_t}, \quad \forall t \geq 1.
\end{aligned} \tag{19}$$

Based on equation (18) and equation (19), we then have

$$
\begin{aligned}
&E[f(z_{t+1}) - f(z_1)] = E\left[\sum_{i=1}^{t} f(z_{i+1}) - f(z_i)\right] \\
\leq & E\left[\sum_{i=1}^{t}\langle\nabla f(z_i), d_i\rangle + \frac{L}{2}\|d_i\|^2\right] \\
= & \underbrace{-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(z_i), \frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)\odot m_{i-1}\rangle\right]}_{T_1} \\
& - E\left[\sum_{i=1}^{t}\alpha_i\left(1+\frac{\beta_{1,i+1}\mu}{1-\beta_{1,i+1}}\right)\langle\nabla f(z_i), g_i/\sqrt{\hat{v}_i}\rangle\right] + E\left[\sum_{i=1}^{t}\alpha_{i-1}\frac{\beta_{1,i}\mu}{1-\beta_{1,i}}\langle\nabla f(z_i), g_{i-1}/\sqrt{\hat{v}_{i-1}}\rangle\right] \\
& \underbrace{-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(z_i), \left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}} - \frac{\beta_{1,i}}{1-\beta_{1,i}}\right)\alpha_i m_i/\sqrt{\hat{v}_i}\rangle\right]}_{T_3} + E\left[\sum_{i=1}^{t}\frac{L}{2}\|d_i\|^2\right] \\
= & T_1 \underbrace{-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(z_i), g_i/\sqrt{\hat{v}_i}\rangle\right]}_{T_2} + T_3 \\
& - \mu E\left[\sum_{i=1}^{t}\left\langle\nabla f(z_i), \left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i} - \frac{\beta_{1,i}}{1-\beta_{1,i}}\alpha_{i-1}g_{i-1}/\sqrt{\hat{v}_{i-1}}\right)\right\rangle\right] + E\left[\sum_{i=1}^{t}\frac{L}{2}\|d_i\|^2\right] \\
= & T_1 + T_2 + T_3 + E\left[\sum_{i=1}^{t}\frac{L}{2}\|d_i\|^2\right] \\
& + \mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(z_{i+1}) - \nabla f(z_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i}\right\rangle\right] - \mu E\left[\left\langle\nabla f(z_t), \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\alpha_t g_t/\sqrt{\hat{v}_t}\right\rangle\right] \\
\leq & T_1 + T_2 + T_3 + E\left[\sum_{i=1}^{t}\frac{L}{2}\|d_i\|^2\right] + \underbrace{\mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(z_{i+1}) - \nabla f(z_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i}\right\rangle\right] + \frac{\beta_1\mu HG}{1-\beta_1}}_{T_{10}} \\
= & T_1 + T_2 + T_3 + T_{10} + E\left[\sum_{i=1}^{t}\frac{L}{2}\|d_i\|^2\right],
\end{aligned}
\tag{20}
$$

where $\{T_i\}$ have been defined in equation (17), the last inequity is due to the assumption.

Further, using inequality $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and equation (20), we have

$$
\begin{aligned}
&\frac{L}{2}E\left[\sum_{i=1}^{t}\|d_i\|^2\right] \leq T_4 + T_5 + \frac{3}{2}E\left[\sum_{i=1}^{t}\left\|\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}g_{t-1}/\sqrt{\hat{v}_{t-1}} - \alpha_t\left(1+\frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)g_t/\sqrt{\hat{v}_t}\right\|^2\right] \\
\leq & T_4 + T_5 + 3\left(\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)^2 + \left(\frac{\beta_1\mu}{1-\beta_1}\right)^2\right)LE\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right] \\
\leq & T_4 + T_5 + \underbrace{\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2} + 6\right)LE\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]}_{T_6}.
\end{aligned}
\tag{21}
$$

Substituting the above inequality into equation (20), we then obtain equation (16).

The proof is complete.

The next series of lemmas separately bound the terms on RHS of equation (16) in Lemma 2.

Lemma 3. [Chen et al., 2019] Suppose that the conditions in Theorem 1 hold, T_1 in equation (17) can be bounded as

$$\begin{aligned}\frac{1}{1-\mu}T_1 =& -E\left[\sum_{i=1}^{t}\langle\nabla f(z_i), \frac{\beta_{1,t}}{1-\beta_{1,t}}\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)\odot m_{i-1}\rangle\right]\\ \leq& H^2\frac{\beta_1}{1-\beta_1}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right]\end{aligned} \tag{22}$$

Proof. Since $\|g_t\| \leq H$, by the update rule of m_t, we have $\|m_t\| \leq H$, this can be proved by induction as below.

