# Supplementary Materials

## A Theoretical analysis details

### A.1 Proof of Theorem 1

According to the expression (15), the eigenvalues of the dynamic system (14) without gradient noise can be considered as functions of $\tau = \alpha\lambda$, as

$$\begin{aligned} &r_1(\tau) = \frac{1}{2}\left(\rho(\tau) - \sqrt{\chi(\tau)}\right), \quad r_2(\tau) = \frac{1}{2}\left(\rho(\tau) + \sqrt{\chi(\tau)}\right), \\ &\text{where } \rho(\tau) = 1 + \beta - \tau(1 - \beta(1 - \mu)), \chi(\tau) = \rho(\tau)^2 - 4\beta(1 - \mu\tau). \end{aligned} \tag{A1}$$

Define the gain factor as

$$r_g(\tau) = \max\left(|r_1(\tau)|, |r_2(\tau)|\right). \tag{A2}$$

The convergence rate $r_c$ is the maximum $r_g(\tau)$ for

$$\tau \in [\tau_{\min}, \tau_{\max}], \text{where } \tau_{\min} = \alpha\lambda_{\min}, \tau_{\max} = \alpha\lambda_{\max}. \tag{A3}$$

In order to prove Theorem 1, we firstly approximate $r_g(\tau_{\min})$, and then prove $r_g(\tau_{\min})$ is the maximum or approximate maximum $r_g(\tau)$ for $\tau \in [\tau_{\min}, \tau_{\max}]$.

#### A.1.1 Approximation of the gain factor for the minimum eigenvalue

Now, we approximate the gain factor $r_g(\tau_{\min})$, where $\tau_{\min} = \alpha\lambda_{\min}$.

From the assumption $\kappa = \lambda_{\max}/\lambda_{\min}$, $\alpha = c_\alpha\sqrt{\kappa}/\lambda_{\max}$, and the definition in (A3) we obtain

$$\tau_{\min} = c_\alpha/\sqrt{\kappa}, \tau_{\max} = c_\alpha\sqrt{\kappa}. \tag{A4}$$

Combining (A1), (A4), and the assumption $\kappa \gg 1$ we obtain

$$\rho(\tau_{\min}) = 1 + \beta + O(1/\sqrt{\kappa}) > 0. \tag{A5}$$

From (A1), (A2), and (A5), we obtain

$$r_g(\tau_{\min}) = |r_2(\tau_{\min})|. \tag{A6}$$

From (A1), (A3), and the assumption $\beta = 1 - c_\beta/\sqrt{\kappa}$, $\mu = c_\mu/\sqrt{\kappa}$, through Taylor expansion, we obtain

$$r_2(\tau_{\min}) = 1 - \frac{1}{2}\left(c_\beta - \sqrt{c_\beta(c_\beta - 4c_\alpha)}\right)/\sqrt{\kappa} + O(1/\kappa). \tag{A7}$$

From (A6), (A7), and the assumption $c_\beta > 0, c_\mu > 0$, we obtain

$$r_g(\tau_{\min}) \approx \begin{cases} 1 - \left(c_\beta - \sqrt{c_\beta(c_\beta - 4c_\alpha)}\right)/(2\sqrt{\kappa}), \text{ if } 4c_\alpha < c_\beta \\ 1 - c_\beta/(2\sqrt{\kappa}), \text{ if } 4c_\alpha \geq c_\beta. \end{cases} \tag{A8}$$

Preprint. Under review.

### A.1.2 Proving that the gain factor is maximized or approximately maximized at the minimum eigenvalue

From the definition (A1) and (A2), the gain factor $r_g(\tau)$ is continuous. Before further discussion, we analyze 3 groups of critical points.

The first group of the critical points satisfy $r_g(\tau) = 1$, which can be considered as the boundary between convergence and divergence. Through long and tedious computation, we find out the following results. The solution of $r_1(\tau) = -1$ is $\tau = 2(1+\beta)/(1-\beta+2\mu\beta)$, and $r_1(\tau) = 1$ has no solution. The solution of $r_2(\tau) = 1$ is $\tau = 0$, and $r_2(\tau) = -1$ has no solution. When $\chi(\tau) < 0$, $r_g(\tau) = 1$ has no solution since $0 \leq \beta < 1$ by the assumption. Consequently, we define the convergence bound as

$$\tau_{\text{rb}} = \frac{2(1+\beta)}{1-\beta+2\beta\mu} \approx \frac{4\sqrt{\kappa}}{c_\beta + 2c_\mu}, \tag{A9}$$

where the approximation follows from the assumption. From $r_g(\tau)$ is continuous, $r_g(\tau_{\min}) = 1 - O(1/\sqrt{\kappa}) < 1$ by (A8), and the solutions of $r_g(\tau) = 1$ are $\tau = 0, \tau_{\text{rb}}$, we obtain $r_g(\tau) < 1$ when

$$\tau \in R_{\text{convergence}} = [\tau_{\min}, \tau_{\text{rb}}), \tag{A10}$$

which provides the region of convergence.

The second group of the critical points satisfy $\rho(\tau) = 0$. The solution is unique, as

$$\tau_{\text{sym}} = \frac{1+\beta}{1-\beta+\beta\mu} \approx \frac{2\sqrt{\kappa}}{c_\beta + c_\mu}. \tag{A11}$$

Because of the definition (A1), (A9), (A11) and the assumption $\kappa \ll 1$, $\tau_{\min} \ll \tau_{\text{sym}}$. From $0 < \beta < 1$, $0 \leq \mu < 1$ by the assumption, we obtain

$$\tau_{\text{rb}} - \tau_{\text{sym}} = \frac{1-\beta^2}{(1-\beta+\beta\mu)(1-\beta+2\beta\mu)} > 0. \tag{A12}$$

Consequently, $[\tau_{\min}, \tau_{\text{sym}}] \subset R_{\text{convergence}}$. The assumption $c_\alpha \leq 2/(c_\beta + c_\mu)$ corresponds to $\tau_{\max} \leq \tau_{\text{sym}}$, that ensures convergence. Combining (A1), (A11), and the assumption, we obtain

$$\begin{aligned} r_g(\tau_{\text{sym}}) = |r_1(\tau_{\text{sym}})| = |r_2(\tau_{\text{sym}})| &= \left| \sqrt{\frac{\beta(\beta+\mu-1)}{1-\beta+\beta\mu}} \right| \\ &\approx \begin{cases} \sqrt{\left|\frac{c_\mu - c_\beta}{c_\mu + c_\beta}\right|}, \text{ if } c_\mu > 0 \\ 1 - c_\beta/(2\sqrt{\kappa}), \text{ if } c_\mu = 0. \end{cases} \end{aligned} \tag{A13}$$

From (A8) and (A13), if $c_\mu > 0$, $r_g(\tau_{\text{sym}}) < r_g(\tau_{\min})$; if $c_\mu = 0$, $r_g(\tau_{\text{sym}}) < r_g(\tau_{\min})$ or $r_g(\tau_{\text{sym}}) \approx r_g(\tau_{\min})$.

The third group of the critical points satisfy $\chi(\tau) = 0$, which can be considered as the boundary between real and complex gain factors. The solutions are

$$\tau_{\text{cb1}} = \frac{(1-\beta)\left(1+\beta-\beta\mu-2\sqrt{\beta(1-\mu)}\right)}{(1-\beta+\beta\mu)^2}, \tau_{\text{cb2}} = \frac{(1-\beta)\left(1+\beta-\beta\mu+2\sqrt{\beta(1-\mu)}\right)}{(1-\beta+\beta\mu)^2}. \tag{A14}$$

From the definition (A1) and the assumption, $\chi(\tau)$ satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\chi(\tau) = 2(1-\beta+\beta\mu)^2 > 0. \tag{A15}$$

Consequently, If $\tau \in (\tau_{\text{cb1}}, \tau_{\text{cb2}})$, $r_1(\tau)$ and $r_2(\tau)$ are complex; else, $r_1(\tau)$ and $r_2(\tau)$ are real. When $\tau \in (\tau_{\text{cb1}}, \tau_{\text{cb2}})$, by the definition (A1) we obtain from the definition (A1) and the assumption, $r_g(\tau)$ satisfies

$$r_g(\tau) = |r_1(\tau)| = |r_2(\tau)| = \sqrt{\rho^2(\tau) - \chi(\tau)}/2 = \sqrt{\beta(1-\mu\tau)}, \tag{A16}$$

where $1-\mu\tau > 0$ due to $\chi(\tau) < 0$. Consequently, $r_g(\tau)$ is nonincreasing on $(\tau_{\text{cb1}}, \tau_{\text{cb2}})$, and $r_g(\tau_{\text{cb2}}) \leq r_g(\tau_{\text{cb1}})$. From the definition (A1), (A14), and the assumption, we obtain

$$r_g(\tau_{\text{cb1}}) = \frac{(1-\beta)\sqrt{\beta(1-\mu)} + \beta\mu}{1-\beta+\beta\mu} \approx 1 - c_\beta/(2\sqrt{\kappa}). \tag{A17}$$

Consequently, $r_g(\tau_{\text{cb2}}) \leq r_g(\tau_{\text{cb1}}) < r_g(\tau_{\min})$ or $r_g(\tau_{\text{cb2}}) \leq r_g(\tau_{\text{cb1}}) \approx r_g(\tau_{\min})$.

From the definition (A4), (A11), (A14), and the assumption, we obtain

$$\tau_{\min} \ll \frac{\tau_{\text{cb1}} + \tau_{\text{cb2}}}{2} \approx \frac{2c_\beta\sqrt{\kappa}}{(c_\beta + c_\mu)^2} \leq \tau_{\text{sym}}. \tag{A18}$$

To sum up, there are 5 critical points. $\tau = 0, \tau_{\text{rb}}$ are the boundary of convergence. $\tau = \tau_{\text{sym}}$ corresponds to the upper bound of step size in the assumption. $\tau = \tau_{\text{cb1}}, \tau_{\text{cb2}}$ are the boundary between real and complex gain factors. $\tau_{\text{sym}}, \tau_{\text{cb1}}, \tau_{\text{cb2}}$ satisfy

$$\begin{aligned} &\tau_{\min} \ll \min(\tau_{\text{sym}}, \tau_{\text{cb2}}), \tau_{\text{sym}} > \tau_{\text{cb1}}, \\ &r_g(\tau) < r_g(\tau_{\min}), \text{ or } r_g(\tau) \approx r_g(\tau_{\min}), \text{ if } \tau = \tau_{\text{sym}}, \tau_{\text{cb1}}, \tau_{\text{cb2}}. \end{aligned} \tag{A19}$$

Then, we prove $r_g(\tau_{\min})$ is the maximum or approximately maximum of $r_g(\tau)$ for $\tau \in [\tau_{\min}, \tau_{\max}]$.

According to the definition (A1) and the assumption, the derivatives of $\rho(\tau)$ and $\chi(\tau)$ are

$$\begin{aligned} \rho'(\tau) &= -(1-\beta+\beta\mu) < 0, \\ \chi'(\tau) &= 2(1-\beta+\beta\mu)^2\tau + 2\left(\beta^2(1-\mu) + \beta\mu - 1\right). \end{aligned} \tag{A20}$$

From (A20) and the assumption, we further obtain

$$4\chi(\tau)(\rho'(\tau))^2 - (\chi'(\tau))^2 = -16(1-\beta)^2\beta(1-\mu) < 0. \tag{A21}$$

From (A21), when $\chi(\tau) \geq 0$,

$$|\rho'(\tau)| < \left|\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{\chi(\tau)}\right|. \tag{A22}$$

By the definition (A1), (A14), and the assumption, we obtain

$$\tau_{\text{cb1}} \approx c_\beta/(4\sqrt{\kappa}). \tag{A23}$$

Similar to (A8), by the definition (A1) and the assumption, we obtain

$$\begin{aligned} r_2(\tau) &\approx 1 - \left(c_\beta - \sqrt{c_\beta\left(c_\beta - 4\tau\sqrt{\kappa}\right)}\right) / \left(2\sqrt{\kappa}\right) \\ &= 1 - O\left(1/\sqrt{\kappa}\right) > 0, \text{ if } 0 < \tau \leq \tau_{\text{cb1}}. \end{aligned} \tag{A24}$$

According to the definition (A1), the relations (A15), (A22), (A24), and the assumption, when $0 < \tau \leq \tau_{\text{cb1}}$,

$$r_g'(\tau) = r_2'(\tau) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\rho(\tau) + \sqrt{\chi(\tau)}\right) = \frac{1}{2}\left(\rho'(\tau) - \left|\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{\chi(\tau)}\right|\right) < 0. \tag{A25}$$

Then, if $\tau_{\text{sym}} \leq \tau_{\text{cb2}}$, from the relation (A16), (A25) and $\tau_{\max} \leq \tau_{\text{sym}}$ by the assumption, $r_g(\tau)$ is nonincreasing on $[\tau_{\min}, \tau_{\max}]$. Consequently, $r_g(\tau_{\min})$ is the maximum.

If $\tau_{\text{sym}} > \tau_{\text{cb2}}$, more analysis is required. When $\tau \geq \tau_{\text{cb2}}$, from the definition (A1) and the derivative relation (A20, A22),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}r_1(\tau) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\rho(\tau) - \sqrt{\chi(\tau)}\right) = \frac{1}{2}\left(\rho'(\tau) - \left|\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{\chi(\tau)}\right|\right) < 0, \\ \frac{\mathrm{d}}{\mathrm{d}\tau}r_2(\tau) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\rho(\tau) + \sqrt{\chi(\tau)}\right) = \frac{1}{2}\left(\rho'(\tau) + \left|\frac{\mathrm{d}}{\mathrm{d}\tau}\sqrt{\chi(\tau)}\right|\right) > 0. \end{aligned} \tag{A26}$$

From the definition (A1) and the relation (A26), $r_1(\tau)$ and $r_2(\tau)$ are continuous monotonous functions on $[\tau_{\text{cb2}}, \tau_{\text{sym}}]$. Consequently, when $\tau \in [\tau_{\text{cb2}}, \tau_{\text{sym}}]$

$$\max_\tau r_g(\tau) = \max\left(\max_\tau |r_1(\tau)|, \max_\tau |r_2(\tau)|\right) = \max\left(r_g(\tau_{\text{cb2}}), r_g(\tau_{\text{sym}})\right). \tag{A27}$$

Combining the relations (A16), (A19), (A25), (A27), and $\tau_{\max} \leq \tau_{\text{sym}}$ by the assumption, $r_g(\tau_{\min})$ is the maximum or an approximate maximum on $[\tau_{\min}, \tau_{\max}]$ when $\tau_{\text{sym}} > \tau_{\text{cb2}}$.

