Supplementary Materials

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 $\tau \in [\tau_{\min}, \tau_{\max}]$.

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

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

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

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

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

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

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 (A15),

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

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.

2 Proof of Theorem 2

In this proof, we use $\boldsymbol{y}_i$ to denote the i^{th} coordinate of a vector $\boldsymbol{y}$.

From Algorithm 1,

$$\begin{aligned}\boldsymbol{x}_{t+1} =& \Pi_{\mathcal{F},\sqrt{\hat{\boldsymbol{V}}_t}}\left(\boldsymbol{x}_t - \alpha_t \hat{\boldsymbol{V}}_t^{-1/2}\left(\left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t\right)\right) \\ =& \arg\min_{\boldsymbol{x}\in\mathcal{F}} \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x} - \left(\boldsymbol{x}_t - \alpha_t \hat{\boldsymbol{V}}_t^{-1/2}\left(\left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t\right)\right)\right)\right\|\end{aligned} \tag{A28}$$

Furthermore, $\Pi_{\mathcal{F},\sqrt{\hat{\boldsymbol{V}}_t}}(\boldsymbol{x}^*) = \boldsymbol{x}^*$ for all $\boldsymbol{x}^* \in \mathcal{F}$. Using Lemma A1 with $\hat{\boldsymbol{u}}_1 = \boldsymbol{x}_{t+1}$ and $\hat{\boldsymbol{u}}_2 = \boldsymbol{x}^*$, we have

$$\begin{aligned}&\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)\right\|^2 \le \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \alpha_t \hat{\boldsymbol{V}}_t^{-1/2}\left(\left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t\right) - \boldsymbol{x}^*\right)\right\|^2 \\ =& \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 + \alpha_t^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\left(\left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t\right)\right\|^2 - 2\alpha_t \left\langle \left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\ =& \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 + \alpha_t^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\left(\left(1-\mu_t\right)\boldsymbol{m}_t + \mu_t \boldsymbol{g}_t\right)\right\|^2 \\ &- 2\alpha_t \left\langle \left(1-\mu_t\right)\beta_{1t}\boldsymbol{m}_{t-1} + \left(\mu_t + \left(1-\mu_t\right)\left(1-\beta_{1t}\right)\right)\boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\ \le& \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 + 2\alpha_t^2 \left(\left(1-\mu_t\right)^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 + \mu_t^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\ &- 2\alpha_t \left\langle \left(1-\mu_t\right)\beta_{1t}\boldsymbol{m}_{t-1} + \left(1-\beta_{1t}+\beta_{1t}\mu_t\right)\boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle.\end{aligned} \tag{A29}$$

Since $0 \le \beta_{1t} < 1$, from the assumptions,

$$0 < 1 - \beta_1 \le \mu_t = \mu < 1. \tag{A30}$$

Rearrange the inequity (A29), we obtain

$$\begin{aligned}&\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\ \le& \frac{1}{2\alpha_t\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 - \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)\right\|^2\right) \\ &+ \frac{\alpha_t}{1-\beta_{1t}\left(1-\mu\right)}\left(\left(1-\mu\right)^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 + \mu^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\ &- \frac{\left(1-\mu\right)\beta_{1t}}{1-\beta_{1t}\left(1-\mu\right)}\left\langle \boldsymbol{m}_{t-1}, \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\ \le& \frac{1}{2\alpha_t\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 - \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)\right\|^2\right) \\ &+ \frac{\alpha_t}{1-\beta_{1t}\left(1-\mu\right)}\left(\left(1-\mu\right)^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 + \mu^2 \left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2\right) \\ &+ \frac{\left(1-\mu\right)\beta_{1t}\alpha_t}{2\left(1-\beta_{1t}\left(1-\mu\right)\right)}\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_{t-1}\right\|^2 + \frac{\left(1-\mu\right)\beta_{1t}}{2\left(1-\beta_{1t}\left(1-\mu\right)\right)\alpha_t}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2.\end{aligned} \tag{A31}$$