Recall that $m_t = \beta_{1,t}m_{t-1} + (1-\beta_{1,t})g_t$, suppose $\|m_{t-1}\| \leq H$, we have

$$\|m_t\| \leq (\beta_{1,t} + (1-\beta_{1,t}))\max(\|g_t\|, \|m_{t-1}\|) = \max(\|g_t\|, \|m_{t-1}\|) \leq H, \tag{23}$$

then since $m_0 = 0$, we have $\|m_0\| \leq H$ which completes the induction.

Given $\|m_t\| \leq H$, we further have

$$\begin{aligned}\frac{1}{1-\mu}T_1 =& -E\left[\sum_{i=2}^{t}\langle\nabla f(z_i), \frac{\beta_{1,t}}{1-\beta_{1,t}}\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)\odot m_{i-1}\rangle\right]\\ \leq& E\left[\sum_{i=1}^{t}\|\nabla f(z_i)\|\,\|m_{i-1}\|\left(\frac{1}{1-\beta_{1,t}}-1\right)\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right]\\ \leq& H^2\frac{\beta_1}{1-\beta_1}E\left[\sum_{i=1}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right]\end{aligned} \tag{24}$$

where the first equality holds due to equation (15), and the last inequality is due to $\beta_1 \geq \beta_{1,i}$.

The proof is complete.

Lemma 4. [Chen et al., 2019] Suppose the conditions in Theorem 1 hold. For T_3 in equation (17), we have

$$\begin{aligned}\frac{1}{1-\mu}T_3 =& -E\left[\sum_{i=1}^{t}\langle\nabla f(z_i), \left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right)\alpha_i m_i/\sqrt{\hat{v}_i}\rangle\right]\\ \leq& \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\left(H^2+G^2\right).\end{aligned} \tag{25}$$

Proof.

$$\begin{aligned}\frac{1}{1-\mu}T_3 \leq& E\left[\sum_{i=1}^{t}\left|\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right|\frac{1}{2}\left(\|\nabla f(z_i)\|^2+\|\alpha_i m_i/\sqrt{\hat{v}_i}\|^2\right)\right]\\ \leq& E\left[\sum_{i=1}^{t}\left|\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right|\frac{1}{2}\left(H^2+G^2\right)\right]\\ =& \sum_{i=1}^{t}\left(\frac{\beta_{1,i}}{1-\beta_{1,i}}-\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\right)\frac{1}{2}\left(H^2+G^2\right)\\ \leq& \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\left(H^2+G^2\right),\end{aligned} \tag{26}$$

where the first inequality is due to$\langle a, b\rangle \leq \frac{1}{2}(\|a\|^2+\|b\|^2)$, the second inequality is using due to upper bound on $\|\nabla f(x_t)\| \leq H$ and $\|\alpha_i m_i/\sqrt{\hat{v}_i}\| \leq G$ given by the assumptions in Theorem 1, the

third equality is because $\beta_{1,t} \leq \beta_1$ and $\beta_{1,t}$ is non-increasing, the last inequality is due to telescope sum.

The proof is complete.

Lemma 5. [Chen et al., 2019] Suppose the assumptions in Theorem 1 hold. For T_4 in equation (17), we have

$$\begin{aligned} \frac{2}{3L}T_4 &= E\left[\sum_{i=1}^{t}\left\|\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t m_t/\sqrt{v_t}\right\|^2\right] \\ &\leq \left(\frac{\beta_1}{1-\beta_1} - \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2 \end{aligned} \tag{27}$$

Proof. The proof is similar to the previous lemma.

$$\begin{aligned} \frac{2}{3L}T_4 =& E\left[\sum_{i=1}^{t}\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)^2 \|\alpha_t m_t/\sqrt{v_t}\|^2\right] \\ \leq& E\left[\sum_{i=1}^{t}\left(\frac{\beta_{1,t}}{1-\beta_{1,t}} - \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2\right] \\ \leq& \left(\frac{\beta_1}{1-\beta_1} - \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\sum_{i=1}^{t}\left(\frac{\beta_{1,t}}{1-\beta_{1,t}} - \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right) G^2 \\ \leq& \left(\frac{\beta_1}{1-\beta_1} - \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2 \end{aligned} \tag{28}$$

where the first inequality is due to $\|\alpha_t m_t/\sqrt{v_t}\| \leq G$ by our assumptions, the second inequality is due to non-decreasing property of $\beta_{1,t}$ and $\beta_1 \geq \beta_{1,t}$, the last inequality is due to telescoping sum.

The proof is complete.