The proof is complete.

### A.2 Proof of Theorem 2

#### A.2.1 Main Results

In this subsection, we prove the convergence properties of ARSG in non-convex optimization, which is more applicable for training deep neural networks than the convergence analysis for convex problems.

**Notations:** In addition to the notations in the paper, we use $x \odot y$ as element-wise product. The vector $\alpha_t/\sqrt{\boldsymbol{v}_t}$ in generalized ARSG and $\alpha_t/\sqrt{\hat{\boldsymbol{v}}_t}$ in ARSG are referred to as the effective stepsize.

Firstly, we study the convergence bound of generalized ARSG by extending the analysis of [1]. It is shown that generalized ARSG shares the form of the convergence bound of generalized ADAM [1], and improves the coefficients for typical hyper-parameters settings.

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

**Theorem A1.** Suppose that Assumptions A1-A3 are satisfied, $\beta_1$ 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 $G > 0$, $\left\|\alpha_t m_t/\sqrt{\boldsymbol{v}_t}\right\| \leq G, \left\|\alpha_t \boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\right\| \leq G, \forall\, t$. Then generalized ARSG (presented by equation (3), omitting the projection operation) yields

$$\begin{aligned}
&E\left[\sum_{t=1}^{T} \alpha_t \langle \nabla f(\boldsymbol{x}_t), \nabla f(\boldsymbol{x}_t)/\sqrt{\boldsymbol{v}_t}\rangle\right] \\
\leq &E\left[C_1 \sum_{t=1}^{T} \|\alpha_t \boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\|^2 + C_2 \sum_{t=2}^{T} \left\| \frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}} \right\|_1 + C_3 \sum_{t=2}^{T-1} \left\| \frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}} \right\|^2\right] + C_4,
\end{aligned} \tag{A28}$$

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 $\{\boldsymbol{g}_t\}$.

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

$$\min_{t\in[T]} E\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = O\left(\frac{s_1(T)}{s_2(T)}\right), \tag{A29}$$

where $s_1(T)$ is defined through the upper bound of RHS of equation (A28), namely, $O(s_1(T))$, and $\sum_{t=1}^{T} \gamma_t = \Omega(s_2(T))$.

**Proof.** See A.2.2 and A.2.3.

Theorem A1 inherits the form of Theorem 3.1 in [1], that shows convergence properties of generalized ADAM. However, the coefficients are slightly different.

The coefficients in Theorem A1 are

$$\begin{aligned}
C_1 =& L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4} + \frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right) + L\left(\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right) + \frac{1}{2}\frac{\beta_1\mu}{1-\beta_1}\right) + \frac{1}{2} \\
C_2 =& G_2^2\left(\frac{\beta_1(1-\mu)}{1-\beta_1} + 2\right) \\
C_3 =& L^2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6} + \frac{3}{2}LG_2^2\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\frac{\beta_1^2}{(1-\beta_1)^2} \\
C_4 =& \frac{\beta_1(1-\mu)}{1-\beta_1}\left(G_2^2+G^2\right) + \frac{3}{2}LG^2\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2 \\
&+ \frac{\beta_1\mu}{1-\beta_1}G_2G + 2G_2^2E\left[\|\alpha_1/\sqrt{\boldsymbol{v}_1}\|_1\right] + E\left[f(\boldsymbol{z}_1) - f(\boldsymbol{z}^*)\right],
\end{aligned} \tag{A30}$$

where $\boldsymbol{z}^*$ is an optimal of $f$, i.e. $\boldsymbol{z}^* \in \arg\min_{\boldsymbol{z}} f(\boldsymbol{z})$.

The coefficients in Theorem 3.1 in [1] are [1]

$$
\begin{aligned}
C_1 =& L^2 \frac{\beta_1^2}{(1-\beta_1)^4} + \frac{3}{2}L + \frac{1}{2} \\
C_2 =& G_2^2 \left( \frac{\beta_1}{1-\beta_1} + 2 \right) \\
C_3 =& L^2 G_2^2 \frac{\beta_1^4}{(1-\beta_1)^6} + \frac{3}{2} L G_2^2 \frac{\beta_1^2}{(1-\beta_1)^2} \\
C_4 =& \left( \frac{\beta_1}{1-\beta_1} \right) \left( G_2^2 + G^2 \right) + \frac{3}{2} L G^2 \left( \frac{\beta_1}{1-\beta_1} \right)^2 + 2 G_2^2 E \left[ \| \alpha_1 / \sqrt{\boldsymbol{v}_1} \|_1 \right] + E \left[ f(\boldsymbol{z}_1) - f(\boldsymbol{z}^*) \right],
\end{aligned}
\tag{A31}
$$

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

It should be noted that although the effect of a large $\beta_1$ is negative in these coefficients in the bounds for the worst cases in general non-convex settings, $\beta_1$ close to 1 is required to gain fast convergence in the local quadratic approximation problems (as shown by Theorem 1). Compared with generalized ADAM, when $d \gg 1, 1-\beta_1 \ll 1, L \gg 1, G \gg 1, G_2 \gg 1$ 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.

Then, we study the convergence bound of ARSG and prove Theorem 2 based on theorem A1. By considering the effect of $\epsilon$ [8], we propose the $O(1/\sqrt{T})$ convergences rate for both ARSG and AMSGRAD, which are better than the $O(\log(T)/\sqrt{T})$ convergence rate of AMSGRAD proposed by [1]. The proof are presented in A.2.4.

Finally, we compare the convergence bounds of ARSG (A32) and AMSGRAD (A33) in typical cases where $d \gg 1, 1-\beta_1 \ll 1, L \gg 1, G_2 \gg 1, \epsilon \ll 1$.

According to Theorem 2, under the assumption of the theorem ARSG satisfies

$$
\min_{t\in[T]} E\left[ \|\nabla f(\boldsymbol{x}_t)\|^2 \right] \leq D_1 \frac{\dot{\alpha}}{\sqrt{T}\epsilon} + D_2 \frac{1}{\dot{\alpha}\sqrt{T}} + (D_3 d + D_4) \frac{1}{T\sqrt{\epsilon}} + (D_5 d + D_6) \frac{\dot{\alpha}}{T^{3/2}\epsilon},
\tag{A32}
$$

A.2.4 shows that under the assumption of Theorem 2 AMSGRAD satisfies

$$
\min_{t\in[T]} E\left[ \|\nabla f(\boldsymbol{x}_t)\|^2 \right] \leq D_1 \frac{\dot{\alpha}}{\sqrt{T}\epsilon^2} + D_2 \frac{1}{\dot{\alpha}\sqrt{T}} + D_3 d \frac{1}{T\epsilon} + (D_5 d + D_6) \frac{\dot{\alpha}}{T^{3/2}\epsilon^2},
\tag{A33}
$$

The constants $D_1, D_2, D_3, D_4, D_5, D_6$ in the bounds (A32) and (A33) are defined in A.2.4. They are independent of $T, \epsilon, d$.

In the bound (A32), the term $D_2/(\dot{\alpha}\sqrt{T})$ represents the influence of initial error and momentum, that is independent of $\epsilon$. The other terms are related to the error introduce by the adaptive preconditioner, so they are sensitive to $\epsilon$. The modified definition of $\epsilon$ in ARSG improves the bound of ARSG compared with AMSGRAD, when the same $\epsilon$ is used (the default value is $\epsilon = 10^{-8}$ for both methods). A larger $\epsilon$ (e.g. $\epsilon = 10^{-3}$ in the fine mode of ARSG to improve generalization) further lowers the bound.

Then, we compare the coefficients. ARSG generates lower $D_1, D_2, D_3, D_5$, but has a higher $D_6$ and an additional coefficient $D_4$ (we can denote $D_4 = 0$ for AMSGRAD). However, since the dimension $d \gg 1$, we have $D_4/(T\sqrt{\epsilon}) \ll D_3 d/(T\sqrt{\epsilon})$ and $D_6/(T^{3/2}\epsilon) \ll D_5 d/(T^{3/2}\epsilon)$ in (A32), showing that the increment in $D_6$ and $D_4$ is neglectable. Consequently, in typical cases ARSG improves the coefficients in the bound.

[1] In [1], 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.

### A.2.2 Proof of auxiliary lemmas

**Lemma 1.** Let $\boldsymbol{x}_0 = \boldsymbol{x}_1$ in Algorithm 2, consider the sequence

$$\boldsymbol{z}_t = \boldsymbol{x}_t + \frac{\beta_{1,t}}{1-\beta_{1,t}}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}),\ \forall t \geq 1. \tag{A34}$$

Then the following holds true

$$\begin{aligned}\boldsymbol{z}_{t+1} - \boldsymbol{z}_t =& -\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)\boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t} \\ &- \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot \boldsymbol{m}_{t-1} \\ &+ \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}} - \alpha_t\left(1 + \frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}, \quad \forall t > 1\end{aligned} \tag{A35}$$

and

$$\boldsymbol{z}_2 - \boldsymbol{z}_1 = -\alpha_1\left(\frac{\beta_{1,2}}{1-\beta_{1,2}} - \frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)\boldsymbol{m}_1/\sqrt{\boldsymbol{v}_1} - \alpha_1\left(1 + \frac{\beta_{1,2}\mu}{1-\beta_{1,2}}\right)\boldsymbol{g}_1/\sqrt{\boldsymbol{v}_1}. \tag{A36}$$

**Proof.** By the update rules (7), we have when $t > 1$,

$$\begin{aligned}&\boldsymbol{x}_{t+1} - \boldsymbol{x}_t \\ =& -\alpha_t\left((1-\mu)\boldsymbol{m}_t + \mu\boldsymbol{g}_t\right)/\sqrt{\boldsymbol{v}_t} \\ =& -\alpha_t\left(\beta_{1,t}(1-\mu)\boldsymbol{m}_{t-1} + ((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t\right)/\sqrt{\boldsymbol{v}_t} \\ =& \beta_{1,t}\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}\odot\left(\frac{\sqrt{\boldsymbol{v}_{t-1}}}{\alpha_{t-1}}\odot(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + \mu\boldsymbol{g}_{t-1}\right) - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t} \\ =& \beta_{1,t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + \beta_{1,t}\left(\frac{\alpha_t}{\alpha_{t-1}}\frac{\sqrt{\boldsymbol{v}_{t-1}}}{\sqrt{\boldsymbol{v}_t}} - 1\right)\odot(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) \\ &+ \alpha_t\beta_{1,t}\mu\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t} \\ =& \beta_{1,t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - \beta_{1,t}\left(\frac{\alpha_t}{\alpha_{t-1}}\frac{\sqrt{\boldsymbol{v}_{t-1}}}{\sqrt{\boldsymbol{v}_t}} - 1\right)\odot\alpha_{t-1}\left((1-\mu)\boldsymbol{m}_{t-1} + \mu\boldsymbol{g}_{t-1}\right)/\sqrt{\boldsymbol{v}_{t-1}} \\ &+ \alpha_t\beta_{1,t}\mu\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t} \\ =& \beta_{1,t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - \beta_{1,t}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\left((1-\mu)\boldsymbol{m}_{t-1} + \mu\boldsymbol{g}_{t-1}\right) \\ &+ \alpha_t\beta_{1,t}\mu\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_t} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t} \\ =& \beta_{1,t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - \beta_{1,t}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1} \\ &+ \alpha_{t-1}\beta_{1,t}\mu\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}.\end{aligned} \tag{A37}$$

Since $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = (1-\beta_{1,t})\boldsymbol{x}_{t+1} + \beta_{1,t}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) - (1-\beta_{1,t})\boldsymbol{x}_t$, based on equation (A37) we have

$$\begin{aligned}&(1-\beta_{1,t})\boldsymbol{x}_{t+1} + \beta_{1,t}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) \\ =& (1-\beta_{1,t})\boldsymbol{x}_t + \beta_{1,t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - \beta_{1,t}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1} \\ &+ \alpha_{t-1}\beta_{1,t}\mu\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}} - \alpha_t((1-\beta_{1,t})(1-\mu)+\mu)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}.\end{aligned} \tag{A38}$$

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

$$
\begin{aligned}
&\boldsymbol{x}_{t+1}+\frac{\beta_{1,t}}{1-\beta_{1,t}}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)\\
=&\boldsymbol{x}_t+\frac{\beta_{1,t}}{1-\beta_{1,t}}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})-\frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1}\\
&+\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left(1-\mu+\frac{\mu}{1-\beta_{1,t}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}.
\end{aligned}
\tag{A39}
$$

According to the definition (A34), Then equation (A39) can be written as

$$
\begin{aligned}
&\boldsymbol{z}_{t+1}\\
=&\boldsymbol{z}_t+\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)-\frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1}\\
&+\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left(1-\mu+\frac{\mu}{1-\beta_{1,t}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\\
&=\boldsymbol{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\boldsymbol{g}_t\right)/\sqrt{\boldsymbol{v}_t}\right)\\
&-\frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1}\\
&+\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left(1-\mu+\frac{\mu}{1-\beta_{1,t}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\\
&=\boldsymbol{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{\boldsymbol{v}_t}\\
&-\frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1}\\
&+\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left((1-\mu)+\frac{\mu}{1-\beta_{1,t+1}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\\
&=\boldsymbol{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{\boldsymbol{v}_t}\\
&-\frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot\boldsymbol{m}_{t-1}\\
&+\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left(1+\frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t},\quad\forall t>1.
\end{aligned}
\tag{A40}
$$