For simplicity, denote

$$\begin{aligned}P_1 =& \sum_{t=1}^{T}\left(\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2 - \left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right)\right\|^2 + \left(1-\mu\right)\beta_{1t}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t - \boldsymbol{x}^*\right)\right\|^2\right) \\ &/\left(2\alpha_t\left(1-\beta_{1t}(1-\mu)\right)\right) \\ P_2 =& \sum_{t=1}^{T}\alpha_t\left(\left(1-\mu\right)^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 + \mu^2\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{g}_t\right\|^2 + \frac{1}{2}(1-\mu)\beta_{1t}\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_{t-1}\right\|^2\right) \\ &/(1-\beta_{1t}(1-\mu))\end{aligned} \tag{A32}$$

Because of the convexity of the objective function, the regret satisfies

$$R_T = \sum_{i=1}^{T} \left(f_t\left(\boldsymbol{x}_t\right) - f_t\left(\boldsymbol{x}^*\right) \right) \leq \sum_{t=1}^{T} \left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle \leq P_1 + P_2. \tag{A33}$$

The first inequity follows from the convexity of function f_t. The second inequality is due to (A31).

We now bound the term $\sum_{t=1}^{T} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2$. We have

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 = \sum_{t=1}^{T-1} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 + \alpha_T \sum_{i=1}^{d} \frac{\boldsymbol{g}_{T,i}^2}{\sqrt{\hat{\boldsymbol{v}}_{T,i}}} \\
\leq &\sum_{t=1}^{T-1} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 + \alpha_T \sum_{i=1}^{d} \frac{\boldsymbol{g}_{T,i}^2}{\sqrt{\boldsymbol{v}_{T,i}}} \leq \sum_{t=1}^{T-1} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 + \frac{\alpha}{\sqrt{T}} \sum_{i=1}^{d} \frac{\boldsymbol{g}_{T,i}^2}{\sqrt{\left(1-\beta_2\right) \sum_{j=1}^{T} \beta_2^{T-j} \boldsymbol{g}_{j,i}^2}} \\
\leq &\sum_{t=1}^{T-1} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 + \frac{\alpha}{\sqrt{T\left(1-\beta_2\right)}} \sum_{i=1}^{d} \left| \boldsymbol{g}_{T,i} \right| \leq \frac{\alpha}{\sqrt{1-\beta_2}} \sum_{t=1}^{T} \left(\frac{1}{\sqrt{t}} \sum_{i=1}^{d} \left| \boldsymbol{g}_{t,i} \right| \right) \\
\leq &\frac{\alpha}{\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 \sqrt{\sum_{t=1}^{T} \frac{1}{t}} \leq \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 .
\end{aligned} \tag{A34}$$

In (A34), the second inequity is follows from the definition of v_t , the fifth inequality is due to Cauchy-Schwarz inequality. The final inequality is due to the following bound on harmonic sum: $\sum_{t=1}^{T} 1/t \leq 1 + \log(T)$.

From (A30), (A34), and Lemma A2, which bound $\sum_{t=1}^{T} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2$, we further bound the term P_2 as

$$\begin{aligned}
P_2 \leq &\frac{1}{1-\beta_1(1-\mu)} \left(\sum_{t=1}^{T} \alpha_t \left((1-\mu)^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2}(1-\mu)\beta_{1t} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\| \right) \right. \\
&\left. + \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \mu^2 \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 \right) \\
\leq &\frac{1}{1-\beta_1(1-\mu)} \left(\beta_1^2 \sum_{t=1}^{T} \alpha_t \left(\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\| \right) \right. \\
&\left. + \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \mu^2 \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 \right) \\
\leq &\frac{1}{1-\beta_1(1-\mu)} \left(\beta_1^2 \left(\sum_{t=1}^{T} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \sum_{t=1}^{T-1} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 \right) \right. \\
&\left. + \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \mu^2 \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 \right) \\
\leq &\frac{1}{1-\beta_1(1-\mu)} \left(\frac{3}{2} \beta_1^2 \sum_{t=1}^{T} \alpha_t \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{\alpha \sqrt{1+\log(T)}}{\sqrt{1-\beta_2}} \mu^2 \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 \right) \\
\leq &\left(\frac{3\beta_1^2}{2\left(1-\beta_1\right)\left(1-\gamma\right)} + \mu^2 \right) \frac{\alpha \sqrt{1+\log(T)}}{\left(1-\beta_1(1-\mu)\right)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left\| \boldsymbol{g}_{1:T,i} \right\|_2 .
\end{aligned} \tag{A35}$$