Lemma 6. [Chen et al., 2019] Suppose the assumptions in Theorem 1 hold. For T_5 in equation (17), we have

$$\begin{aligned} \frac{2}{3L}T_5 =& E\left[\sum_{i=1}^{t}\left\|\frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)\odot m_{i-1}\right\|^2\right] \\ \leq& \left(\frac{\beta_1}{1-\beta_1}\right)^2 H^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j^2\right] \end{aligned} \tag{29}$$

Proof.

$$\begin{aligned} \frac{2}{3L}T_5 \leq& E\left[\sum_{i=2}^{t}\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{j=1}^{d}\left(\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j^2 (m_{i-1})_j^2\right)\right] \\ \leq& \left(\frac{\beta_1}{1-\beta_1}\right)^2 H^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j^2\right] \end{aligned} \tag{30}$$

where the fist inequality is due to $\beta_1 \geq \beta_{1,t}$ and equation (15), the second inequality is due to $\|m_i\| < H$.

The proof is complete.

Lemma 7. Suppose the assumptions in Theorem 1 hold. For T_{10} in equation (16), we have

$$\begin{aligned} T_{10} =& \mu E\left[\sum_{i=1}^{t-1}\left\langle \nabla f(z_{i+1}) - \nabla f(z_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i}\right\rangle\right] + \frac{\beta_1\mu}{1-\beta_1}HG \\ \leq& \frac{\beta_1\mu}{1-\beta_1}\left(T_4 + T_5 + T_6 + \frac{L}{2}E\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]\right) + \frac{\beta_1\mu}{1-\beta_1}HG. \end{aligned} \tag{31}$$

Proof. Combining the assumptions A1, A2, the assumptions in Theorem 1, and the definition of T_{10} in equation (17), we obtain

$$\begin{aligned} T_{10} =& \mu E\left[\sum_{i=1}^{t-1}\left\langle \nabla f(z_{i+1})-\nabla f(z_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i g_i/\sqrt{\hat{v}_i}\right\rangle\right]+\frac{\beta_1\mu}{1-\beta_1}HG \\ \leq& \frac{\beta_1\mu L}{1-\beta_1}E\left[\sum_{i=1}^{t}\|d_i\|\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|\right]+\frac{\beta_1\mu}{1-\beta_1}HG \\ \leq& \frac{\beta_1\mu L}{1-\beta_1}E\left[\sum_{i=1}^{t}\frac{1}{2}\left(\|d_i\|^2+\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right)\right]+\frac{\beta_1\mu}{1-\beta_1}HG \\ \leq& \frac{\beta_1\mu}{1-\beta_1}\left(T_4+T_5+T_6+\frac{L}{2}E\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}HG, \end{aligned} \tag{32}$$

where the last inequity is due to equation (21).

The proof is complete.

lemma 8 Suppose the assumptions in Theorem 1 hold. For T_2 in equation (17), we have

$$\begin{aligned} T_2 =& -E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(z_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\ \leq& \left(L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+\frac{1}{2}\right)E\left[\sum_{i=1}^{t}\|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\ &+L^2H^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right] \\ &+2H^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right] \\ &+2H^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\hat{v}_1})_j\right]-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(x_i), \nabla f(x_t)/\sqrt{\hat{v}_i}\rangle\right]. \end{aligned} \tag{33}$$

Proof. Recall from the definition equation (7), we have

$$z_i-x_i=\frac{\beta_{1,i}}{1-\beta_{1,i}}(x_i-x_{i-1})=-\frac{\beta_{1,i}}{1-\beta_{1,i}}\alpha_{i-1}\left((1-\mu)m_{i-1}+\mu g_{t-1}\right)/\sqrt{\hat{v}_{i-1}}. \tag{34}$$

Further we have $z_1=x_1$ by definition of z_1. We have

$$\begin{aligned} T_2 =& -E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(z_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\ =& -E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right]-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(z_i)-\nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right]. \end{aligned} \tag{35}$$

The second term of equation (35) can be bounded as

$$\begin{aligned} &-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(z_i)-\nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\ \leq& E\left[\sum_{i=2}^{t}\frac{1}{2}\|\nabla f(z_i)-\nabla f(x_i)\|^2+\frac{1}{2}\|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\ \leq& \frac{L^2}{2}T_7+\frac{1}{2}E\left[\sum_{i=2}^{t}\|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right], \end{aligned} \tag{36}$$

where the first inequality is because $\langle a,b\rangle \leq \frac{1}{2}\left(\|a\|^2+\|b\|^2\right)$ and the fact that $z_1 = x_1$, the second inequality is because $\|\nabla f(z_i) - \nabla f(x_i)\| \leq L\|z_i - x_i\| = L\left\|\frac{\beta_{1,t}}{1-\beta_{1,t}}\alpha_{i-1}\left((1-\mu)m_{i-1}+\mu g_{t-1}\right)/\sqrt{\hat{v}_{i-1}}\right\|$, and T_7 is defined as

$$T_7 = E\left[\sum_{i=2}^{t}\left\|\frac{\beta_{1,i}}{1-\beta_{1,i}}\alpha_{i-1}\left((1-\mu)m_{i-1}+\mu g_{t-1}\right)/\sqrt{\hat{v}_{i-1}}\right\|^2\right]. \tag{37}$$