For $t=1$, we have $\boldsymbol{z}_1=\boldsymbol{x}_1$ (due to $\boldsymbol{x}_1=\boldsymbol{x}_0$), and

$$\begin{aligned}
&\boldsymbol{z}_2-\boldsymbol{z}_1\\
=&\boldsymbol{x}_2+\frac{\beta_{1,2}}{1-\beta_{1,2}}(\boldsymbol{x}_2-\boldsymbol{x}_1)-\boldsymbol{x}_1\\
=&\boldsymbol{x}_2+\left(\frac{\beta_{1,2}}{1-\beta_{1,2}}-\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(\boldsymbol{x}_2-\boldsymbol{x}_1)+\frac{\beta_{1,1}}{1-\beta_{1,1}}(\boldsymbol{x}_2-\boldsymbol{x}_1)-\boldsymbol{x}_1\\
=&\left(\frac{\beta_{1,2}}{1-\beta_{1,2}}-\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(-\alpha_1((1-\mu)\boldsymbol{m}_1+\mu\boldsymbol{g}_1)/\sqrt{\boldsymbol{v}_1})+\left(\frac{\beta_{1,1}}{1-\beta_{1,1}}+1\right)(\boldsymbol{x}_2-\boldsymbol{x}_1)\\
=&\left(\frac{\beta_{1,2}}{1-\beta_{1,2}}-\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(-\alpha_1((1-\mu)\boldsymbol{m}_1+\mu\boldsymbol{g}_1)/\sqrt{\boldsymbol{v}_1})\\
&+\frac{1}{1-\beta_{1,1}}(-\alpha_1((1-\beta_{1,1})(1-\mu)+\mu)\boldsymbol{g}_1/\sqrt{\hat{v}_1})\\
=&-\left(\frac{\beta_{1,2}}{1-\beta_{1,2}}-\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)(\alpha_1\boldsymbol{m}_1/\sqrt{\boldsymbol{v}_1})-\alpha_1\left((1-\mu)+\frac{1}{1-\beta_{1,2}}\mu\right)\boldsymbol{g}_1/\sqrt{\boldsymbol{v}_1}\\
=&-\left(\frac{\beta_{1,2}}{1-\beta_{1,2}}-\frac{\beta_{1,1}}{1-\beta_{1,1}}\right)(1-\mu)(\alpha_1\boldsymbol{m}_1/\sqrt{\boldsymbol{v}_1})-\alpha_1\left(1+\frac{\beta_{1,2}\mu}{1-\beta_{1,2}}\right)\boldsymbol{g}_1/\sqrt{\boldsymbol{v}_1}.
\end{aligned}\tag{A41}$$

The proof is complete.

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

$$\left(\frac{\alpha_1}{\sqrt{\boldsymbol{v}_1}}-\frac{\alpha_0}{\sqrt{\boldsymbol{v}_0}}\right)\odot\boldsymbol{m}_0=0,\quad \left(\frac{\alpha_1}{\sqrt{\boldsymbol{v}_1}}-\frac{\alpha_0}{\sqrt{\boldsymbol{v}_0}}\right)\odot\boldsymbol{g}_0=0.\tag{A42}$$

**Lemma 2.** Suppose that the conditions in Theorem A1 hold, then

$$E\left[f(\boldsymbol{z}_{t+1})-f(\boldsymbol{z}_1)\right]\leq\sum_{i=1}^{6}T_i+T_{10},\tag{A43}$$

where

$$\begin{aligned}
T_1&=-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i),\frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot\boldsymbol{m}_{i-1}\rangle\right],\\
T_2&=-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{z}_i),\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right],\\
T_3&=-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i),\left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right)\alpha_i\boldsymbol{m}_i/\sqrt{\boldsymbol{v}_i}\rangle\right],\\
T_4&=E\left[\sum_{i=1}^{t}\frac{3}{2}L\left\|\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t\boldsymbol{m}_t/\sqrt{v_t}\right\|^2\right],\\
T_5&=E\left[\sum_{i=1}^{t}\frac{3}{2}L\left\|\frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot\boldsymbol{m}_{i-1}\right\|^2\right],\\
T_6&=\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)LE\left[\sum_{i=1}^{t}\left\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\|^2\right]\\
T_{10}&=\mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(\boldsymbol{z}_{i+1})-\nabla f(\boldsymbol{z}_i),\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\rangle\right]+\frac{\beta_1\mu}{1-\beta_1}G_2G.
\end{aligned}\tag{A44}$$

**Proof.** By the Lipschitz smoothness of $\nabla f$, we obtain

$$f(\boldsymbol{z}_{t+1})\leq f(\boldsymbol{z}_t)+\langle\nabla f(\boldsymbol{z}_t),\boldsymbol{d}_t\rangle+\frac{L}{2}\left\|\boldsymbol{d}_t\right\|^2,\tag{A45}$$

where $\boldsymbol{d}_t = \boldsymbol{z}_{t+1} - \boldsymbol{z}_t$, and Lemma 1 together with equation (A45) yield

$$\begin{aligned}\boldsymbol{d}_t =& -\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}} - \frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t(1-\mu)\boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t} \\ &- \frac{\beta_{1,t}}{1-\beta_{1,t}}(1-\mu)\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right)\odot \boldsymbol{m}_{t-1} \\ &+ \frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}} - \alpha_t\left(1+\frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}, \quad \forall t \geq 1.\end{aligned} \tag{A46}$$

Based on equation (A45) and equation (A46), we then have

$$\begin{aligned}&E[f(\boldsymbol{z}_{t+1}) - f(\boldsymbol{z}_1)] = E\left[\sum_{i=1}^{t} f(\boldsymbol{z}_{i+1}) - f(\boldsymbol{z}_i)\right] \\ \leq& E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i), \boldsymbol{d}_i\rangle + \frac{L}{2}\|\boldsymbol{d}_i\|^2\right] \\ =& \underbrace{-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i), \frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}} - \frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot \boldsymbol{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(\boldsymbol{z}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{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(\boldsymbol{z}_i), \boldsymbol{g}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\rangle\right] \\ &\underbrace{-(1-\mu)E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{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{\boldsymbol{v}_i}\rangle\right]}_{T_3} + E\left[\sum_{i=1}^{t}\frac{L}{2}\|\boldsymbol{d}_i\|^2\right] \\ =& T_1 \underbrace{-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{z}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right]}_{T_2} + T_3 \\ &- \mu E\left[\sum_{i=1}^{t}\left\langle\nabla f(\boldsymbol{z}_i), \left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i} - \frac{\beta_{1,i}}{1-\beta_{1,i}}\alpha_{i-1}\boldsymbol{g}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\right)\right\rangle\right] + E\left[\sum_{i=1}^{t}\frac{L}{2}\|\boldsymbol{d}_i\|^2\right] \\ =& T_1 + T_2 + T_3 + E\left[\sum_{i=1}^{t}\frac{L}{2}\|\boldsymbol{d}_i\|^2\right] \\ &+ \mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(\boldsymbol{z}_{i+1}) - \nabla f(\boldsymbol{z}_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\rangle\right] - \mu E\left[\left\langle\nabla f(\boldsymbol{z}_t), \frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\alpha_t\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\right\rangle\right] \\ \leq& T_1 + T_2 + T_3 + E\left[\sum_{i=1}^{t}\frac{L}{2}\|\boldsymbol{d}_i\|^2\right] + \underbrace{\mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(\boldsymbol{z}_{i+1}) - \nabla f(\boldsymbol{z}_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\rangle\right] + \frac{\beta_1\mu G_2 G}{1-\beta_1}}_{T_{10}} \\ =& T_1 + T_2 + T_3 + T_{10} + E\left[\sum_{i=1}^{t}\frac{L}{2}\|\boldsymbol{d}_i\|^2\right],\end{aligned} \tag{A47}$$

where $\{T_i\}$ have been defined in equation (A44), 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 (A47), we have

$$\begin{aligned}
&\frac{L}{2}E\left[\sum_{i=1}^{t}\|\boldsymbol{d}_i\|^2\right] \leq T_4+T_5+\frac{3}{2}E\left[\sum_{i=1}^{t}\left\|\frac{\alpha_{t-1}\beta_{1,t}\mu}{1-\beta_{1,t}}\boldsymbol{g}_{t-1}/\sqrt{\boldsymbol{v}_{t-1}}-\alpha_t\left(1+\frac{\beta_{1,t+1}\mu}{1-\beta_{1,t+1}}\right)\boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\right\|^2\right] \\
\leq& T_4+T_5+3\left(\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)^2+\left(\frac{\beta_1\mu}{1-\beta_1}\right)^2\right)LE\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
\leq& T_4+T_5+\underbrace{\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)LE\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]}_{T_6}.
\end{aligned} \tag{A48}$$

Substituting the above inequality into equation (A47), we then obtain equation (A43).

The proof is complete.

The next series of lemmas separately bound the terms on RHS of equation (A43) in Lemma 2.

**Lemma 3.** [1] Suppose that the conditions in Theorem A1 hold, $T_1$ in equation (A44) can be bounded as

$$\begin{aligned}
&\frac{1}{1-\mu}T_1=-E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i),\frac{\beta_{1,t}}{1-\beta_{1,t}}\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot\boldsymbol{m}_{i-1}\rangle\right] \\
\leq& G_2^2\frac{\beta_1}{1-\beta_1}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j\right|\right]
\end{aligned} \tag{A49}$$

**Proof.** Since $\|\boldsymbol{g}_t\| \leq G_2$, by the update rule of $\boldsymbol{m}_t$, we have $\|\boldsymbol{m}_t\| \leq G_2$, this can be proved by induction as below.

Recall that $\boldsymbol{m}_t=\beta_{1,t}\boldsymbol{m}_{t-1}+(1-\beta_{1,t})\boldsymbol{g}_t$, suppose $\|\boldsymbol{m}_{t-1}\| \leq G_2$, we have

$$\|\boldsymbol{m}_t\| \leq (\beta_{1,t}+(1-\beta_{1,t}))\max(\|\boldsymbol{g}_t\|,\|\boldsymbol{m}_{t-1}\|)=\max(\|\boldsymbol{g}_t\|,\|\boldsymbol{m}_{t-1}\|) \leq G_2, \tag{A50}$$

then since $\boldsymbol{m}_0=0$, we have $\|m_0\| \leq G_2$ which completes the induction.

Given $\|\boldsymbol{m}_t\| \leq G_2$, we further have

$$\begin{aligned}
\frac{1}{1-\mu}T_1=&-E\left[\sum_{i=2}^{t}\langle\nabla f(\boldsymbol{z}_i),\frac{\beta_{1,t}}{1-\beta_{1,t}}\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot\boldsymbol{m}_{i-1}\rangle\right] \\
\leq& E\left[\sum_{i=1}^{t}\|\nabla f(\boldsymbol{z}_i)\|\,\|\boldsymbol{m}_{i-1}\|\left(\frac{1}{1-\beta_{1,t}}-1\right)\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j\right|\right] \\
\leq& G_2^2\frac{\beta_1}{1-\beta_1}E\left[\sum_{i=1}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j\right|\right]
\end{aligned} \tag{A51}$$

where the first equality holds due to equation (A42), and the last inequality is due to $\beta_1 \geq \beta_{1,i}$.

The proof is complete.

**Lemma 4.** [1] Suppose the conditions in Theorem A1 hold. For $T_3$ in equation (A44), we have

$$\begin{aligned}
&\frac{1}{1-\mu}T_3=-E\left[\sum_{i=1}^{t}\langle\nabla f(\boldsymbol{z}_i),\left(\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right)\alpha_i m_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
\leq& \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\left(G_2^2+G^2\right).
\end{aligned} \tag{A52}$$

**Proof.**

$$\begin{aligned}\frac{1}{1-\mu}T_3 \leq & E\left[\sum_{i=1}^{t}\left|\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right|\frac{1}{2}\left(\|\nabla f(\boldsymbol{z}_i)\|^2+\|\alpha_i m_i/\sqrt{\boldsymbol{v}_i}\|^2\right)\right]\\ \leq & E\left[\sum_{i=1}^{t}\left|\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}-\frac{\beta_{1,i}}{1-\beta_{1,i}}\right|\frac{1}{2}\left(G_2^2+G^2\right)\right]\\ = & \sum_{i=1}^{t}\left(\frac{\beta_{1,i}}{1-\beta_{1,i}}-\frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\right)\frac{1}{2}\left(G_2^2+G^2\right)\\ \leq & \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\left(G_2^2+G^2\right),\end{aligned} \tag{A53}$$

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 $\|\nabla f(\boldsymbol{x}_t)\| \leq G_2$ and $\|\alpha_i m_i/\sqrt{\boldsymbol{v}_i}\| \leq G$ given by the assumptions in Theorem A1, 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.** [1] Suppose the assumptions in Theorem A1 hold. For $T_4$ in equation (A44), we have

$$\begin{aligned}\frac{2}{3L}T_4 = & E\left[\sum_{i=1}^{t}\left\|\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\alpha_t m_t/\sqrt{v_t}\right\|^2\right]\\ \leq & \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2\end{aligned} \tag{A54}$$

**Proof.** The proof is similar to the previous lemma.

$$\begin{aligned}\frac{2}{3L}T_4 = & E\left[\sum_{i=1}^{t}\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)^2\|\alpha_t m_t/\sqrt{\boldsymbol{v}_t}\|^2\right]\\ \leq & E\left[\sum_{i=1}^{t}\left(\frac{\beta_{1,t}}{1-\beta_{1,t}}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2\right]\\ \leq & \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)\sum_{i=1}^{t}\left(\frac{\beta_{1,t}}{1-\beta_{1,t}}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)G^2\\ \leq & \left(\frac{\beta_1}{1-\beta_1}-\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}\right)^2 G^2\end{aligned} \tag{A55}$$

where the first inequality is due to $\|\alpha_t m_t/\sqrt{\boldsymbol{v}_t}\| \leq G$ 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.** [1] Suppose the assumptions in Theorem A1 hold. For $T_5$ in equation (A44), we have

$$\begin{aligned}\frac{2}{3L}T_5 = & E\left[\sum_{i=1}^{t}\left\|\frac{\beta_{1,i}}{1-\beta_{1,i}}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)\odot \boldsymbol{m}_{i-1}\right\|^2\right]\\ \leq & \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_2^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j^2\right]\end{aligned} \tag{A56}$$

**Proof.**

$$\begin{aligned}\frac{2}{3L}T_5 \leq & E\left[\sum_{i=2}^{t}\left(\frac{\beta_1}{1-\beta_1}\right)^2\sum_{j=1}^{d}\left(\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j^2(\boldsymbol{m}_{i-1})_j^2\right)\right]\\ \leq & \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_2^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j^2\right]\end{aligned} \tag{A57}$$

where the fist inequality is due to $\beta_1 \geq \beta_{1,t}$ and equation (A42), the second inequality is due to $\|m_i\| < G_2$.