The third inequity is due to $\beta_{1t} \geq \beta_{1t+1}$ and $\hat{\boldsymbol{v}}_{t,i}^{1/2} / \alpha_t \geq \hat{\boldsymbol{v}}_{t-1,i}^{1/2} / \alpha_{t-1}$ by definition.

We also have

$$
\begin{aligned}
&P_1 \\
&\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+\sum_{t=2}^{T}\left(\frac{1}{2\alpha_t\left(1-\beta_{1t}(1-\mu)\right)}\left\|\hat{\boldsymbol{V}}_1^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2-\right. \\
&\left.\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{\beta_{1t}(1-\mu)}{2\alpha_t\left(1-\beta_{1t}(1-\mu)\right)}\left\|\hat{\boldsymbol{V}}_1^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\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+\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\right.\right. \\
&\left.\left.-\frac{1}{\alpha_{t-1}}\left\|\hat{\boldsymbol{V}}_{t-1}^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2\right)+\sum_{t=1}^{T}\frac{\beta_{1t}(1-\mu)}{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. \\
&\leq \frac{1}{2\left(1-\beta_1(1-\mu)\right)}\left(\frac{1}{\alpha_1}\left\|\hat{\boldsymbol{V}}_1^{1/4}\left(\boldsymbol{x}_1-\boldsymbol{x}^*\right)\right\|^2+\sum_{t=2}^{T}\left(\frac{1}{\alpha_t}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2\right.\right. \\
&\left.\left.-\frac{1}{\alpha_{t-1}}\left\|\hat{\boldsymbol{V}}_{t-1}^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2\right)+\sum_{t=1}^{T}\frac{\beta_{1t}(1-\mu)}{\alpha_t}\left\|\hat{\boldsymbol{V}}_t^{1/4}\left(\boldsymbol{x}_t-\boldsymbol{x}^*\right)\right\|^2\right) \\
&=\frac{1}{2\left(1-\beta_1(1-\mu)\right)}\left(\frac{1}{\alpha_1}\sum_{i=1}^{d}\hat{\boldsymbol{v}}_{1,i}^{1/2}\left(\boldsymbol{x}_{1,i}-\boldsymbol{x}_i^*\right)^2+\sum_{t=2}^{T}\left(\sum_{i=1}^{d}\left(\boldsymbol{x}_{t,i}-\boldsymbol{x}_i^*\right)^2\left(\frac{\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_t}-\frac{\hat{\boldsymbol{v}}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right)\right)\right. \\
&\left.+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}(1-\mu)\left(\boldsymbol{x}_{t,i}-\boldsymbol{x}_i^*\right)^2\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_t}\right) \\
&\leq \frac{D_\infty^2}{2\left(1-\beta_1(1-\mu)\right)}\left(\frac{1}{\alpha_1}\sum_{i=1}^{d}\hat{\boldsymbol{v}}_{1,i}^{1/2}+\sum_{t=2}^{T}\left(\sum_{i=1}^{d}\left(\frac{\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_t}-\frac{\hat{\boldsymbol{v}}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right)\right)+\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}(1-\mu)\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_t}\right) \\
&=\frac{D_\infty^2\sqrt{T}}{2\left(1-\beta_1(1-\mu)\right)\alpha}\sum_{i=1}^{d}\hat{\boldsymbol{v}}_{t,i}^{1/2}+\frac{(1-\mu)D_\infty^2}{2\left(1-\beta_1(1-\mu)\right)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_t}
\end{aligned}
\tag{A36}
$$