By expanding equation (37), we obtain

$$\begin{aligned}
T_7 \leq & \frac{\beta_1^2(1-\mu)^2}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\left\|\alpha_{i-1}m_{i-1}/\sqrt{\hat{v}_{i-1}}\right\|^2\right] + \frac{\beta_1^2\mu^2}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\left\|\alpha_{i-1}g_{i-1}/\sqrt{\hat{v}_{i-1}}\right\|^2\right] \\
& + \frac{2\beta_1^2\mu(1-\mu)}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\left|\left\langle \alpha_{i-1}m_{i-1}/\sqrt{\hat{v}_{i-1}}, \alpha_{i-1}g_{i-1}/\sqrt{\hat{v}_{i-1}}\right\rangle\right|\right] \\
\leq & \underbrace{\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\left\|\alpha_{i-1}m_{i-1}/\sqrt{\hat{v}_{i-1}}\right\|^2\right]}_{T_{7A}} + \underbrace{\frac{\beta_1^2\mu}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\left\|\alpha_{i-1}g_{i-1}/\sqrt{\hat{v}_{i-1}}\right\|^2\right]}_{T_{7B}},
\end{aligned} \tag{38}$$

where the last inequity is due to $2\langle a,b\rangle \leq \|a\|^2+\|b\|^2$.

We next bound the T_{7A} in equation (38), by update rule $m_i = \beta_{1,i}m_{i-1} + (1-\beta_{1,i}g_i)$, we have $m_i = \sum_{k=1}^{i}[(\prod_{l=k+1}^{i}\beta_{1,l})(1-\beta_{1,k})g_k]$. Based on that, we obtain

$$\begin{aligned}
T_{7A} \leq & \frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_{i-1}m_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j^2\right] \\
= & \frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\frac{\alpha_{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})g_k}{\sqrt{\hat{v}_{i-1}}}\right)_j^2\right] \\
\leq & 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}\underbrace{E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\frac{\alpha_{k}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})g_k}{\sqrt{\hat{v}_{k}}}\right)_j^2\right]}_{T_8} \\
& + 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}\underbrace{E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})(g_k)_j\left(\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\hat{v}_k}}\right)_j\right)^2\right]}_{T_9},
\end{aligned} \tag{39}$$

where the first inequality is due to $\beta_{1,t} \leq \beta_1$, the second equality is by substituting expression of m_t, the last inequality is because $(a+b)^2 \leq 2(\|a\|^2+\|b\|^2)$, and we have introduced T_8 and T_9 for ease of notation.

In equation (39), we first bound T_8 as below

$$
\begin{aligned}
T_8 =& E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\sum_{p=1}^{i-1}\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)(1-\beta_{1,k})\left(\frac{\alpha_p g_p}{\sqrt{\hat{v}_p}}\right)_j\left(\prod_{q=p+1}^{i-1}\beta_{1,p}\right)(1-\beta_{1,p})\right]\\
\overset{(i)}{\leq}& E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\sum_{p=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\beta_1^{i-1-p}\right)\frac{1}{2}\left(\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j^2+\left(\frac{\alpha_p g_p}{\sqrt{\hat{v}_p}}\right)_j^2\right)\right]\\
\overset{(ii)}{=}& E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j^2\sum_{p=1}^{i-1}\left(\beta_1^{i-1-p}\right)\right]\\
\overset{(iii)}{\leq}& \frac{1}{1-\beta_1}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j^2\right]\\
\overset{(iv)}{=}& \frac{1}{1-\beta_1}E\left[\sum_{k=1}^{t-1}\sum_{j=1}^{d}\sum_{i=k+1}^{t}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j^2\right]\\
\leq& \left(\frac{1}{1-\beta_1}\right)^2 E\left[\sum_{k=1}^{t-1}\sum_{j=1}^{d}\left(\frac{\alpha_k g_k}{\sqrt{\hat{v}_k}}\right)_j^2\right]=\left(\frac{1}{1-\beta_1}\right)^2 E\left[\sum_{i=1}^{t-1}\|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right],
\end{aligned}
\tag{40}
$$

where (i) is due to $ab < \frac{1}{2}(a^2+b^2)$ and follows from $\beta_{1,t} \leq \beta_1$ and $\beta_{1,t} \in [0,1)$, (ii) is due to symmetry of p and k in the summation, (iii) is because of $\sum_{p=1}^{i-1}\left(\beta_1^{i-1-p}\right) \leq \frac{1}{1-\beta_1}$, (iv) is exchanging order of summation, and the second-last inequality is due to the similar reason as (iii).