The proof is complete.

**Lemma 7.** Suppose the assumptions in Theorem A1 hold. For $T_{10}$ in equation (A43), we have

$$\begin{aligned}T_{10} = & \mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(\boldsymbol{z}_{i+1})-\nabla f(\boldsymbol{z}_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\rangle\right]+\frac{\beta_1\mu}{1-\beta_1}G_2G\\ \leq & \frac{\beta_1\mu}{1-\beta_1}\left(T_4+T_5+T_6+\frac{L}{2}E\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}G_2G.\end{aligned} \tag{A58}$$

**Proof.** Combining the assumptions A1, A2, the assumptions in Theorem A1, and the definition of $T_{10}$ in equation (A44), we obtain

$$\begin{aligned}T_{10} = & \mu E\left[\sum_{i=1}^{t-1}\left\langle\nabla f(\boldsymbol{z}_{i+1})-\nabla f(\boldsymbol{z}_i), \frac{\beta_{1,i+1}}{1-\beta_{1,i+1}}\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\right\rangle\right]+\frac{\beta_1\mu}{1-\beta_1}G_2G\\ \leq & \frac{\beta_1\mu L}{1-\beta_1}E\left[\sum_{i=1}^{t}\|\boldsymbol{d}_i\|\,\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|\right]+\frac{\beta_1\mu}{1-\beta_1}G_2G\\ \leq & \frac{\beta_1\mu L}{1-\beta_1}E\left[\sum_{i=1}^{t}\frac{1}{2}\left(\|\boldsymbol{d}_i\|^2+\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right)\right]+\frac{\beta_1\mu}{1-\beta_1}G_2G\\ \leq & \frac{\beta_1\mu}{1-\beta_1}\left(T_4+T_5+T_6+\frac{L}{2}E\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}G_2G,\end{aligned} \tag{A59}$$

where the last inequity is due to equation (A48).

The proof is complete.

**lemma 8** Suppose the assumptions in Theorem A1 hold. For $T_2$ in equation (A44), we have

$$\begin{aligned}T_2 = & -E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{z}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right]\\ \leq & \left(L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+\frac{1}{2}\right)E\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]\\ & +L^2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right]\\ & +2G_2^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j\right|\right]\\ & +2G_2^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\boldsymbol{v}_1})_j\right]-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \nabla f(\boldsymbol{x}_t)/\sqrt{\boldsymbol{v}_i}\rangle\right].\end{aligned} \tag{A60}$$

**Proof.** Recall from the definition equation (A34), we have

$$\boldsymbol{z}_i-\boldsymbol{x}_i = \frac{\beta_{1,i}}{1-\beta_{1,i}}(\boldsymbol{x}_i-\boldsymbol{x}_{i-1}) = -\frac{\beta_{1,i}}{1-\beta_{1,i}}\alpha_{i-1}\left((1-\mu)\boldsymbol{m}_{i-1}+\mu\boldsymbol{g}_{t-1}\right)/\sqrt{\boldsymbol{v}_{i-1}}. \tag{A61}$$

Further we have $\boldsymbol{z}_1 = \boldsymbol{x}_1$ by definition of $\boldsymbol{z}_1$. We have

$$\begin{aligned} T_2 =& - E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(\boldsymbol{z}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\ =& - E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(\boldsymbol{x}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] - E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(\boldsymbol{z}_i) - \nabla f(\boldsymbol{x}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right]. \end{aligned} \tag{A62}$$

The second term of equation (A62) can be bounded as

$$\begin{aligned} & - E\left[\sum_{i=1}^{t} \alpha_i \langle \nabla f(\boldsymbol{z}_i) - \nabla f(\boldsymbol{x}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\ \leq & E\left[\sum_{i=2}^{t} \frac{1}{2}\|\nabla f(\boldsymbol{z}_i) - \nabla f(\boldsymbol{x}_i)\|^2 + \frac{1}{2}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\ \leq & \frac{L^2}{2} T_7 + \frac{1}{2} E\left[\sum_{i=2}^{t} \|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right], \end{aligned} \tag{A63}$$

where the first inequality is because $\langle a, b\rangle \leq \frac{1}{2}\left(\|a\|^2 + \|b\|^2\right)$ and the fact that $\boldsymbol{z}_1 = \boldsymbol{x}_1$, the second inequality is because $\|\nabla f(\boldsymbol{z}_i) - \nabla f(\boldsymbol{x}_i)\| \leq L\|\boldsymbol{z}_i - \boldsymbol{x}_i\| = L\left\|\frac{\beta_{1,t}}{1-\beta_{1,t}}\alpha_{i-1}\left((1-\mu)\boldsymbol{m}_{i-1} + \mu \boldsymbol{g}_{t-1}\right)/\sqrt{\boldsymbol{v}_{i-1}}\right\|$, and $T_7$ is defined as

$$T_7 = E\left[\sum_{i=2}^{t} \left\| \frac{\beta_{1,i}}{1-\beta_{1,i}} \alpha_{i-1}\left((1-\mu)\boldsymbol{m}_{i-1} + \mu \boldsymbol{g}_{t-1}\right)/\sqrt{\boldsymbol{v}_{i-1}}\right\|^2\right]. \tag{A64}$$

By expanding equation (A64), we obtain

$$\begin{aligned} T_7 \leq & \frac{\beta_1^2(1-\mu)^2}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t} \|\alpha_{i-1}\boldsymbol{m}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\|^2\right] + \frac{\beta_1^2\mu^2}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t} \|\alpha_{i-1}\boldsymbol{g}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\|^2\right] \\ & + \frac{2\beta_1^2\mu(1-\mu)}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t} |\langle \alpha_{i-1}\boldsymbol{m}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}, \alpha_{i-1}\boldsymbol{g}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\rangle|\right] \\ \leq & \underbrace{\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t} \|\alpha_{i-1}\boldsymbol{m}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\|^2\right]}_{T_{7A}} + \underbrace{\frac{\beta_1^2\mu}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t} \|\alpha_{i-1}\boldsymbol{g}_{i-1}/\sqrt{\boldsymbol{v}_{i-1}}\|^2\right]}_{T_{7B}}, \end{aligned} \tag{A65}$$

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 (A65), by update rule $m_i = \beta_{1,i}\boldsymbol{m}_{i-1} + (1-\beta_{1,i}\boldsymbol{g}_i)$, we have $m_i = \sum_{k=1}^{i}[(\prod_{l=k+1}^{i}\beta_{1,l})(1-\beta_{1,k})\boldsymbol{g}_k]$. Based on that, we obtain

$$
\begin{aligned}
T_{7A} \leq & \frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_{i-1}\boldsymbol{m}_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j^2\right] \\
= & \frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2} E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\frac{\alpha_{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})\boldsymbol{g}_k}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j^2\right] \\
\leq & 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}\underbrace{E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\frac{\alpha_{k}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})\boldsymbol{g}_k}{\sqrt{\boldsymbol{v}_{k}}}\right)_j^2\right]}_{T_8} \\
& + 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^2}\underbrace{E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,l}\right)(1-\beta_{1,k})(\boldsymbol{g}_k)_j\left(\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\boldsymbol{v}_k}}\right)_j\right)^2\right]}_{T_9},
\end{aligned}
\tag{A66}
$$

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 (A66), we first bound $T_8$ as below

$$
\begin{aligned}
T_8 = & E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\sum_{p=1}^{i-1}\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)(1-\beta_{1,k})\left(\frac{\alpha_p g_p}{\sqrt{\boldsymbol{v}_p}}\right)_j\left(\prod_{q=p+1}^{i-1}\beta_{1,p}\right)(1-\beta_{1,p})\right] \\
\overset{(i)}{\leq} & E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\sum_{p=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\beta_1^{i-1-p}\right)\frac{1}{2}\left(\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j^2+\left(\frac{\alpha_p g_p}{\sqrt{\boldsymbol{v}_p}}\right)_j^2\right)\right] \\
\overset{(ii)}{=} & E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j^2\sum_{p=1}^{i-1}\left(\beta_1^{i-1-p}\right)\right] \\
\overset{(iii)}{\leq} & \frac{1}{1-\beta_1}E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\sum_{k=1}^{i-1}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j^2\right] \\
\overset{(iv)}{=} & \frac{1}{1-\beta_1}E\left[\sum_{k=1}^{t-1}\sum_{j=1}^{d}\sum_{i=k+1}^{t}\left(\beta_1^{i-1-k}\right)\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j^2\right] \\
\leq & \left(\frac{1}{1-\beta_1}\right)^2 E\left[\sum_{k=1}^{t-1}\sum_{j=1}^{d}\left(\frac{\alpha_k g_k}{\sqrt{\boldsymbol{v}_k}}\right)_j^2\right] = \left(\frac{1}{1-\beta_1}\right)^2 E\left[\sum_{i=1}^{t-1}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right],
\end{aligned}
\tag{A67}
$$

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 (A66), we have

$$\begin{aligned}
T_9 =& E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)(1-\beta_{1,k})(\boldsymbol{g}_k)_j\left(\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\boldsymbol{v}_k}}\right)_j\right)^2\right] \\
\leq& G_2^2 E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\sum_{k=1}^{i-1}\left(\prod_{l=k+1}^{i-1}\beta_{1,k}\right)\left|\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}-\frac{\alpha_k}{\sqrt{\boldsymbol{v}_k}}\right|_j\right)^2\right] \\
\leq& G_2^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\left|\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_k}{\sqrt{\boldsymbol{v}_k}}\right|_j\right)^2\right] \\
\leq& G_2^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\sum_{l=k+1}^{i}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j\right)^2\right],
\end{aligned} \tag{A68}$$

where the first inequality holds due to $\beta_{1,k}<1$ and $|(g_k)_j|\leq G_2$, the second inequality holds due to $\beta_{1,k}\leq\beta_1$, and the last inequality applied the triangle inequality. For RHS of equation (A68), using Lemma 9 (that will be proved later) with $a_i=\left|\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right|_j$, we further have

$$\begin{aligned}
T_9 \leq& G_2^2 E\left[\sum_{i=1}^{t-1}\sum_{j=1}^{d}\left(\sum_{k=1}^{i}\beta_1^{i-k}\sum_{l=k+1}^{i}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j\right)^2\right] \\
\leq& G_2^2\left(\frac{1}{1-\beta_1}\right)^2\left(\frac{\beta_1}{1-\beta_1}\right)^2 E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right]
\end{aligned} \tag{A69}$$

Based on equation (A66), equation (A67) and equation (A69), we can then bound $T_{7A}$ as

$$\begin{aligned}
T_{7A} \leq& 2\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}E\left[\sum_{i=1}^{t-1}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
&+2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right].
\end{aligned} \tag{A70}$$

According to equation (A65) and equation (A70), we obtain

$$\begin{aligned}
T_7 \leq& T_{7A}+T_{7B} \\
\leq& \left(\frac{2\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{(1-\beta_1)^2}\right)E\left[\sum_{i=1}^{t-1}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
&+2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right].
\end{aligned} \tag{A71}$$

Based on equation (A63), equation (A66), equation (A71), we can then bound the second term of equation (A62) as

$$\begin{aligned}
&-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{z}_i)-\nabla f(\boldsymbol{x}_i),\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
\leq& \left(L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+\frac{1}{2}\right)E\left[\sum_{i=1}^{t}\|\alpha_i\boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
&+L^2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}E\left[\sum_{j=1}^{d}\sum_{i=2}^{t-1}\left|\frac{\alpha_l}{\sqrt{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right].
\end{aligned} \tag{A72}$$

Let us turn to the first term in equation (A62). By assumption A3 as $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_t) + \zeta_t$ with $E[\zeta_t] = 0$, we have

$$\begin{aligned}
&E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
=&E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), (\nabla f(\boldsymbol{x}_i)+\zeta_i)/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
=&E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \nabla f(\boldsymbol{x}_i)/\sqrt{\boldsymbol{v}_i}\rangle\right] + E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \zeta_i/\sqrt{\boldsymbol{v}_i}\rangle\right].
\end{aligned} \tag{A73}$$