In (A36), the second inequity follows from the assumption $\beta_{1t} < \beta_{1t-1}$, the third and the last inequality is due to $\hat{\boldsymbol{v}}_{t,i}^{1/2}/\alpha_t \geq \hat{\boldsymbol{v}}_{t-1,i}^{1/2}/\alpha_{t-1}$ by definition and the assumption $\alpha_t = \alpha/\sqrt{t}$.

Combining (A33), (A35), and (A36), we obtain

$$
\begin{aligned}
R_T \leq & \frac{1}{(1-\beta_1(1-\mu))}\left(\frac{D_\infty^2\sqrt{T}}{2\alpha}\sum_{i=1}^{d}\hat{\boldsymbol{v}}_{t,i}^{1/2}+\frac{(1-\mu)D_\infty^2}{2}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}\hat{\boldsymbol{v}}_{t,i}^{1/2}}{\alpha_i}\right. \\
&\left.+\left(\frac{3\beta_1^2}{2\left(1-\beta_1\right)\left(1-\gamma\right)}+\mu^2\right)\frac{\alpha\sqrt{1+\log(T)}}{\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:T,i}\right\|_2\right).
\end{aligned}
\tag{A37}
$$

The proof is complete.

The Lemmas used in the proof are as follows:

Lemma A1. [McMahan and Streeter, 2010]
For any $\boldsymbol{Q} \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \in R^d$, suppose $\hat{\boldsymbol{u}}_1 = \min_{\boldsymbol{x}\in\mathcal{F}}\left\|\boldsymbol{Q}^{1/2}\left(\boldsymbol{x}-\boldsymbol{z}_1\right)\right\|$ and $\hat{\boldsymbol{u}}_2 = \min_{\boldsymbol{x}\in\mathcal{F}}\left\|\boldsymbol{Q}^{1/2}\left(\boldsymbol{x}-\boldsymbol{z}_2\right)\right\|$ then we have $\left\|\boldsymbol{Q}^{1/2}\left(\hat{\boldsymbol{u}}_1-\hat{\boldsymbol{u}}_2\right)\right\| \leq \left\|\boldsymbol{Q}^{1/2}\left(\boldsymbol{z}_1-\boldsymbol{z}_2\right)\right\|$.

Lemma A2. [Reddi et al., 2018]
For the parameter settings and conditions assumed in Theorem 1, which is the same as Theorem 4 in Reddi et al. [2018], we have

$$
\sum_{t=1}^{T}\alpha_t\left\|\hat{\boldsymbol{V}}_t^{-1/4}\boldsymbol{m}_t\right\|^2 \leq \frac{\alpha\sqrt{1+\log T}}{\left(1-\beta_1\right)\left(1-\gamma\right)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:T,i}\right\|_2.
$$

The proofs of Lemma A1 and A2 are described in Reddi et al. [2018].

3 Proof of Theorem 3

Because of the objective function is strongly convex, from (A30) and (A31) the regret satisfies

$$
\begin{aligned}
R_T &= \sum_{i=1}^{T} \left(f_t\left(\boldsymbol{x}_t\right) - f_t\left(\boldsymbol{x}^*\right) \right) \leq \sum_{t=1}^{T} \left(\left\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^* \right\rangle - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right) \\
&\leq \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t \left(1 - \beta_{1t}(1-\mu)\right)} \left(\left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^* \right) \right\|^2 - \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^* \right) \right\|^2 \right) - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right. \\
&+ \frac{\alpha_t}{\left(1 - \beta_{1t}(1-\mu)\right)} \left(\beta_1^2 \left(\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\|^2 \right) + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) \\
&\left. + \frac{(1-\mu)\beta_{1t}}{2\alpha_t \left(1 - \beta_{1t}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^* \right) \right\|^2 \right).
\end{aligned} \tag{A38}
$$

