Supplementary Material for
“ A Sufficient Condition for Convergences of Adam and RMSProp ”

In this supplementary we give the complete proof of our main results. Section A introduces the necessary lemmas for the proof and Section B prove the maim propositions, theorems and corollaries. Section C describes the architectures of LeNet and ResNet-18, and the statistics of the training datasets and validation datasets of MNIST and CIFAR-100.

Notations We use bold letters to represent vectors. The k-th component of a vector $\boldsymbol{v}_t$ is denoted as $v_{t,k}$. The inner product between two vectors $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ is denoted as $\langle \boldsymbol{v}_t, \boldsymbol{w}_t \rangle$. Other than that, all computations that involve vectors shall be understood in the component-wise way. We say a vector $\boldsymbol{v}_t \geq 0$ if every component of $\boldsymbol{v}_t$ is non-negative, and $\boldsymbol{v}_t \geq \boldsymbol{w}_t$ if $v_{t,k} \geq w_{t,k}$ for all $k = 1, 2, \ldots, d$. The ℓ_1 norm of a vector $\boldsymbol{v}_t$ is defined as $\|\boldsymbol{v}_t\|_1 = \sum_{k=1}^d |v_{t,k}|$. The ℓ_2 norm is defined as $\|\boldsymbol{v}_t\|^2 = \langle \boldsymbol{v}_t, \boldsymbol{v}_t \rangle = \sum_{k=1}^d |v_{t,k}|^2$. Given a positive vector $\hat{\boldsymbol{\eta}}_t$, it will be helpful defining the following weighted norm: $\|\boldsymbol{v}_t\|_{\boldsymbol{\eta}_t}^2 = \langle \boldsymbol{v}_t, \hat{\boldsymbol{\eta}}_t \boldsymbol{v}_t \rangle = \sum_{k=1}^d \hat{\eta}_{t,k} |v_{t,k}|^2$.

A. Key Lemmas

In this section we provide the necessary lemmas for the proof of the main theorem.

Lemma 15. *Given $S_0 > 0$ and a non-negative sequence $\{s_t\}$, let $S_t = S_0 + \sum_{i=1}^t s_i$ for $t \geq 1$. Then the following estimate holds*

$$\sum_{t=1}^T \frac{s_t}{S_t} \leq \log(S_T) - \log(S_0). \tag{10}$$

Proof. The finite sum $\sum_{t=1}^T s_t/S_t$ can be interpreted as a Riemann sum $\sum_{t=1}^T (S_t - S_{t-1})/S_t$. Since $1/x$ is decreasing on the interval $(0, \infty)$, we have

$$\sum_{t=1}^T \frac{S_t - S_{t-1}}{S_t} \leq \int_{S_0}^{S_T} \frac{1}{x} dx = \log(S_T) - \log(S_0).$$

The proof is finished. □

Lemma 16 (Abel's Lemma - Summation by parts)**.** *Let $\{u_t\}$ and $\{s_t\}$ be two non-negative sequences. Let $S_t = \sum_{i=1}^t s_i$ for $t \geq 1$. Then*

$$\sum_{t=1}^T u_t s_t = \sum_{t=1}^{T-1} (u_t - u_{t+1}) S_t + u_T S_T. \tag{11}$$

Proof. Let $S_0 = 0$. Then

$$\sum_{t=1}^T u_t s_t = \sum_{t=1}^T u_t (S_t - S_{t-1}) = \sum_{t=1}^{T-1} u_t S_t - \sum_{t=1}^{T-1} u_{t+1} S_t + u_T S_T = \sum_{t=1}^{T-1} (u_t - u_{t+1}) S_t + u_T S_T. \tag{12}$$

The proof is finished. □

Lemma 17. *Let $\{\theta_t\}$ and $\{\alpha_t\}$ satisfy the restrictions (R2) and (R3). For any $i \leq t$ we have*

$$\chi_t \leq C_0 \chi_i \;\; \textit{and} \;\; \alpha_t \leq C_0 \alpha_i. \tag{13}$$