It can be seen that the first term in RHS of equation (A73) is the desired descent quantity, the second term is a bias term to be bounded. For the second term in RHS of equation (A73), we have

$$\begin{aligned}
&E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \zeta_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
=&E\left[\sum_{i=2}^{t}\langle\nabla f(\boldsymbol{x}_i), \zeta_i\odot(\alpha_i/\sqrt{\boldsymbol{v}_i}-\alpha_{i-1}/\sqrt{\boldsymbol{v}_{i-1}})\rangle\right] + E\left[\sum_{i=2}^{t}\alpha_{i-1}\langle\nabla f(\boldsymbol{x}_i), \zeta_i\odot(1/\sqrt{\boldsymbol{v}_{i-1}})\rangle\right] \\
&+ E\left[\alpha_1\langle\nabla f(\boldsymbol{x}_1), \zeta_1/\sqrt{\boldsymbol{v}_1}\rangle\right] \\
\geq&E\left[\sum_{i=2}^{t}\langle\nabla f(\boldsymbol{x}_i), \zeta_i\odot(\alpha_i/\sqrt{\boldsymbol{v}_i}-\alpha_{i-1}/\sqrt{\boldsymbol{v}_{i-1}})\rangle\right] - 2G_2^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\boldsymbol{v}_1})_j\right],
\end{aligned} \tag{A74}$$

where the last equation is because given $\boldsymbol{x}_i, \boldsymbol{v}_{i-1}$, $E\left[\zeta_i\odot(1/\sqrt{\boldsymbol{v}_{i-1}})|\boldsymbol{x}_i, \boldsymbol{v}_{i-1}\right] = 0$ and $\|\zeta_i\| \leq 2G_2$ due to $\|\boldsymbol{g}_i\| \leq G_2$ and $\|\nabla f(\boldsymbol{x}_i)\| \leq G_2$ based on Assumptions A2 and A3. Further, we have

$$\begin{aligned}
&E\left[\sum_{i=2}^{t}\langle\nabla f(\boldsymbol{x}_i), \zeta_t\odot(\alpha_i/\sqrt{\boldsymbol{v}_i}-\alpha_{i-1}/\sqrt{\boldsymbol{v}_{i-1}})\rangle\right] \\
=&E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}(\nabla f(\boldsymbol{x}_i))_j(\zeta_t)_j(\alpha_i/(\sqrt{\boldsymbol{v}_i})_j-\alpha_{i-1}/(\sqrt{\boldsymbol{v}_{i-1}})_j)\right] \\
\geq&-E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}|(\nabla f(\boldsymbol{x}_i))_j|\,|(\zeta_t)_j|\,|(\alpha_i/(\sqrt{\boldsymbol{v}_i})_j-\alpha_{i-1}/(\sqrt{\boldsymbol{v}_{i-1}})_j)|\right] \\
\geq&-2G_2^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}|(\alpha_i/(\sqrt{\boldsymbol{v}_i})_j-\alpha_{i-1}/(\sqrt{\boldsymbol{v}_{i-1}})_j)|\right]
\end{aligned} \tag{A75}$$

Substituting equation (A74) and equation (A75) into equation (A73), we then bound the first term of equation (A62) as

$$\begin{aligned}
&-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\rangle\right] \\
\leq&2G_2^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}|(\alpha_i/(\sqrt{\boldsymbol{v}_i})_j-\alpha_{i-1}/(\sqrt{\boldsymbol{v}_{i-1}})_j)|\right] + 2G_2^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\boldsymbol{v}_1})_j\right] \\
&-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i), \nabla f(\boldsymbol{x}_i)/\sqrt{\boldsymbol{v}_i}\rangle\right].
\end{aligned} \tag{A76}$$

We finally apply equation (A76) and equation (A72) to obtain equation (A60).

The proof is complete.

**Lemma 9.** [1] 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

$$\sum_{i=1}^{t} b_i^2 \leq \left(\frac{1}{1-\beta}\right)^2 \left(\frac{\beta}{1-\beta}\right)^2 \sum_{i=2}^{t} a_i^2. \tag{A77}$$

**Proof.** The result is proved by following

$$\begin{aligned}
&\sum_{i=1}^{t} b_i^2 = \sum_{i=1}^{t} \left(\sum_{k=1}^{i} \beta^{i-k} \sum_{l=k+1}^{i} a_l\right)^2 \\
&\overset{(i)}{=} \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \sum_{k=1}^{l-1} \beta^{i-k} a_l\right)^2 = \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \beta^{i-l+1} a_l \sum_{k=1}^{l-1} \beta^{l-1-k}\right)^2 \\
&\overset{(ii)}{\leq} \left(\frac{1}{1-\beta}\right)^2 \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \beta^{i-l+1} a_l\right)^2 \\
&= \left(\frac{1}{1-\beta}\right)^2 \sum_{i=1}^{t} \left(\sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} a_l \beta^{i-m+1} a_m\right) \\
&\overset{(iii)}{\leq} \left(\frac{1}{1-\beta}\right)^2 \sum_{i=1}^{t} \sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} \beta^{i-m+1} \frac{1}{2} \left(a_l^2 + a_m^2\right) \\
&\overset{(iv)}{=} \left(\frac{1}{1-\beta}\right)^2 \sum_{i=1}^{t} \sum_{l=2}^{i} \sum_{m=2}^{i} \beta^{i-l+1} \beta^{i-m+1} a_l^2 \overset{(v)}{\leq} \left(\frac{1}{1-\beta}\right)^2 \frac{\beta}{1-\beta} \sum_{l=2}^{t} \sum_{i=l}^{t} \beta^{i-l+1} a_l^2 \\
&\leq \left(\frac{1}{1-\beta}\right)^2 \left(\frac{\beta}{1-\beta}\right)^2 \sum_{l=2}^{t} a_l^2,
\end{aligned} \tag{A78}$$

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.

### A.2.3 Proof of Theorem A1

**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(\boldsymbol{z}_{t+1})-f(\boldsymbol{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}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]\right) + \frac{\beta_1\mu}{1-\beta_1}G_2 G \\
= &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}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]+\frac{\beta_1\mu}{1-\beta_1}G_2 G \\
\leq &G_2^2\frac{\beta_1}{1-\beta_1}(1-\mu)E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{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 \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
&+L^2G_2^2\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{\boldsymbol{v}_l}}-\frac{\alpha_{l-1}}{\sqrt{\boldsymbol{v}_{l-1}}}\right|_j^2\right] \\
&+2G_2^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left|\left(\frac{\alpha_i}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{v}_{i-1}}}\right)_j\right|\right] \\
&+2G_2^2E\left[\sum_{j=1}^{d}(\alpha_1/\sqrt{\boldsymbol{v}_1})_j\right]-E\left[\sum_{i=1}^{t}\alpha_i\langle\nabla f(\boldsymbol{x}_i),\nabla f(\boldsymbol{x}_t)/\sqrt{\boldsymbol{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(G_2^2+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)^2G_2^2E\left[\sum_{i=2}^{t}\sum_{j=1}^{d}\left(\frac{\alpha_t}{\sqrt{\boldsymbol{v}_i}}-\frac{\alpha_{i-1}}{\sqrt{\boldsymbol{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}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right] \\
&+\frac{\beta_1\mu}{1-\beta_1}\left(\frac{L}{2}E\left[\sum_{i=1}^{t}\|\alpha_i \boldsymbol{g}_i/\sqrt{\boldsymbol{v}_i}\|^2\right]\right)+\frac{\beta_1\mu}{1-\beta_1}G_2 G.
\end{aligned}
\tag{A79}
$$

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

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T}\alpha_t\langle\nabla f(\boldsymbol{x}_t),\nabla f(\boldsymbol{x}_t)/\sqrt{\boldsymbol{v}_t}\rangle\right] \\
\leq &E\left[C_1\sum_{t=1}^{T}\|\alpha_t \boldsymbol{g}_t/\sqrt{\boldsymbol{v}_t}\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\boldsymbol{v}_t}}-\frac{\alpha_{t-1}}{\sqrt{\boldsymbol{v}_{t-1}}}\right\|^2\right]+C_4,
\end{aligned}
\tag{A80}
$$

where the coefficients are

$$
\begin{aligned}
C_1 =& L^2\left(\frac{\beta_1^2(1-\mu)}{(1-\beta_1)^4}+\frac{\beta_1^2\mu}{2(1-\beta_1)^2}\right)+L\left(\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\left(\frac{9\beta_1^2\mu^2}{(1-\beta_1)^2}+6\right)+\frac{1}{2}\frac{\beta_1\mu}{1-\beta_1}\right)+\frac{1}{2}\\
C_2 =& G_2^2\left(\frac{\beta_1(1-\mu)}{1-\beta_1}+2\right)\\
C_3 =& L^2G_2^2\frac{\beta_1^4(1-\mu)}{(1-\beta_1)^6}+\frac{3}{2}LG_2^2\frac{1-\beta_1+\beta_1\mu}{1-\beta_1}\frac{\beta_1^2}{(1-\beta_1)^2}\\
C_4 =& \frac{\beta_1(1-\mu)}{1-\beta_1}\left(G_2^2+G^2\right)+\frac{3}{2}LG^2\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2\\
&+\frac{\beta_1\mu}{1-\beta_1}G_2G+2G_2^2E\left[\|\alpha_1/\sqrt{\boldsymbol{v}_1}\|_1\right]+E\left[f(\boldsymbol{z}_1)-f(\boldsymbol{z}^*)\right],
\end{aligned}
\tag{A81}
$$

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

Using the fact that $(\alpha_i/\sqrt{\boldsymbol{v}_i})_j\geq\gamma_i,\forall j$ by definition, inequality equation (A29) directly follows.

This completes the proof.

### A.2.4 Proof of Theorem 2

**Proof.**

First, in order to apply Theorem A1 in ARSG, we prove for some constant $G>0$, $\left\|\alpha_t\boldsymbol{m}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|\leq G, \left\|\alpha_t\boldsymbol{g}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|\leq G,\forall\, t$.

By $\hat{\boldsymbol{v}}_t\geq\epsilon$ from definition and $\alpha_t=\dot{\alpha}/\sqrt{T}$ in the assumptions of Theorem 2, we can obtain

$$
\begin{aligned}
&\left\|\alpha_t\boldsymbol{g}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|\leq\left\|\alpha_t\boldsymbol{g}_t/\sqrt{\epsilon}\right\|=\left\|\dot{\alpha}\boldsymbol{g}_t/\sqrt{T\epsilon}\right\|\leq G,\ \text{where } G=\dot{\alpha}G_2/\sqrt{T\epsilon},\\
&\left\|\alpha_t\boldsymbol{m}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|=\left\|\alpha_t\sum_{u=0}^{t}\left((1-\beta_1)\beta_1^{t-u}\boldsymbol{g}_t\right)/\sqrt{\hat{\boldsymbol{v}}_t}\right\|\leq\dot{\alpha}\sum_{u=0}^{t}\left\|(1-\beta_1)\beta_1^{t-u}\boldsymbol{g}_t\right\|/\sqrt{T\epsilon}\leq G.
\end{aligned}
\tag{A82}
$$

Next we bound non-constant terms in RHS of equation (A80), which is given by

$$
E\left[C_1\sum_{t=1}^{T}\left\|\alpha_t\boldsymbol{g}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|^2\right]+C_4,
\tag{A83}
$$

where $\boldsymbol{v}_t$ in generalized ARSG is substituted by $\hat{\boldsymbol{v}}_t$ in ARSG by definition. It should be noted that by the assumption of Theorem 2, $C_4$ is related to $T$ and $\epsilon$ because $G=\dot{\alpha}G_2/\sqrt{T\epsilon}$ is included in $C_4$. Besides, $C_1, C_2, C_3$ are independent of $T$ and $\epsilon$.

For the term with $C_1$ we have

$$
E\left[\sum_{t=1}^{T}\left\|\alpha_t\boldsymbol{g}_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|^2\right]\leq E\left[\sum_{t=1}^{T}\left\|\dot{\alpha}\boldsymbol{g}_t/\sqrt{T\epsilon}\right\|^2\right]\leq\dot{\alpha}^2G_2^2/\epsilon.
\tag{A84}
$$

For the term with $C_2$, by $(\hat{\boldsymbol{v}}_t)_j\geq(\hat{\boldsymbol{v}}_{t-1})_j$ and $\alpha_t=\dot{\alpha}/\sqrt{T}$ we have

$$
\begin{aligned}
E\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|_1\right]&=E\left[\sum_{j=1}^{d}\sum_{t=2}^{T}\left(\frac{\alpha_{t-1}}{(\sqrt{\hat{\boldsymbol{v}}_{t-1}})_j}-\frac{\alpha_t}{(\sqrt{\hat{\boldsymbol{v}}_t})_j}\right)\right]\\
=E\left[\sum_{j=1}^{d}\left(\frac{\alpha_1}{(\sqrt{\hat{\boldsymbol{v}}_1})_j}-\frac{\alpha_T}{(\sqrt{\hat{\boldsymbol{v}}_T})_j}\right)\right]&\leq E\left[\sum_{j=1}^{d}\frac{\alpha_1}{(\sqrt{\hat{\boldsymbol{v}}_1})_j}\right]\leq\dot{\alpha}d/\sqrt{T\epsilon}.
\end{aligned}
\tag{A85}
$$

For the term with $C_3$, we have

$$E\left[\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|^2\right]\leq E\left[\frac{\dot{\alpha}}{\sqrt{T\epsilon}}\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|_1\right]\leq \dot{\alpha}^2 d/(T\epsilon) \tag{A86}$$

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

By the definition of $G$ (A82) the coefficient $C_4$ can be rearranged as

$$\begin{aligned}
C_4 &= \frac{\beta_1(1-\mu)}{1-\beta_1}\left(G_2^2+G^2\right)+\frac{3}{2}LG^2\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2\\
&+\frac{\beta_1\mu}{1-\beta_1}G_2G+2G_2^2E\left[\|\alpha_1/\sqrt{\hat{\boldsymbol{v}}_1}\|_1\right]+E\left[f(\boldsymbol{z}_1)-f(\boldsymbol{z}^*)\right]\\
&\leq \underbrace{\left(\frac{\beta_1\mu}{1-\beta_1}+2d\right)G_2^2}_{C_{4a}}\frac{\dot{\alpha}}{\sqrt{T\epsilon}}+\underbrace{\left(\frac{\beta_1(1-\mu)}{1-\beta_1}+\frac{3}{2}L\left(1+\frac{\beta_1\mu}{1-\beta_1}\right)\left(\frac{\beta_1}{1-\beta_1}\right)^2\right)G_2^2}_{C_{4b}}\frac{\dot{\alpha}^2}{T\epsilon}\\
&+\underbrace{\frac{\beta_1(1-\mu)}{1-\beta_1}G_2^2+E\left[f(\boldsymbol{z}_1)-f(\boldsymbol{z}^*)\right]}_{C_{4c}}
\end{aligned} \tag{A87}$$