For the T_9 in equation (39), we have

$$
\begin{aligned}
T_9 =& E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)(1-\beta_{1,k})(g_k)_j\left(\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\hat{v}_k}}\right)_j\right)^2\right]\\
\leq& H^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)\left|\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\hat{v}_k}}\right|_j\right)^2\right]\\
\leq& H^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\left|\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_k}{\sqrt{\hat{v}_k}}\right|_j\right)^2\right]\\
\leq& H^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\sum_{l=k+1}^{i}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j\right)^2\right],
\end{aligned}
\tag{41}
$$

where the first inequality holds due to $\beta_{1,k} < 1$ and $|(g_k)_j| \leq H$, the second inequality holds due to $\beta_{1,k} \leq \beta_1$, and the last inequality applied the triangle inequality. For RHS of equation (41), using Lemma 9 (that will be proved later) with $a_i = \left|\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right|_j$, we further have

$$
\begin{aligned}
T_9 \leq& H^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\sum_{l=k+1}^{i}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j\right)^2\right]\\
\leq& H^2\left(\frac{1}{1-\beta_1}\right)^2\left(\frac{\beta_1}{1-\beta_1}\right)^2 E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right]
\end{aligned}
\tag{42}
$$

Based on equation (39), equation (40) and equation (42), we can then bound T_{7A} as

$$\begin{aligned} T_{7A} \leq & 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4} E\left[\sum_{i=1}^{t-1} \|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\ & + 2H^2 \frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6} E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}} - \frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right]. \end{aligned} \tag{43}$$

According to equation (38) and equation (43), we obtain

$$\begin{aligned} T_7 \leq & T_{7A} + T_{7B} \\ \leq & \left(\frac{2\beta_1^2(1-\mu)}{(1-\beta_1)^4} + \frac{\beta_1^2\mu}{(1-\beta_1)^2}\right) E\left[\sum_{i=1}^{t-1} \|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\ & + 2H^2 \frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6} E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}} - \frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right]. \end{aligned} \tag{44}$$

Based on equation (36), equation (39), equation (44), we can then bound the second term of equation (35) as

$$\begin{aligned} & - E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(z_i) - \nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\ \leq & \left(L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4} + \frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right) + \frac{1}{2}\right) E\left[\sum_{i=1}^{t} \|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\ & + L^2 H^2 \frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6} E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}} - \frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right]. \end{aligned} \tag{45}$$

Let us turn to the first term in equation (35). Reparameterize g_t as $g_t = \nabla f(x_t) + \delta_t$ with $E[\delta_t] = 0$, we have

$$\begin{aligned} & E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\ = & E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(x_i), (\nabla f(x_i) + \delta_i)/\sqrt{\hat{v}_i}\rangle\right] \\ = & E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(x_i), \nabla f(x_i)/\sqrt{\hat{v}_i}\rangle\right] + E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(x_i), \delta_i/\sqrt{\hat{v}_i}\rangle\right]. \end{aligned} \tag{46}$$

It can be seen that the first term in RHS of equation (46) is the desired descent quantity, the second term is a bias term to be bounded. For the second term in RHS of equation (46), we have

$$\begin{aligned} & E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(x_i), \delta_i/\sqrt{\hat{v}_i}\rangle\right] \\ = & E\left[\sum_{i=2}^{t} \langle \nabla f(x_i), \delta_i \odot (\alpha_i/\sqrt{\hat{v}_i} - \alpha_{i-1}/\sqrt{\hat{v}_{i-1}})\rangle\right] + E\left[\sum_{i=2}^{t} \alpha_{i-1} \langle \nabla f(x_i), \delta_i \odot (1/\sqrt{\hat{v}_{i-1}})\rangle\right] \\ & + E\left[\alpha_1 \langle \nabla f(x_1), \delta_1/\sqrt{\hat{v}_1}\rangle\right] \\ \geq & E\left[\sum_{i=2}^{t} \langle \nabla f(x_i), \delta_i \odot (\alpha_i/\sqrt{\hat{v}_i} - \alpha_{i-1}/\sqrt{\hat{v}_{i-1}})\rangle\right] - 2H^2 E\left[\sum_{j=1}^{d} (\alpha_1/\sqrt{\hat{v}_1})_j\right], \end{aligned} \tag{47}$$