Proof. For any $i \leq t$, since the sequence $\{a_t\}$ is non-increasing, we have $a_t \leq a_i$. Hence,

$$\chi_t = \frac{\alpha_t}{\sqrt{1 - \theta_t}} \leq C_0 a_t \leq C_0 a_i \leq C_0 \frac{\alpha_i}{\sqrt{1 - \theta_i}} = C_0 \chi_i.$$

This proves the first inequality. On the other hand, since $\{\theta_t\}$ is non-decreasing,

$$\alpha_t \leq C_0 \frac{\sqrt{1-\theta_t}}{\sqrt{1-\theta_i}} \alpha_i \leq C_0 \alpha_i = C_0 \alpha_i.$$

The proof is finished. □

Let $\Theta_{(t,i)} = \prod_{j=i+1}^{t} \theta_j$ for $i < t$ and $\Theta_{(t,t)} = 1$ by convention.

Lemma 18. *Fix a constant θ' with $\beta^2 < \theta' < \theta$. Let C_1 be as given as Eq. (6) in the main paper. For any $i \leq t$ we have*

$$\Theta_{(t,i)} \geq C_1 (\theta')^{t-i}. \tag{14}$$

Proof. For any $i \leq t$, Since $\theta_j \geq \theta'$ for $j \geq N$, and $\theta_j < \theta'$ for $j < N$, we have

$$\Theta_{(t,i)} = \prod_{j=i+1}^{t} \theta_j \geq \left(\prod_{j=i+1}^{N} \theta_j \right) (\theta')^{t-N} = \left(\prod_{j=i+1}^{N} (\theta_j/\theta') \right) (\theta')^{t-i} \geq \left(\prod_{j=1}^{N} (\theta_j/\theta') \right) (\theta')^{t-i}.$$

We take the constant $C_1 = \prod_{j=1}^{N} (\theta_j/\theta')$ where N is the maximum of the indices for which $\theta_j < \theta'$. The proof is finished. □

Remark 19. *If $\theta_t = \theta$ is a constant, we have $\Theta_{(t,i)} = \theta^{t-i}$. In this case we can take $\theta' = \theta$ and $C_1 = 1$.*

Lemma 20. *Let $\gamma := \beta^2/\theta'$. We have the following estimate*

$$\boldsymbol{m_t^2} \leq \frac{1}{C_1(1-\gamma)(1-\theta_t)} \boldsymbol{v}_t, \ \forall t. \tag{15}$$

Proof. Let $B_{(t,i)} = \prod_{j=i+1}^{t} \beta_j$ for $i < t$ and $B_{(t,t)} = 1$ by convention. By the iteration formula $\boldsymbol{m}_t = \beta_t \boldsymbol{m}_{t-1} + (1-\beta_t)\boldsymbol{g}_t$ and $\boldsymbol{m}_0 = \boldsymbol{0}$, we have

$$\boldsymbol{m}_t = \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \beta_j \right) (1-\beta_i) \boldsymbol{g}_i = \sum_{i=1}^{t} B_{(t,i)} (1-\beta_i) \boldsymbol{g}_i.$$

Similarly, by $\boldsymbol{v}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t^2$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$, we have

$$\boldsymbol{v}_t = \left(\prod_{j=1}^{t} \theta_j \right) \boldsymbol{\epsilon} + \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \theta_j \right) (1-\theta_i)\, \boldsymbol{g}_i^2 \geq \sum_{i=1}^{t} \Theta_{(t,i)} (1-\theta_i) \boldsymbol{g}_i^2.$$

It follows by arithmetic inequality that

$$\begin{aligned} \boldsymbol{m}_t^2 &= \left(\sum_{i=1}^{t} \frac{(1-\beta_i) B_{(t,i)}}{\sqrt{(1-\theta_i)\Theta_{(t,i)}}} \sqrt{(1-\theta_i)\Theta_{(t,i)}} \boldsymbol{g}_i \right)^2 \\ &\leq \left(\sum_{i=1}^{t} \frac{(1-\beta_i)^2 B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}} \right) \left(\sum_{i=1}^{t} \Theta_{(t,i)} (1-\theta_i) \boldsymbol{g}_i^2 \right) \leq \left(\sum_{i=1}^{t} \frac{(1-\beta_i)^2 B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}} \right) \boldsymbol{v}_t. \end{aligned}$$