Then we have for ARSG,

$$\begin{aligned}
&E\left[C_1\sum_{t=1}^{T}\left\|\alpha_t g_t/\sqrt{\hat{\boldsymbol{v}}_t}\right\|^2+C_2\sum_{t=2}^{T}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|_1+C_3\sum_{t=2}^{T-1}\left\|\frac{\alpha_t}{\sqrt{\hat{\boldsymbol{v}}_t}}-\frac{\alpha_{t-1}}{\sqrt{\hat{\boldsymbol{v}}_{t-1}}}\right\|^2\right]+C_4\\
\leq& C_1G_2^2\dot{\alpha}^2/\epsilon+(C_2d+C_{4a})\dot{\alpha}/\sqrt{T\epsilon}+(C_3d+C_{4b})\dot{\alpha}^2/(T\epsilon)+C_{4c}
\end{aligned} \tag{A88}$$

Now we lower bound the effective stepsizes, since $\hat{\boldsymbol{v}}_t$ is exponential moving average of $\boldsymbol{g}_t^2$ and $\|\boldsymbol{g}_t\|\leq G_2$, we have $(\hat{\boldsymbol{v}}_t)_j\leq G_2^2$, we have

$$\alpha_t/(\sqrt{\hat{\boldsymbol{v}}_t})_j\geq\frac{\dot{\alpha}}{G_2\sqrt{T}} \tag{A89}$$

And thus

$$E\left[\sum_{t=1}^{T}\alpha_i\langle\nabla f(\boldsymbol{x}_t),\nabla f(\boldsymbol{x}_t)/\sqrt{\hat{\boldsymbol{v}}_t}\rangle\right]\geq E\left[\sum_{t=1}^{T}\frac{\dot{\alpha}}{G_2\sqrt{T}}\|\nabla f(\boldsymbol{x}_t)\|^2\right]\geq\frac{\sqrt{T}\dot{\alpha}}{G_2}\min_{t\in[T]}E\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \tag{A90}$$

Then by equation (A80), equation (A88) and equation (A90), we have

$$\frac{\dot{\alpha}\sqrt{T}}{G_2}\min_{t\in[T]}E\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\leq C_1G_2^2\dot{\alpha}^2/\epsilon+(C_2d+C_{4a})\dot{\alpha}/\sqrt{T\epsilon}+(C_3d+C_{4b})\dot{\alpha}^2/(T\epsilon)+C_{4c} \tag{A91}$$

which is equivalent to

$$\begin{aligned}
&\min_{t\in[T]}E\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\\
\leq& C_1G_2^3\frac{\dot{\alpha}}{\sqrt{T}\epsilon}+G_2C_{4c}\frac{1}{\dot{\alpha}\sqrt{T}}+G_2(C_2d+C_{4a})\frac{1}{T\sqrt{\epsilon}}+G_2(C_3d+C_{4b})\frac{\dot{\alpha}}{T^{3/2}\epsilon}.
\end{aligned} \tag{A92}$$

For simplicity, we rewrite the inequality (A92) as

$$\min_{t\in[T]}E\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\leq D_1\frac{\dot{\alpha}}{\sqrt{T}\epsilon}+D_2\frac{1}{\dot{\alpha}\sqrt{T}}+(D_3d+D_4)\frac{1}{T\sqrt{\epsilon}}+(D_5d+D_6)\frac{\dot{\alpha}}{T^{3/2}\epsilon}, \tag{A93}$$

where the constants $D_1, D_2, D_3, D_4, D_5, D_6$ are constants independent of $T, \epsilon, d$, defined as

$$
\begin{aligned}
D_1 &= L^2 G_2^3 \left( \frac{\beta_1^2 (1-\mu)}{(1-\beta_1)^4} + \frac{\beta_1^2 \mu}{2(1-\beta_1)^2} \right) + L G_2^3 \left( \frac{1-\beta_1+\beta_1\mu}{1-\beta_1} \left( \frac{9\beta_1^2\mu^2}{(1-\beta_1)^2} + 6 \right) + \frac{1}{2} \frac{\beta_1 \mu}{1-\beta_1} \right) + \frac{G_2^3}{2}, \\
D_2 &= \frac{\beta_1 (1-\mu)}{1-\beta_1} G_2^3 + G_2 E \left[ f(\boldsymbol{z}_1) - f(\boldsymbol{z}^*) \right], \\
D_3 &= G_2^3 \left( \frac{\beta_1 (1-\mu)}{1-\beta_1} + 4 \right), \\
D_4 &= \frac{\beta_1 \mu}{1-\beta_1} G_2^3, \\
D_5 &= L^2 G_2^3 \frac{\beta_1^4 (1-\mu)}{(1-\beta_1)^6} + \frac{3}{2} L G_2^3 \frac{1-\beta_1+\beta_1\mu}{1-\beta_1} \frac{\beta_1^2}{(1-\beta_1)^2}, \\
D_6 &= \left( \frac{\beta_1 (1-\mu)}{1-\beta_1} + \frac{3}{2} L \left( 1 + \frac{\beta_1 \mu}{1-\beta_1} \right) \left( \frac{\beta_1}{1-\beta_1} \right)^2 \right) G_2^3.
\end{aligned}
\tag{A94}
$$

This completes the proof.

The same proof also applies to AMSGRAD under the assumption of Theorem 2, except for the coefficients $C_1, C_2, C_3, C_4$ are defined in equation (A31), $\mu = 0$ and $\epsilon$ in ARSG corresponds to $\epsilon^2$ in AMSGRAD. Consequently, by the assumption of Theorem 2, AMSGRAD satisfies

$$
\min_{t \in [T]} E \left[ \|\nabla f(\boldsymbol{x}_t)\|^2 \right] \leq D_1 \frac{\dot{\alpha}}{\sqrt{T}\epsilon^2} + D_2 \frac{1}{\dot{\alpha}\sqrt{T}} + D_3 d \frac{1}{T\epsilon} + (D_5 d + D_6) \frac{\dot{\alpha}}{T^{3/2}\epsilon^2}, \tag{A95}
$$

where the constants $D_1, D_2, D_3, D_5, D_6$ are defined as

$$
\begin{aligned}
D_1 &= L^2 G_2^3 \frac{\beta_1^2}{(1-\beta_1)^4} + \frac{3}{2} L G_2^3 + \frac{G_2^3}{2}, \\
D_2 &= \frac{\beta_1}{1-\beta_1} G_2^3 + G_2 E \left[ f(\boldsymbol{z}_1) - f(\boldsymbol{z}^*) \right], \\
D_3 &= G_2^3 \left( \frac{\beta_1}{1-\beta_1} + 4 \right), \\
D_5 &= L^2 G_2^3 \frac{\beta_1^4}{(1-\beta_1)^6} + \frac{3}{2} L G_2^3 \frac{\beta_1^2}{(1-\beta_1)^2}, \\
D_6 &= \left( \frac{\beta_1}{1-\beta_1} + \frac{3}{2} L \left( \frac{\beta_1}{1-\beta_1} \right)^2 \right) G_2^3.
\end{aligned}
\tag{A96}
$$

### A.3 Proof of Theorem 3

In order to process the projection operation in Algorithm 1, we introduce the following Lemma.

**Lemma 10 [5].** For any $\boldsymbol{Q} \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \in R^d$, suppose $\hat{\boldsymbol{u}}_1 = \min_{\boldsymbol{x} \in \mathcal{F}} \left\| \boldsymbol{Q}^{1/2} (\boldsymbol{x} - \boldsymbol{z}_1) \right\|$ and $\hat{\boldsymbol{u}}_2 = \min_{\boldsymbol{x} \in \mathcal{F}} \left\| \boldsymbol{Q}^{1/2} (\boldsymbol{x} - \boldsymbol{z}_2) \right\|$ then we have $\left\| \boldsymbol{Q}^{1/2} (\hat{\boldsymbol{u}}_1 - \hat{\boldsymbol{u}}_2) \right\| \leq \left\| \boldsymbol{Q}^{1/2} (\boldsymbol{z}_1 - \boldsymbol{z}_2) \right\|$.

The proofs is described in [5] and [7].

From Algorithm 1,

$$
\begin{aligned}
\boldsymbol{x}_{t+1} =& \Pi_{\mathcal{F}, \sqrt{\hat{\boldsymbol{V}}_t}} \left( \boldsymbol{x}_t - \alpha_t \hat{\boldsymbol{V}}_t^{-1/2} \left( (1-\mu_t) \boldsymbol{m}_t + \mu_t \boldsymbol{g}_t \right) \right) \\
=& \arg\min_{\boldsymbol{x} \in \mathcal{F}} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left( \boldsymbol{x} - \left( \boldsymbol{x}_t - \alpha_t \hat{\boldsymbol{V}}_t^{-1/2} \left( (1-\mu_t) \boldsymbol{m}_t + \mu_t \boldsymbol{g}_t \right) \right) \right) \right\|
\end{aligned}
\tag{A97}
$$

Furthermore, $\Pi_{\mathcal{F},\sqrt{\hat{\boldsymbol{V}}_t}}(\boldsymbol{x}^*) = \boldsymbol{x}^*$ for all $\boldsymbol{x}^* \in \mathcal{F}$. Using Lemma 10 with $\hat{\boldsymbol{u}}_1 = \boldsymbol{x}_{t+1}$ and $\hat{\boldsymbol{u}}_2 = \boldsymbol{x}^*$, we have

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\right)\right\|^2 \leq \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\alpha_t\hat{\boldsymbol{V}}_t^{-1/2}\left(\left(1-\mu_t\right)\boldsymbol{m}_t+\mu_t\boldsymbol{g}_t\right)-\boldsymbol{x}^*\right)\right\|^2 \\
=&\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2+\alpha_t^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\left(\left(1-\mu_t\right)\boldsymbol{m}_t+\mu_t\boldsymbol{g}_t\right)\right\|^2-2\alpha_t\left\langle\left(1-\mu_t\right)\boldsymbol{m}_t+\mu_t\boldsymbol{g}_t,\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle \\
=&\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2+\alpha_t^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\left(\left(1-\mu_t\right)\boldsymbol{m}_t+\mu_t\boldsymbol{g}_t\right)\right\|^2 \\
&-2\alpha_t\left\langle\left(1-\mu_t\right)\beta_{1t}\boldsymbol{m}_{t-1}+\left(\mu_t+\left(1-\mu_t\right)\left(1-\beta_{1t}\right)\right)\boldsymbol{g}_t,\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle \\
\leq&\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2+2\alpha_t^2\left(\left(1-\mu_t\right)^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2+\mu_t^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\
&-2\alpha_t\left\langle\left(1-\mu_t\right)\beta_{1t}\boldsymbol{m}_{t-1}+\left(1-\beta_{1t}+\beta_{1t}\mu_t\right)\boldsymbol{g}_t,\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle.
\end{aligned}
\tag{A98}
$$

From the assumptions,

$$
0<1-\beta_1\leq\mu_t=\mu<1, \tag{A99}
$$

rearrange the inequity (A98), we obtain

$$
\begin{aligned}
&\left\langle\boldsymbol{g}_t,\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle \\
\leq&\frac{1}{2\alpha_t\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2-\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\right)\right\|^2\right) \\
&+\frac{\alpha_t}{1-\beta_{1t}\left(1-\mu\right)}\left(\left(1-\mu\right)^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2+\mu^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\
&-\frac{\left(1-\mu\right)\beta_{1t}}{1-\beta_{1t}\left(1-\mu\right)}\left\langle\boldsymbol{m}_{t-1},\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle \\
\leq&\frac{1}{2\alpha_t\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2-\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\right)\right\|^2\right) \\
&+\frac{\alpha_t}{1-\beta_{1t}\left(1-\mu\right)}\left(\left(1-\mu\right)^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2+\mu^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\
&+\frac{\left(1-\mu\right)\beta_{1t}\alpha_t}{2\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_{t-1}\right\|^2+\frac{\left(1-\mu\right)\beta_{1t}}{2\left(1-\beta_{1t}\left(1-\mu\right)\right)\alpha_t}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2.
\end{aligned}
\tag{A100}
$$

Because of the strongly convex assumption in Theorem 3, from (A99) and (A100) the regret satisfies

$$
\begin{aligned}
&R_T=\sum_{i=1}^{T}\left(f_t\left(\boldsymbol{x}_t\right)-f_t\left(\boldsymbol{x}^*\right)\right)\leq\sum_{t=1}^{T}\left(\left\langle\boldsymbol{g}_t,\boldsymbol{x}_t-\boldsymbol{x}^*\right\rangle-\frac{\lambda}{2}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2\right) \\
&\leq\sum_{t=1}^{T}\left(\frac{1}{2\alpha_t\left(1-\beta_{1t}(1-\mu)\right)}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2-\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\right)\right\|^2\right)-\frac{\lambda}{2}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2\right. \\
&+\frac{\alpha_t}{\left(1-\beta_{1t}(1-\mu)\right)}\left(\beta_1^2\left(\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2+\frac{1}{2}\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_{t-1}\right\|^2\right)+\mu^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\
&\left.+\frac{\left(1-\mu\right)\beta_{1t}}{2\alpha_t\left(1-\beta_{1t}(1-\mu)\right)}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2\right).
\end{aligned}
\tag{A101}
$$

We divide RHS of (A101) to three parts, as

$$
\begin{aligned}
Q_1 =& \sum_{t=1}^{T} \left( \frac{1}{2\alpha_t \left(1-\beta_{1t}(1-\mu)\right)} \left( \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 - \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right) \right\|^2 \right) - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right) \\
Q_2 =& \sum_{t=1}^{T} \frac{\alpha_t}{\left(1-\beta_{1t}(1-\mu)\right)} \left( \beta_1^2 \left( \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\|^2 \right) + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) \\
Q_3 =& \sum_{t=1}^{T} \frac{(1-\mu)\beta_{1t}}{2\alpha_t \left(1-\beta_{1t}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 .
\end{aligned}
\tag{A102}
$$

Firstly, we bound the term $Q_1$.