where the last equation is because given x_i, $\hat{v}_{i-1}$, $E\left[\delta_i \odot (1/\sqrt{\hat{v}_{i-1}})|x_i, \hat{v}_{i-1}\right] = 0$ and $\|\delta_i\| \leq 2H$ due to $\|g_i\| \leq H$ and $\|\nabla f(x_i)\| \leq H$ based on Assumptions A2 and A3. Further, we have

$$\begin{aligned}
&E\left[\sum_{i=2}^{t}\langle\nabla f(x_i), \delta_t \odot (\alpha_i/\sqrt{\hat{v}_i} - \alpha_{i-1}/\sqrt{\hat{v}_{i-1}})\rangle\right] \\
=&E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}(\nabla f(x_i))_j(\delta_t)_j(\alpha_i/(\sqrt{\hat{v}_i})_j - \alpha_{i-1}/(\sqrt{\hat{v}_{i-1}})_j)\right] \\
\geq& -E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}|(\nabla f(x_i))_j|\,|(\delta_t)_j|\left|(\alpha_i/(\sqrt{\hat{v}_i})_j - \alpha_{i-1}/(\sqrt{\hat{v}_{i-1}})_j)\right|\right] \\
\geq& -2H^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|(\alpha_i/(\sqrt{\hat{v}_i})_j - \alpha_{i-1}/(\sqrt{\hat{v}_{i-1}})_j)\right|\right]
\end{aligned} \tag{48}$$

Substituting equation (47) and equation (48) into equation (46), we then bound the first term of equation (35) as

$$\begin{aligned}
&-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(x_i), g_i/\sqrt{\hat{v}_i}\rangle\right] \\
\leq& 2H^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|(\alpha_i/(\sqrt{\hat{v}_i})_j - \alpha_{i-1}/(\sqrt{\hat{v}_{i-1}})_j)\right|\right] + 2H^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\hat{v}_1})_j\right] \\
&-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(x_i), \nabla f(x_i)/\sqrt{\hat{v}_i}\rangle\right].
\end{aligned} \tag{49}$$

We finally apply equation (49) and equation (45) to obtain equation (33).

The proof is complete.

Lemma 9. [Chen et al., 2019] For $a_i \geq 0$, $\beta \in [0, 1)$, and $b_i = \sum_{k=1}^{i}\beta^{i-k}\sum_{l=k+1}^{i}a_l$, we have

$$\sum_{i=1}^{t}b_i^2 \leq \left(\frac{1}{1-\beta}\right)^2\left(\frac{\beta}{1-\beta}\right)^2\sum_{i=2}^{t}a_i^2. \tag{50}$$

Proof. The result is proved by following

$$
\begin{aligned}
&\sum_{i=1}^{t} b_i^2 = \sum_{i=1}^{t} \left(\sum_{k=1}^{i} \beta^{i-k} \sum_{l=k+1}^{i} a_l \right)^2 \\
&\overset{(i)}{=} \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \sum_{k=1}^{l-1} \beta^{i-k} a_l \right)^2 = \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \beta^{i-l+1} a_l \sum_{k=1}^{l-1} \beta^{l-1-k} \right)^2 \\
&\overset{(ii)}{\leq} \left(\frac{1}{1-\beta} \right)^2 \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \beta^{i-l+1} a_l \right)^2 \\
&= \left(\frac{1}{1-\beta} \right)^2 \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} a_l \beta^{i-m+1} a_m \right) \\
&\overset{(iii)}{\leq} \left(\frac{1}{1-\beta} \right)^2 \sum_{i=1}^{t} \sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} \beta^{i-m+1} \frac{1}{2} \left(a_l^2 + a_m^2 \right) \\
&\overset{(iv)}{=} \left(\frac{1}{1-\beta} \right)^2 \sum_{i=1}^{t} \sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} \beta^{i-m+1} a_l^2 \overset{(v)}{\leq} \left(\frac{1}{1-\beta} \right)^2 \frac{\beta}{1-\beta} \sum_{l=2}^{t} \sum_{i=l}^{t} \beta^{i-l+1} a_l^2 \\
&\leq \left(\frac{1}{1-\beta} \right)^2 \left(\frac{\beta}{1-\beta} \right)^2 \sum_{l=2}^{t} a_l^2,
\end{aligned}
\tag{51}
$$

where (i) is by changing order of summation, (ii) is due to $\sum_{k=1}^{l-1} \beta^{l-1-k} \leq \frac{1}{1-\beta}$, (iii) is by the fact that $ab \leq \frac{1}{2}(a^2 + b^2)$, (iv) is due to symmetry of a_l and a_m in the summation, (v) is because $\sum_{m=2}^{i} \beta^{i-m+1} \leq \frac{\beta}{1-\beta}$ and the last inequality is for similar reason.