We divide the righthand side of (A38) to three parts, as

$$
\begin{aligned}
Q_1 &= \sum_{t=1}^{T} \left(\frac{1}{2\alpha_t \left(1 - \beta_{1t}(1-\mu)\right)} \left(\left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^* \right) \right\|^2 - \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_{t+1} - \boldsymbol{x}^* \right) \right\|^2 \right) - \frac{\lambda}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}^* \right\|^2 \right) \\
Q_2 &= \sum_{t=1}^{T} \frac{\alpha_t}{\left(1 - \beta_{1t}(1-\mu)\right)} \left(\beta_1^2 \left(\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\|^2 \right) + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) \\
Q_3 &= \sum_{t=1}^{T} \frac{(1-\mu)\beta_{1t}}{2\alpha_t \left(1 - \beta_{1t}(1-\mu)\right)} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^* \right) \right\|^2.
\end{aligned} \tag{A39}
$$

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) \\
&= \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{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{A40}
$$

The first inequity follows from β_t is nonincreasing, the second equity follows from $\alpha_t = \alpha/t$. The last inequity is because of the assumption $\alpha \geq \max_{i \in \{1, \cdots, d\}} \left(t\hat{\boldsymbol{v}}_{t,i}^{1/2} - (t-1)\hat{\boldsymbol{v}}_{t-1,i}^{1/2} \right) / \left(\left(1 - \beta_1(1-\mu)\right)\lambda \right)$, and $\epsilon \to 0^+$.

Then, we bound the term Q_2.

$$
\begin{aligned}
Q_2 &\leq \frac{\alpha}{1-\beta_1(1-\mu)} \sum_{t=1}^{T} \frac{1}{t} \left(\beta_1^2 \left(\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \frac{1}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_{t-1} \right\|^2 \right) + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) \\
&\leq \frac{\alpha}{1-\beta_1(1-\mu)} \left(\sum_{t=1}^{T} \frac{1}{t} \left(\beta_1^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right) + \sum_{t=1}^{T-1} \frac{1}{t+1} \frac{\beta_1^2}{2} \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 \right) \\
&\leq \frac{\alpha}{1-\beta_1(1-\mu)} \sum_{t=1}^{T} \frac{1}{t} \left(\frac{3}{2} \beta_1^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 + \mu^2 \left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 \right)
\end{aligned} \tag{A41}
$$

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

We further bound the two terms in the righthand side of (A41).

$$
\begin{aligned}
&\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{m}_t \right\|^2 \\
\leq &\sum_{i=1}^{d} \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{1(t-k+1)} \right) \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{1(t-k+1)} \boldsymbol{g}_{j,i}^2 \right) / \sqrt{(1-\beta_2) \sum_{j=1}^{t} \left(\beta_2^{t-j} \boldsymbol{g}_{j,i}^2 \right)} \\
\leq &\sum_{i=1}^{d} \left(\sum_{j=1}^{t} \beta_{1(t-k+1)}^{t-j} \right) \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{1(t-k+1)} \boldsymbol{g}_{j,i}^2 \right) / \sqrt{(1-\beta_2) \sum_{j=1}^{t} \left(\beta_2^{t-j} \boldsymbol{g}_{j,i}^2 \right)} \\
\leq &\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{1(t-k+1)} \boldsymbol{g}_{j,i}^2 \right) / \sqrt{\sum_{j=1}^{t} \left(\beta_2^{t-j} \boldsymbol{g}_{j,i}^2 \right)} \\
\leq &\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \sum_{j=1}^{t} \beta_1^{t-j} \boldsymbol{g}_{j,i}^2 / \sqrt{\beta_2^{t-j} \boldsymbol{g}_{j,i}^2} \\
= &\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}} \sum_{i=1}^{d} \sum_{j=1}^{t} \gamma^{t-j} \left| \boldsymbol{g}_{j,i} \right| \\
\leq &\frac{1}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)} G_1
\end{aligned} \tag{A42}
$$