$$
\begin{aligned}
Q_1 =& \frac{1}{2\alpha_1 \left(1-\beta_1(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_1^{1/4} \left(\boldsymbol{x}_1 - \boldsymbol{x}^*\right) \right\|^2 - \frac{1}{2\alpha_T \left(1-\beta_{1T}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_T^{1/4} \left(\boldsymbol{x}_{T+1} - \boldsymbol{x}^*\right) \right\|^2 \\
&+ \sum_{t=2}^{T} \left( \frac{1}{2\alpha_t \left(1-\beta_{1t}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 - \frac{1}{2\alpha_{t-1} \left(1-\beta_{1t-1}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_{t-1}^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 \right) \\
&- \sum_{t=1}^{T} \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \\
\leq& \frac{1}{2\alpha_1 \left(1-\beta_1(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_1^{1/4} \left(\boldsymbol{x}_1 - \boldsymbol{x}^*\right) \right\|^2 - \frac{\lambda}{2} \left\| \boldsymbol{x}_1 - \boldsymbol{x}^* \right\|^2 \\
&+ \sum_{t=2}^{T} \left( \frac{1}{2\left(1-\beta_{1t}(1-\mu)\right)} \left( \frac{1}{\alpha_t} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 - \frac{1}{\alpha_{t-1}} \left\| \hat{\boldsymbol{V}}_{t-1}^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 \right) - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right) \\
=& \sum_{t=1}^{T} \left( \frac{1}{2\left(1-\beta_{1t}(1-\mu)\right)} \left( \frac{t}{\alpha} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 - \frac{t-1}{\alpha} \left\| \hat{\boldsymbol{V}}_{t-1}^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 \right) - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right) \\
\leq& 0
\end{aligned}
\tag{A103}
$$

The first inequity follows from $\beta_t$ is nonincreasing, the second equity follows from $\alpha_t = \alpha/t$. The last inequity is because of the assumption $\alpha \geq \max_{i\in\{1,\cdots,d\}} \left( t\hat{\boldsymbol{v}}_{t,i}^{1/2} - (t-1)\hat{\boldsymbol{v}}_{t-1,i}^{1/2} \right) / \left( \left(1-\beta_1(1-\mu)\right)\lambda \right)$.

Then, we bound the term $Q_2$.

$$
\begin{aligned}
Q_2 \leq& \frac{\alpha}{1-\beta_1(1-\mu)} \sum_{t=1}^{T} \frac{1}{t} \left( \beta_1^2 \left( \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\|^2 \right) + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) \\
\leq& \frac{\alpha}{1-\beta_1(1-\mu)} \left( \sum_{t=1}^{T} \frac{1}{t} \left( \beta_1^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) + \sum_{t=1}^{T-1} \frac{1}{t+1} \frac{\beta_1^2}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 \right) \\
\leq& \frac{\alpha}{1-\beta_1(1-\mu)} \sum_{t=1}^{T} \frac{1}{t} \left( \frac{3}{2} \beta_1^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right)
\end{aligned}
\tag{A104}
$$

The second inequity is because of $\boldsymbol{m}_0 = 0$, and $\hat{\boldsymbol{V}}_t$ is nondecreasing.

We further bound the two terms in RHS of (A104).

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 \\
\leq&\sum_{i=1}^{d}\left(\sum_{j=1}^{t}\prod_{k=1}^{t-j}\beta_{1(t-k+1)}\right)\left(\sum_{j=1}^{t}\prod_{k=1}^{t-j}\beta_{1(t-k+1)}\boldsymbol{g}_{j,i}^2\right)/\sqrt{(1-\beta_2)\sum_{j=1}^{t}\left(\beta_2^{t-j}\boldsymbol{g}_{j,i}^2\right)} \\
\leq&\sum_{i=1}^{d}\left(\sum_{j=1}^{t}\beta_1^{t-j}\right)\left(\sum_{j=1}^{t}\prod_{k=1}^{t-j}\beta_{1(t-k+1)}\boldsymbol{g}_{j,i}^2\right)/\sqrt{(1-\beta_2)\sum_{j=1}^{t}\left(\beta_2^{t-j}\boldsymbol{g}_{j,i}^2\right)} \\
\leq&\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left(\sum_{j=1}^{t}\prod_{k=1}^{t-j}\beta_{1(t-k+1)}\boldsymbol{g}_{j,i}^2\right)/\sqrt{\sum_{j=1}^{t}\left(\beta_2^{t-j}\boldsymbol{g}_{j,i}^2\right)} \\
\leq&\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\sum_{j=1}^{t}\beta_1^{t-j}\boldsymbol{g}_{j,i}^2/\sqrt{\beta_2^{t-j}\boldsymbol{g}_{j,i}^2} \\
=&\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\sum_{j=1}^{t}\gamma^{t-j}\left|\boldsymbol{g}_{j,i}\right| \\
\leq&\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)}G_1
\end{aligned}
\tag{A105}
$$

The first inequality follows from Cauchy-Schwarz inequality.

$$
\begin{aligned}
\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2 &\leq \sum_{i=1}^{d}\frac{\boldsymbol{g}_{ti}^2}{\sqrt{(1-\beta_2)\sum_{j=1}^{t}\left(\beta_2^{t-j}\boldsymbol{g}_{j,i}^2\right)}} \\
&\leq \sum_{i=1}^{d}\frac{\boldsymbol{g}_{t,i}^2}{\sqrt{1-\beta_2}\left|\boldsymbol{g}_{t,i}\right|} \leq \frac{1}{\sqrt{1-\beta_2}}G_1.
\end{aligned}
\tag{A106}
$$

Combining (A104), (A105), and (A106), we obtain

$$
\begin{aligned}
Q_2 &\leq \frac{\alpha G_1}{(1-\beta_1(1-\mu))\sqrt{1-\beta_2}}\sum_{t=1}^{T}\frac{1}{t}\left(\frac{3}{2}\frac{\beta_1^2}{(1-\beta_1)(1-\gamma)}+\mu^2\right) \\
&\leq \frac{\alpha G_1}{(1-\beta_1(1-\mu))\sqrt{1-\beta_2}}\left(\frac{3}{2}\frac{\beta_1^2}{(1-\beta_1)(1-\gamma)}+\mu^2\right)(\log(T)+1).
\end{aligned}
\tag{A107}
$$

The first inequity follows from the assumptions $\alpha_t = \alpha/t$.

Finally, we bound the term $Q_3$.

$$
\begin{aligned}
Q_3 &\leq \frac{1}{2(1-\beta_1(1-\mu))}\sum_{t=1}^{T}\left(\frac{(1-\mu)\beta_{1t}}{\alpha_t}\left\|\hat{\boldsymbol{V}}_t^{1/4}(\boldsymbol{x}_t-\boldsymbol{x}^*)\right\|^2\right) \\
&= \frac{(1-\mu)\beta_1}{2\alpha(1-\beta_1(1-\mu))}\sum_{t=1}^{T}\left(\frac{1}{t}\left\|\hat{\boldsymbol{V}}_t^{1/4}(\boldsymbol{x}_t-\boldsymbol{x}^*)\right\|^2\right) \\
&\leq \frac{(1-\mu)\beta_1}{2\alpha(1-\beta_1(1-\mu))}\sum_{t=1}^{T}\frac{1}{t\sqrt{\epsilon}}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|^2 \\
&\leq \frac{(1-\mu)\beta_1 D_{\mathcal{F}}^2}{2\alpha(1-\beta_1(1-\mu))}\frac{1+\log(T)}{\sqrt{\epsilon}}.
\end{aligned}
\tag{A108}
$$

Both the first equity and the first inequity follow from the assumptions $\alpha_t = \alpha/t$ and $\beta_{1t} = \beta_1/t^2$. The last inequity is due to $\hat{\boldsymbol{v}}_{t,i}$ is nondecreasing by definition.

**Algorithm A1** ASGD Algorithm

**Input**: initial parameter vector $\boldsymbol{x}_1$, short step $\ddot{\alpha}$, long step parameter $\ddot{\kappa} \geq 1$, statistical advantage parameter $\ddot{\xi} \leq \sqrt{\ddot{\kappa}}$, iteration number $T$

**Output**: parameter vector $\boldsymbol{x}_T$

1: Set $\bar{x}_1 = \boldsymbol{x}_1$, $\ddot{\beta} = 1 - 0.7^2\ddot{\xi}/\ddot{\kappa}$.
2: **for** $t = 1$ to $T - 1$ **do**
3: $\quad \boldsymbol{g}_t = \nabla f_t(\boldsymbol{x}_t)$.
4: $\quad \bar{\boldsymbol{x}}_{t+1} = \ddot{\beta}\bar{\boldsymbol{x}}_t + (1 - \ddot{\beta})\left(\boldsymbol{x}_t - \frac{\ddot{\kappa}\ddot{\alpha}}{0.7}\boldsymbol{g}_t\right)$.
5: $\quad \boldsymbol{x}_{t+1} = \frac{0.7}{0.7+(1-\ddot{\beta})}\left(\boldsymbol{x}_t - \ddot{\alpha}\boldsymbol{g}_t\right) + \frac{1-\ddot{\beta}}{0.7+(1-\ddot{\beta})}\bar{\boldsymbol{x}}_{t+1}$.
6: **end for**

Combining (A101), (A102), (A103), (A107), and (A108), we obtain

$$R_T \leq \left(\frac{\alpha G_1}{\sqrt{1-\beta_2}}\left(\frac{3}{2}\frac{\beta_1^2}{(1-\beta_1)(1-\gamma)} + \mu^2\right) + \frac{(1-\mu)\beta_1 D_{\mathcal{F}}^2}{2\alpha\sqrt{\epsilon}}\right)\frac{1+\log(T)}{1-\beta_1(1-\mu)}. \tag{A109}$$

The proof is complete.

As a special case of ARSG with $\mu = 0$, AMSGRAD shares the bound (A109) under the same assumption except for $\mu = 0$, and $\epsilon$ is substituted by $\epsilon^2$. Compared with AMSGRAD, ARSG improves the bound by redefining $\epsilon$, and also improves the coefficients in typical cases where $1 - \beta_1 \ll 1, \epsilon \ll 1$.

### A.4 Equivalence of RSG and ASGD

Jain et al. [3] shows that ASGD [4, 3] improves on SGD in any information-theoretically admissible regime. By taking a long step as well as short step and an appropriate average of both of them, ASGD tries to make similar progress on different eigen-directions.

The pseudo code of ASGD is shown in Algorithm A1. It maintains two iterates: descent iterate $\boldsymbol{x}_t$ and a running average $\bar{x}_t$. The running average is a weighted average of the previous average and a long gradient step from the descent iterate, while the descent iterate is updated as a convex combination of short gradient step from the descent iterate and the running average. The method takes 3 hyper-parameters: short step $\ddot{\alpha}$, long step parameter $\ddot{\kappa}$, and statistical advantage parameter $\ddot{\xi}$. $\ddot{\alpha}$ is the same as the step size in SGD. For convex functions, $\ddot{\kappa}$ is an estimation of the condition number. The statistical advantage parameter $\ddot{\xi} \leq \sqrt{\ddot{\kappa}}$ captures trade off between statistical and computational condition numbers, and $\ddot{\xi} \ll \sqrt{\ddot{\kappa}}$ in high stochasticity regimes.

Now we demonstrate that RSG is a more efficient equivalent form of ASGD.

We rewrite the update of Algorithm A1 as

$$\begin{aligned}
&\begin{bmatrix}\bar{\boldsymbol{x}}_{t+1}\\ \boldsymbol{x}_{t+1}\end{bmatrix} = \ddot{A}\begin{bmatrix}\bar{\boldsymbol{x}}_t\\ \boldsymbol{x}_t\end{bmatrix} + \ddot{b}\boldsymbol{g}_t\\
&\ddot{A} = \begin{bmatrix}\ddot{\beta} & 1-\ddot{\beta}\\ \frac{(1-\ddot{\beta})\ddot{\beta}}{(1-\ddot{\beta})+0.7} & \frac{(1-\ddot{\beta})^2+0.7}{(1-\ddot{\beta})+0.7}\end{bmatrix}, \ddot{b} = \begin{bmatrix}\frac{\ddot{\beta}-1}{0.7}\ddot{\kappa}\ddot{\alpha}\\ -\frac{0.7^2+(1-\ddot{\beta})^2\ddot{\kappa}}{0.7((1-\ddot{\beta})+0.7)}\ddot{\alpha}\end{bmatrix}.
\end{aligned} \tag{A110}$$

Define the variable transform as

$$\begin{bmatrix}\tilde{\boldsymbol{m}}_t\\ \boldsymbol{x}_t\end{bmatrix} = \ddot{T}\begin{bmatrix}\bar{\boldsymbol{x}}_t\\ \boldsymbol{x}_t\end{bmatrix}, \ddot{T} = \begin{bmatrix}\ddot{l} & \ddot{k}\ddot{l}\\ 0 & 1\end{bmatrix}, \tag{A111}$$

where $\ddot{k}$ are $\ddot{l}$ are adjustable coefficients.