The proof is complete.

2.2 Proof of Theorem 1

Proof. We combine Lemma 2, Lemma 3, Lemma 4, Lemma 5, Lemma 6, Lemma 7, and Lemma 8 to bound the overall expected descent of the objective. We obtain

$$
\begin{aligned}
&E\left[f(z_{t+1})-f(z_1)\right] \leq \sum_{i=1}^{6} T_i + T_{10} \\
\leq &\sum_{i=1}^{6} T_i + \frac{\beta_1\mu}{1-\beta_1}\left(T_4+T_5+T_6+\frac{L}{2}E\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}HG \\
= &T_1+T_2+T_3+\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}(T_4+T_5+T_6)+\frac{\beta_1\mu}{1-\beta_1}\frac{L}{2}E\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]+\frac{\beta_1\mu}{1-\beta_1}HG \\
\leq &H^2\frac{\beta_1}{1-\beta_1}(1-\mu)E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right] \\
&+\left(L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+\frac{1}{2}\right)E\left[\sum_{i=1}^{t}\|\alpha_i g_i/\sqrt{\hat{v}_i}\|^2\right] \\
&+L^2H^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\hat{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\hat{v}_{l-1}}}\right|_j^2\right] \\
&+2H^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j\right|\right] \\
&+2H^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\hat{v}_1})_j\right]-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(x_i),\nabla f(x_t)/\sqrt{\hat{v}_i}\rangle\right] \\
&+\left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)(1-\mu)\left(H^2+G^2\right) \\
&+\frac{3}{2}L\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2G^2 \\
&+\frac{3}{2}L\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{\beta_1}{1-\beta_1}\right)^2H^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\hat{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\hat{v}_{i-1}}}\right)_j^2\right] \\
&+\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)LE\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right] \\
&+\frac{\beta_1\mu}{1-\beta_1}\left(\frac{L}{2}E\left[\sum_{i=1}^{t}\left\|\alpha_i g_i/\sqrt{\hat{v}_i}\right\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}HG.
\end{aligned}
\tag{52}
$$

By rearranging the above inequality and merging similar terms, we further have

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T}\alpha_t\langle\nabla f(x_t),\nabla f(x_t)/\sqrt{\hat{v}_t}\rangle\right] \\
\leq &E\left[C_1\sum_{t=1}^{T}\left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|^2\right]+C_4,
\end{aligned}
\tag{53}
$$

where the coefficients are

$$
\begin{aligned}
C_1 \leq & L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+L\left(\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)+\frac{1}{2}\frac{\beta_1\mu}{1-\beta_1}\right)+\frac{1}{2}\\
C_2 \leq & H^2\left(\frac{\beta_1(1-\mu)}{1-\beta_1}+2\right)\\
C_3 \leq & L^2H^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}+\frac{3}{2}LH^2\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\frac{\beta_1^2}{(1-\beta_1)^2}\\
C_4 \leq & \frac{\beta_1(1-\mu)}{1-\beta_1}\left(H^2+G^2\right)+\frac{3}{2}LG^2\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2\\
& +\frac{\beta_1\mu}{1-\beta_1}HG+2H^2E\left[\|\alpha_1/\sqrt{\hat{v}_1}\|_1\right]+E\left[f(z_1)-f(z^*)\right],
\end{aligned}
\tag{54}
$$

where z^* is an optimal of f, i.e. $z^* \in \arg\min_z f(z)$.

Using the fact that $(\alpha_i/\sqrt{\hat{v}_i})_j \geq \gamma_i, \forall j$ by definition, inequality equation (2) directly follows.

This completes the proof.

3 Proof of Corollary 1

Proof.[3]

We first bound non-constant terms in RHS of equation (53), which is given by

$$
E\left[C_1\sum_{t=1}^{T}\left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|^2\right]+C_4.
\tag{55}
$$

For the term with C_1, assume $\min_{j\in[d]}(\sqrt{\hat{v}_1})_j \geq c > 0$ (this is natural since if it is 0, division by 0 error will happen), we have

$$
\begin{aligned}
& E\left[\sum_{t=1}^{T}\left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\|^2\right]\\
\leq & E\left[\sum_{t=1}^{T}\|\alpha_t g_t/c\|^2\right]=E\left[\sum_{t=1}^{T}\left\|\frac{1}{\sqrt{t}}g_t/c\right\|^2\right]=E\left[\sum_{t=1}^{T}\left(\frac{1}{c\sqrt{t}}\right)^2\|g_t\|^2\right]\\
\leq & H^2/c^2\sum_{t=1}^{T}\frac{1}{t}\leq H^2/c^2(1+\log T)
\end{aligned}
\tag{56}
$$

where the first inequality is due to $(\hat{v}_t)_j \geq (\hat{v}_{t-1})_j$, and the last inequality is due to $\sum_{t=1}^{T}1/t \leq 1+\log T$.