The first inequality follows from Cauchy-Schwarz inequality.

$$
\begin{aligned}
\left\| \hat{\boldsymbol{V}}_t^{-1/4} \boldsymbol{g}_t \right\|^2 &\leq \sum_{i=1}^{d} \frac{\boldsymbol{g}_{ti}^2}{\sqrt{(1-\beta_2) \sum_{j=1}^{t} \left(\beta_2^{t-j} \boldsymbol{g}_{j,i}^2 \right)}} \\
&\leq \sum_{i=1}^{d} \frac{\boldsymbol{g}_{t,i}^2}{\sqrt{1-\beta_2} \left| \boldsymbol{g}_{t,i} \right|} \leq \frac{1}{\sqrt{1-\beta_2}} G_1.
\end{aligned} \tag{A43}
$$

Combining (A41), (A42), and (A43), we obtain

$$
\begin{aligned}
Q_2 &\leq \frac{\alpha G_1}{(1-\beta_1(1-\mu))\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{1}{t} \left(\frac{3}{2} \frac{\beta_1^2}{(1-\beta_1)(1-\gamma)} + \mu^2 \right) \\
&\leq \frac{\alpha G_1}{(1-\beta_1(1-\mu))\sqrt{1-\beta_2}} \left(\frac{3}{2} \frac{\beta_1^2}{(1-\beta_1)(1-\gamma)} + \mu^2 \right) (\log(T)+1).
\end{aligned} \tag{A44}
$$

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

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**

Finally, we bound the term Q_3.

$$
\begin{aligned}
Q_3 &\leq \frac{1}{2\left(1-\beta_1(1-\mu)\right)} \sum_{t=1}^{T} \left(\frac{(1-\mu)\beta_{1t}}{\alpha_t} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 \right) \\
&= \frac{(1-\mu)\beta_1}{2\alpha\left(1-\beta_1(1-\mu)\right)} \sum_{t=1}^{T} \left(\frac{1}{t} \left\| \hat{\boldsymbol{V}}_t^{1/4} \left(\boldsymbol{x}_t - \boldsymbol{x}^*\right) \right\|^2 \right) \\
&\leq \frac{(1-\mu)\beta_1}{2\alpha\left(1-\beta_1(1-\mu)\right)} \sum_{t=1}^{T} \frac{1}{t} \sum_{i=1}^{d} \hat{\boldsymbol{v}}_{t,i}^{1/2} \left(\boldsymbol{x}_{t,i} - \boldsymbol{x}_i^*\right)^2 \qquad (A45) \\
&\leq \frac{(1-\mu)\beta_1 D_\infty^2}{2\alpha\left(1-\beta_1(1-\mu)\right)} \sum_{t=1}^{T} \frac{1}{t} \sum_{i=1}^{d} \hat{\boldsymbol{v}}_{t,i}^{1/2} \\
&\leq \frac{(1-\mu)\beta_1 D_\infty^2}{2\alpha\left(1-\beta_1(1-\mu)\right)} (1+\log(T)) \sum_{i=1}^{d} \hat{\boldsymbol{v}}_{t,i}^{1/2}.
\end{aligned}
$$

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.

Combining (A38), (A39), (A40), (A44), and (A45), we obtain

$$
R_T \leq \left(\frac{\alpha G_1}{\sqrt{1-\beta_2}} \left(\frac{3}{2} \frac{\beta_1^2}{\left(1-\beta_1\right)\left(1-\gamma\right)} + \mu^2 \right) + \frac{(1-\mu)\beta_1 D_\infty^2}{2\alpha} \sum_{i=1}^{d} \hat{\boldsymbol{v}}_{t,i}^{1/2} \right) \frac{1+\log(T)}{1-\beta_1(1-\mu)}. \qquad (A46)
$$