Combining (A110) and (A111), we obtain

$$\begin{bmatrix}\tilde{\boldsymbol{m}}_{t+1}\\ \boldsymbol{x}_{t+1}\end{bmatrix} = \tilde{T}\begin{bmatrix}\tilde{\boldsymbol{m}}_t\\ \boldsymbol{x}_t\end{bmatrix} + \ddot{T}\ddot{b}\boldsymbol{g}_t, \tilde{T} = \ddot{T}\ddot{A}\ddot{T}^{-1}. \tag{A112}$$

In order to minimize the number of vector computations, we solve the adjustable coefficients $\ddot{k}$ and $\ddot{l}$ by assigning $\tilde{T}_{1,2} = 0, \tilde{T}_{2,1} = 1$. We choose the solution as

$$\ddot{k} = -1, \ddot{l} = \frac{(1-\ddot{\beta})\ddot{\beta}}{(1-\ddot{\beta})+0.7}. \tag{A113}$$

Combining (A112) and (A113), we obtain

$$\begin{bmatrix} \tilde{\boldsymbol{m}}_{t+1} \\ \boldsymbol{x}_{t+1} \end{bmatrix} = \tilde{T} \begin{bmatrix} \tilde{\boldsymbol{m}}_{t} \\ \boldsymbol{x}_{t} \end{bmatrix} + \ddot{T}\ddot{b}\boldsymbol{g}_t, \tilde{T} = \begin{bmatrix} \frac{0.7\ddot{\beta}}{(1-\ddot{\beta})+0.7} & 0 \\ 1 & 1 \end{bmatrix}. \tag{A114}$$

When the hyper-parameters are constant, the concise form of RSG (3) can be rearranged as

$$\begin{aligned} \boldsymbol{m}_t &= \beta \boldsymbol{m}_{t-1} - \alpha(1-\beta)(1-\mu)\nabla f_t(\boldsymbol{x}_t) \\ \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t + \boldsymbol{m}_t - \alpha\mu\nabla f_t(\boldsymbol{x}_t), \end{aligned} \tag{A115}$$

Define $\ddot{\boldsymbol{m}}_t = \beta \boldsymbol{m}_{t-1}$, (A114) is further rearranged as

$$\begin{aligned} \ddot{\boldsymbol{m}}_{t+1} &= \beta \ddot{\boldsymbol{m}}_t - \alpha\beta(1-\beta)(1-\mu)\nabla f_t(\boldsymbol{x}_t) \\ \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t + \ddot{\boldsymbol{m}}_t - \alpha(1-\beta+\beta\mu)\nabla f_t(\boldsymbol{x}_t), \end{aligned} \tag{A116}$$

The update (A114) and (A116) are identical. The momentum coefficient of RSG is

$$\beta = \frac{0.7\ddot{\beta}}{(1-\ddot{\beta})+0.7} = (\ddot{\kappa} - 0.49\ddot{\xi})/(\ddot{\kappa} + 0.7\ddot{\xi}), \tag{A117}$$

where the second equity follows from the definition of $\ddot{\beta}$ in Algorithm A1.

It should be noted that the 3 hyper-parameters of ASGD vary in large ranges, and are difficult to estimate. The huge costs in tuning limits the application of ASGD, while the cost for tuning the hyper-parameters for RSG can be greatly reduced by the analysis in Section 3.

RSG also reduces the computational overheads of ASGD in each iteration. Besides the gradient computation, ASGD requires 6 scalar vector multiplications and 4 vector additions per iteration, while RSG reduces the costs to 3 scalar vector multiplications and 3 vector additions.

# B Experimental details

## B.1 Experiments on MNIST

We compare the performance of SGD, ADAM, NADAM, AMSGRAD, RANGER, RSG, ARSG and ARSGB, for training logistic regression and a simple convolutional neural network (CNN) on the MNIST dataset. The dataset consists of 60k training images and 10k validation images in 10 classes. The image size is $28 \times 28$.

The CNN architecture has two $5 \times 5$ convolutional layers, with 20 and 50 outputs. Each convolutional layer is followed by Batch Normalization (BN) [2] and a $2 \times 2$ max pooling. The network ends with a 500-way fully-connected layer with BN and ReLU [6], a 10-way fully-connected layer, and softmax.

The training of logistic regression runs for 20 epochs, and the CNNs are trained for 10 epochs. ARSGB performs the boost at the 6th epoch. The minibatch size is 256. The hyper-parameters for all the methods are chosen by grid search, which will be described in detail in the B.5, and the best results in training are reported.

## B.2 Experiments on WikiText-2

In the experiment, we train LSTM on the WikiText-2 dataset, that consists of 2088628 tokens (600 articles) as the training set and 217646 tokens (60 articles) as the validation set. We use the codes of the PyTorch official example available at `https://github.com/pytorch/examples/tree/master/word_language_model`.

The LSTM we train has 2 layers each containing 200 hidden units. The dimension of word embeddings is 200. For the input embedding layer, the embedding layer, and each hidden layer, we use dropout with probability 0.2. A weight decay of 0.00001 is used for regularization, and the minibatch size is 128.

The training for each network runs for 40 epochs. We divide the step size by 4 after the 20th and 30th epoch. ARSGB performs the boost at the 11th epoch. Gradient clipping of 0.25 is applied to all optimizers. The hyper-parameters are chosen by grid search to minimizes the training loss before the dropping of step size, which will be described in detail in the B.5. The training finishes at the end of the 40th epoch, since the validation losses for all the methods have been rebounding in the final epochs.

### B.3 Experiments on CIFAR-10

In the experiment, we train ResNet-20 on the CIFAR-10 dataset, that consists of 50k training images and 10k validation images in 10 classes. The image size is $32 \times 32$.

In training, the network inputs are $28 \times 28$ images randomly cropped from the original images or their horizontal flips. The inputs are subtracted by the global mean and divided by the standard deviation. The architecture of the network is as follows: The first layer is $3 \times 3$ convolutions. Then we use a stack of 18 layers with $3 \times 3$ convolutions on the feature maps of sizes $\{28, 14, 7\}$ respectively, with 6 layers for each feature map size. The numbers of filters are $\{16, 32, 64\}$ respectively. A shortcut connection is added to each pair of $3 \times 3$ filters. The subsampling is performed by convolutions with a stride of 2. Batch normalization is adopted right after each convolution and before the ReLU activation. The network ends with a global average pooling, a 10-way fully-connected layer, and softmax. In validation, the original images are used as inputs.

We train ResNet-20 on CIFAR-10 using SGD, ADAM, NADAM, AMSGRAD, RANGER, RSG, ARSG and ARSGB. The training for each network runs for 80 epochs. The hyper-parameters are selected by grid search (see the B.5 for details), and we divide the constant step size by 10 at the 12000th iteration (in the 62th epoch). Two group of hyper-parameters are obtained for each method, one of which is the fast mode to minimizes the training loss before the dropping of step size, and the other is the fine mode to maximizes the best validation accuracy. ARSGB performs the boost at the 21th epoch in the fast mode, and the 51th epoch in the fine mode. A weight decay of 0.001 is used for regularization, and the minibatch size is 256. The training finishes at the end of the 80th epoch, since the validation losses for all the methods have been rebounding in the final epochs.

### B.4 Experiments on ImageNet

In the experiment, we train ResNet-50 on the ImageNet 2012 classification dataset that consists of 1.28 million training images [2] and 50k validation images in 1000 classes. We use the codes of the PyTorch official example available at `https://github.com/pytorch/examples/tree/master/imagenet`. The model used is also the PyTorch official implementation of ResNet-50.

In training, the network inputs are $224 \times 224$ images randomly resized and cropped from the original images or their horizontal flips. The inputs are subtracted by the global mean and divided by the standard deviation. In validation, the input are $224 \times 224$ images center-cropped from the original images which are resized to 256 by the shorter edge.

Exhaustive grid search to obtain the optimal hyper-parameters is not available because the computation is too huge. We choose SGD with the default hyper-parameters of the PyTorch official examples as baseline, in which the initial learning rate is 0.1 and momentum value is 0.9 ($\alpha = 1, \beta = 0.9$). For ARSG, we choose $\alpha = 0.02$ by the concept that the maximum elementwise step size should be close to the default value of SGD. We also choose $\alpha = 0.01$ since a relatively small step size tends to converge fast in the latter echos. The other hyper-parameters are set according to the default values for good generalization (fine mode), as $\beta_1 = 0.999, \beta_2 = 0.99, \mu = 0.1, \epsilon = 0.001$. For ADAM, we set $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ by the default value, and perform grid search of 15 echos to select the step size $\alpha$ from $\{0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002\}$. We choose

[2] In our training set 978 images are corrupted. We are unable to update these files since our request is not approved by the official website. We delete the files for simplicity. Consequently, the generalization may be slightly harmed.

Table A1: The hyper-parameters in the experiments

| | Experiments on MNIST | | Experiments on WikiText-2 |
|---|---|---|---|
| Methods | Logistic regression | CNN | LSTM |
| SGD | $(2.0, 0.99)$ | $(1.0, 0.9)$ | $(50, 0.9)$ |
| ADAM | $(0.005, 0.999, 0.999)$ | $(0.0005, 0.9, 0.999)$ | $(0.002, 0, 0.999)$ |
| NADAM | $(0.01, 0.999, 0.999)$ | $(0.0005, 0.9, 0.999)$ | $(0.002, 0.9, 0.99)$ |
| AMSGRAD | $(0.005, 0.99, 0.999)$ | $(0.005, 0.9, 0.99)$ | $(0.005, 0, 0.99)$ |
| RANGER | $(0.02, 0.99, 0.999, 5, 0.5)$ | $(0.005, 0.9, 0.999, 5, 0.2)$ | $(0.002, 0, 0.999, 5, 0.8)$ |
| RSG | $(5.0, 0.999, 0.1)$ | $(5.0, 0.999, 0.1)$ | $(100, 0.999, 0.1)$ |
| ARSG | $(0.05, 0.999, 0.99, 0.1)$ | $(0.01, 0.999, 0.99, 0.1)$ | $(0.02, 0.999, 0.99, 0.1)$ |
| ARSGB | $(0.1, 0.999, 0.99, 0.05)$ | $(0.05, 0.999, 0.99, 0.05)$ | $(0.05, 0.999, 0.99, 0.05)$ |

| | ResNet-20 on CIFAR-10 | |
|---|---|---|
| Methods | Fast mode | Fine mode |
| SGD | $(0.1, 0.9)$ | $(0.5, 0.9)$ |
| ADAM | $(0.0005, 0.9, 0.99)$ | $(0.001, 0.9, 0.99)$ |
| NADAM | $(0.0005, 0.99, 0.999)$ | $(0.001, 0.9, 0.999)$ |
| AMSGRAD | $(0.0005, 0.9, 0.99)$ | $(0.001, 0.9, 0.999)$ |
| RANGER | $(0.005, 0.99, 0.999, 5, 0.2)$ | $(0.005, 0.9, 0.999, 5, 0.2)$ |
| RSG | $(0.2, 0.999, 0.1)$ | $(1, 0.999, 0.1)$ |
| ARSG | $(0.002, 0.999, 0.99, 0.1)$ | $(0.05, 0.999, 0.99, 0.1)$ |
| ARSGB | $(0.005, 0.999, 0.99, 0.05)$ | $(0.05, 0.999, 0.99, 0.05)$ |

$\alpha = 0.0001, 0.0002$ because they are the fastest and perform roughly the same in the end of the grid search. The training for each network runs for 90 epochs, and the step size $\alpha$ for all the methods is divided by 10 at the 31th and 61 epochs following the SGD baseline. A weight decay of $0.0001$ is used for regularization, and the minibatch size is $256$.

### B.5 Hyper-parameters selection by grid search

We use piecewise constant hyper-parameters in the experiments. For ADAM, NADAM, and AMSGRAD, the hyper-parameters $(\alpha, \beta_1, \beta_2)$ are selected by grid search from $\{0.0005, 0.001, 0.002, 0.005, 0.01, 0.02\} \times \{0, 0.9, 0.99, 0.999\} \times \{0.99, 0.999\}$, where $\beta_1 = 0$ is excluded for NADAM. Although RANGER generally improves the performance compared with the above adaptive methods, it requires two more hyper-parameters for look ahead optimization as the synchronization period $k_{\text{LA}}$ and slow weights step $\alpha_{\text{LA}}$. The hyper-parameters $(\alpha, \beta_1, \beta_2, k_{\text{LA}}, \alpha_{\text{LA}})$ are selected by grid search from $\{0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05\} \times \{0.9, 0.95, 0.99, 0.999\} \times \{0.999\} \times \{5\} \times \{0.2, 0.5, 0.8\}$. For SGD, $\alpha$ is selected by grid search from $\{2, 5, 10, 20, 50, 100, 200\}$ in the WikiText-2 experiment, and $\{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$ in the image classification experiments. $\beta$ is selected from $\{0, 0.9, 0.99, 0.999\}$. For RSG, $\alpha$ is selected from the same grid points as SGD, and $(\beta, \mu)$ are set to $(0.999, 0.1)$ according to the default values. For ARSG and ARSGB, $\alpha$ is selected by grid search from $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$, $(\beta_1, \beta_2, \mu)$ are set to the default values as $(0.999, 0.99, 0.1)$ for ARSG and $(0.999, 0.99, 0.05)$ for ARSGB. The small positive constant $\epsilon$ is set as $10^{-8}$ for all the adaptive methods, excepting for $\epsilon = 0.001$ for ARSG and ARSGB in the fine mode of the CIFAR10 experiment.

Table A1 shows the hyper-parameters selected.

The ImageNet experiment is carried out on a workstation with an Intel Xeon E5-2678 v3 CPU and 8 NVIDIA 2080TI GPUs (We can use at most 4 GPUs). Other experiments are carried out on a workstation with an Intel Xeon Gold 6148 CPU and a NVIDIA V100 GPU. The source code of ARSG and the examples can be downloaded at `https://github.com/rationalspark/NAMSG`. The simulation environment for the experiments on MNIST and CIFAR-10 is MXNET, which can be downloaded at `http://mxnet.incubator.apache.org`. Other experiments are performed using PyTorch, which can be downloaded at `https://pytorch.org/`. The MNIST dataset can be downloaded at `http://yann.lecun.com/exdb/mnist`; the WikiText-2

dataset can be downloaded at https://www.salesforce.com/products/einstein/ai-research/the-wikitext-dependency-language-modeling-dataset/; the CIFAR-10 dataset can be downloaded at http://www.cs.toronto.edu/~kriz/cifar.html; the ImageNet dataset can be downloaded at http://www.image-net.org/.

## References


[1] 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.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[3] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 06–09 Jul 2018.

[4] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *International Conference on Learning Representations*, 2018.

[5] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. *CoRR*, abs/1002.4908, 2010.

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

[7] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

[8] Qianqian Tong, Guannan Liang, and Jinbo Bi. Calibrating the Adaptive Learning Rate to Improve Convergence of ADAM. *arXiv e-prints*, abs/1908.00700, 2019.