For the term with C_2, we have

$$
\begin{aligned}
& E\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1\right]=E\left[\sum_{j=1}^{d}\sum_{t=2}^{T}\left(\frac{\alpha_{t-1}}{(\sqrt{\hat{v}_{t-1}})_j}-\frac{\alpha_t}{(\sqrt{\hat{v}_t})_j}\right)\right]\\
= & E\left[\sum_{j=1}^{d}\left(\frac{\alpha_1}{(\sqrt{\hat{v}_1})_j}-\frac{\alpha_T}{(\sqrt{\hat{v}_T})_j}\right)\right]\leq E\left[\sum_{j=1}^{d}\frac{\alpha_1}{(\sqrt{\hat{v}_1})_j}\right]\leq d/c
\end{aligned}
\tag{57}
$$

where the first equality is due to$(\hat{v}_t)_j \geq (\hat{v}_{t-1})_j$ and $\alpha_t \leq \alpha_{t-1}$, and the second equality is due to telescope sum.

[3]The proof is almost the same as the proof of Corollary 3.1 for AMSGRAD in Chen et al. [2019]. The difference mainly lies in Theorem 1 and Theorem 3.1 in their paper.

For the term with C_3, we have

$$\begin{aligned}
&E\left[\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|^2\right]\\
\leq&E\left[\frac{1}{c}\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1\right]\\
\leq&d/c^2
\end{aligned} \tag{58}$$

where the first inequality is due to $|(\alpha_t/\sqrt{\hat{v}_t}-\alpha_{t-1}/\sqrt{\hat{v}_{t-1}})_j|\leq 1/c$.

Then we have for ARSG,

$$\begin{aligned}
&E\left[C_1\sum_{t=1}^{T}\left\|\alpha_t g_t/\sqrt{\hat{v}_t}\right\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{v}_{t-1}}}\right\|^2\right]+C_4\\
\leq&C_1H^2/c^2(1+\log T)+C_2d/c+C_3(d/c)^2+C_4
\end{aligned} \tag{59}$$

Now we lower bound the effective stepsizes, since $\hat{v}_t$ is exponential moving average of g_t^2 and $\|g_t\|\leq H$, we have $(\hat{v}_t)_j\leq H^2$, we have

$$\alpha/(\sqrt{\hat{v}_t})_j\geq\frac{1}{H\sqrt{t}} \tag{60}$$

And thus

$$E\left[\sum_{t=1}^{T}\alpha_i\langle\nabla f(x_t),\nabla f(x_t)/\sqrt{\hat{v}_t}\rangle\right]\geq E\left[\sum_{t=1}^{T}\frac{1}{H\sqrt{t}}\|\nabla f(x_t)\|^2\right]\geq\frac{\sqrt{T}}{H}\min_{t\in[T]}E\left[\|\nabla f(x_t)\|^2\right] \tag{61}$$

Then by equation (53), equation (59) and equation (61), we have

$$\frac{1}{H}\sqrt{T}\min_{t\in[T]}E\left[\|\nabla f(x_t)\|^2\right]\leq C_1H^2/c^2(1+\log T)+C_2d/c+C_3d/c^2+C_4 \tag{62}$$

which is equivalent to

$$\begin{aligned}
&\min_{t\in[T]}E\left[\|\nabla f(x_t)\|^2\right]\\
\leq&\frac{H}{\sqrt{T}}\left(C_1H^2/c^2(1+\log T)+C_2d/c+C_3d/c^2+C_4\right)\\
=&\frac{1}{\sqrt{T}}\left(Q_1+Q_2\log T\right)
\end{aligned} \tag{63}$$

One more thing is to verify the assumption $\|\alpha_t m_t/\sqrt{\hat{v}_t}\|\leq G$ in Theorem 1, since $\alpha_{t+1}/(\sqrt{\hat{v}_{t+1}})_j\leq\alpha_t/(\sqrt{\hat{v}_t})_j$ and $\alpha_1/(\sqrt{\hat{v}_1})_j\leq 1/c$ in the algorithm, we have $\|\alpha_t m_t/\sqrt{\hat{v}_t}\|\leq\|m_t\|/c\leq H/c$.

This completes the proof.

References

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1x-x309tm`.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.