4 Equivalence of RSG and ASGD

Jain et al. [2018] shows that ASGD [Kidambi et al., 2018, Jain et al., 2018] 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

$$\begin{aligned}
&\begin{bmatrix}\bar{\boldsymbol{x}}_{t+1}\\ \boldsymbol{x}_{t+1}\end{bmatrix} = \ddot{A}\begin{bmatrix}\bar{\boldsymbol{x}}_{t}\\ \boldsymbol{x}_{t}\end{bmatrix} + \ddot{b}\boldsymbol{g}_t \\
&\ddot{A} = \begin{bmatrix}\ddot{\beta} & 1-\ddot{\beta}\\ \frac{(1-\ddot{\beta})\ddot{\beta}}{(1-\ddot{\beta})+0.7} & \frac{(1-\ddot{\beta})^2+0.7}{(1-\ddot{\beta})+0.7}\end{bmatrix}, \ddot{b} = \begin{bmatrix}\frac{\ddot{\beta}-1}{0.7}\ddot{\kappa}\ddot{\alpha}\\ -\frac{0.7^2+(1-\ddot{\beta})^2\ddot{\kappa}}{0.7((1-\ddot{\beta})+0.7)}\ddot{\alpha}\end{bmatrix}.
\end{aligned} \tag{A47}$$

Define the variable transform as

$$\begin{bmatrix}\tilde{\boldsymbol{m}}_{t}\\ \boldsymbol{x}_{t}\end{bmatrix} = \ddot{T}\begin{bmatrix}\bar{\boldsymbol{x}}_{t}\\ \boldsymbol{x}_{t}\end{bmatrix}, \ddot{T} = \begin{bmatrix}\ddot{l} & \ddot{k}\ddot{l}\\ 0 & 1\end{bmatrix}, \tag{A48}$$

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

Combining (A47) and (A48), we obtain

$$\begin{bmatrix}\tilde{\boldsymbol{m}}_{t+1}\\ \boldsymbol{x}_{t+1}\end{bmatrix} = \tilde{T}\begin{bmatrix}\tilde{\boldsymbol{m}}_{t}\\ \boldsymbol{x}_{t}\end{bmatrix} + \ddot{T}\ddot{b}\boldsymbol{g}_t, \tilde{T} = \ddot{T}\ddot{A}\ddot{T}^{-1}. \tag{A49}$$

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

$$\ddot{k} = -1, \ddot{l} = \frac{(1-\ddot{\beta})\ddot{\beta}}{(1-\ddot{\beta})+0.7}. \tag{A50}$$

Combining (A49) and (A50), we obtain

$$\begin{bmatrix}\tilde{\boldsymbol{m}}_{t+1}\\ \boldsymbol{x}_{t+1}\end{bmatrix} = \tilde{T}\begin{bmatrix}\tilde{\boldsymbol{m}}_{t}\\ \boldsymbol{x}_{t}\end{bmatrix} + \ddot{T}\ddot{b}\boldsymbol{g}_t, \tilde{T} = \begin{bmatrix}\frac{0.7\ddot{\beta}}{(1-\ddot{\beta})+0.7} & 0\\ 1 & 1\end{bmatrix}. \tag{A51}$$

When the hyper-parameters are constant, the concise form of RSG (6) can be rearranged as

$$\begin{aligned}
&\boldsymbol{m}_t = \beta\boldsymbol{m}_{t-1} - \alpha(1-\beta)(1-\mu)\nabla f_t(\boldsymbol{x}_t)\\
&\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{m}_t - \alpha\mu\nabla f_t(\boldsymbol{x}_t),
\end{aligned} \tag{A52}$$

The update (A51) and (A52) are identical. The momentum coefficient of RSG is

$$\beta = \frac{0.7\ddot{\beta}}{(1-\ddot{\beta})+0.7} = (\ddot{\kappa} - 0.49\ddot{\xi})/(\ddot{\kappa} + 0.7\ddot{\xi}), \tag{A53}$$

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.

5 More details on experiments

We use 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 $\{0.0005, 0.001, 0.002, 0.005, 0.01, 0.02\} \times \{0, 0.9, 0.99, 0.999, 0.9999\} \times \{0.99, 0.999\}$. 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_{LA} and slow weights step α_{LA}. The hyper-parameters $(\alpha, \beta_1, \beta_2, k_{\text{LA}}, \alpha_{\text{LA}})$ are selected by grid search from $\{0.0005, 0.001, 0.002, 0.005, 0.01, 0.02\} \times \{0.9, 0.95, 0.99, 0.999\} \times \{0.999\} \times$

$\{5\} \times \{0.2, 0.5, 0.8\}$. For SGD, the hyper-parameters (α, β) are selected by grid search from $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0\} \times \{0, 0.9, 0.99, 0.999, 0.9999\}$. For RSG, the hyper-parameters (α, β, μ) are selected by grid search from $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0\} \times \{0, 0.9, 0.99, 0.999, 0.9999\} \times \{0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.9\}$. For ARSG and ARSGB, the hyper-parameters (α) is selected by grid search from $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$, (β_1, β_2, μ) are set according to the default values. The small positive constant ϵ is set as 10^{-8} for all the adaptive methods. In ARSGB, the grid search runs for 5 epochs in the experiments on MNIST, and 20 epochs on CIFAR10. The average convergence rate is computed each 2 epoches on MNIST, and 10 epochs on CIFAR10. α and μ are scaled when the converging rate is halved to achieve fast convergence, and at the 50th epoch (when the loss flattens) to maximize the final generalization.

Table A1 shows the hyper-parameters selected.

Table A1: The hyper-parameters in the experiments

	Experiments on MNIST	
Methods	Logistic regression	CNN
SGD	$(2.0, 0.99)$	$(1.0, 0.9)$
ADAM	$(0.005, 0.999, 0.999)$	$(0.0005, 0.9, 0.999)$
NADAM	$(0.01, 0.999, 0.999)$	$(0.0005, 0.9, 0.999)$
AMSGRAD	$(0.005, 0.99, 0.999)$	$(0.005, 0.9, 0.99)$
RANGER	$(0.02, 0.99, 0.999, 5, 0.5)$	$(0.005, 0.9, 0.999, 5, 0.2)$
RSG	$(5.0, 0.999, 0.1)$	$(5.0, 0.999, 0.1)$
ARSG	$(0.05, 0.999, 0.99, 0.1)$	$(0.01, 0.999, 0.99, 0.1)$
ARSGB	$(0.1, 0.999, 0.99, 0.05)$	$(0.05, 0.999, 0.99, 0.05)$

	Resnet-20 on CIFAR-10	
Methods	Fastest convergence	Best generalization
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	$(0.2, 0.999, 0.1)$	$(0.5, 0.99, 0.3)$
ARSG	$(0.002, 0.999, 0.99, 0.1)$	$(0.005, 0.999, 0.99, 0.2)$
ARSGB	$(0.005, 0.999, 0.99, 0.05)$	$(0.01, 0.999, 0.99, 0.05)$

The experiments are carried out on a workstation with an Intel Xeon Gold 6148 CPU and a NVIDIA V100 GPU. The source code of ARSG can be downloaded at `https://github.com/rationalspark/NAMSG/blob/master/Arsg.py`. The simulation environment is MXNET, which can be downloaded at `http://mxnet.incubator.apache.org`. The MNIST dataset can be downloaded at `http://yann.lecun.com/exdb/mnist`; the CIFAR-10 dataset can be downloaded at `http://www.cs.toronto.edu/~kriz/cifar.html`.

References

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. URL `http://proceedings.mlr.press/v75/jain18a.html`.

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. URL `https://openreview.net/forum?id=rJTutzbA-`.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. *CoRR*, abs/1002.4908, 2010. URL http://arxiv.org/abs/1002.4908.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.