Note that $\{\theta_t\}$ is non-decreasing by (**R**2), and $B_{(t,i)} \leq \beta^{t-i}$ by (**R**1). By Lemma 18, we have

$$\sum_{i=1}^{t} \frac{(1-\beta_i)^2 B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}} \leq \frac{1}{C_1(1-\theta_t)} \sum_{i=1}^{t} \left(\frac{\beta^2}{\theta'} \right)^{t-i} \leq \frac{1}{C_1(1-\theta_t)} \sum_{k=0}^{t-1} \gamma^k \leq \frac{1}{C_1(1-\gamma)(1-\theta_t)}.$$

The proof is finished. □

Let $\boldsymbol{\Delta}_t := \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\alpha_t \boldsymbol{m}_t / \sqrt{\boldsymbol{v}_t}$. Let $\hat{\boldsymbol{v}}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{\sigma_t^2}$ where $\boldsymbol{\sigma_t^2} = \mathbb{E}_t \left[\boldsymbol{g_t^2} \right]$ and let $\hat{\boldsymbol{\eta}}_{\boldsymbol{t}} = \alpha_t / \sqrt{\hat{\boldsymbol{v}}_{\boldsymbol{t}}}$.

Lemma 21. *The following equality holds*

$$\boldsymbol{\Delta}_t - \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}\boldsymbol{\Delta}_{t-1} = -(1-\beta_t)\hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t - \hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\boldsymbol{A}_t - \hat{\boldsymbol{\eta}}_t\boldsymbol{\sigma}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\boldsymbol{B}_t, \tag{16}$$

where

$$\begin{aligned}
\boldsymbol{A}_t &= \frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}} - \frac{(1-\beta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}},\\
\boldsymbol{B}_t &= \left(\frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{\sigma}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) + \frac{(1-\beta_t)\boldsymbol{\sigma}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}}.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
\boldsymbol{\Delta}_t - \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}\boldsymbol{\Delta}_{t-1} = &-\frac{\alpha_t\boldsymbol{m}_{\boldsymbol{t}}}{\sqrt{\boldsymbol{v}_t}} + \frac{\beta_t\alpha_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}} = -\alpha_t\left(\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right)\\
= &\underbrace{-\frac{(1-\beta_t)\alpha_t\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}}_{\text{(I)}}\underbrace{-\beta_t\alpha_t\boldsymbol{m}_{t-1}\left(\frac{1}{\sqrt{\boldsymbol{v}_t}} - \frac{1}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right)}_{\text{(II)}}.
\end{aligned} \tag{17}$$

For (I) we have

$$\begin{aligned}
\text{(I)} &= \frac{(1-\beta_t)\alpha_t\boldsymbol{g}_t}{\sqrt{\hat{\boldsymbol{v}}_t}} + (1-\beta_t)\alpha_t\boldsymbol{g}_t\left(\frac{1}{\sqrt{\boldsymbol{v}_t}} - \frac{1}{\sqrt{\hat{\boldsymbol{v}}_t}}\right)\\
&= (1-\beta_t)\hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t + (1-\beta_t)\alpha_t\boldsymbol{g}_t\frac{(1-\theta_t)(\boldsymbol{\sigma}_t^2-\boldsymbol{g}_t^2)}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}(\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t})}\\
&= (1-\beta_t)\hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t + \hat{\boldsymbol{\eta}}_t\boldsymbol{\sigma}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\frac{(1-\beta_t)\boldsymbol{\sigma}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}} - \hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\frac{(1-\beta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}}.
\end{aligned} \tag{18}$$

For (II) we have

$$\begin{aligned}
&\text{(II)} = \beta_t\alpha_t\boldsymbol{m}_{t-1}\frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\theta_t\boldsymbol{v}_{t-1}}(\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}})}\\
= &\ \beta_t\alpha_t\boldsymbol{m}_{t-1}\frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}(\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}})} + \beta_t\alpha_t\boldsymbol{m}_{t-1}\frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}(\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}})}\left(\frac{1}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}} - \frac{1}{\sqrt{\hat{\boldsymbol{v}}_t}}\right)\\
= &\ \hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\left(\frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) + \frac{\beta_t\alpha_t\boldsymbol{m}_{t-1}(1-\theta_t)^2\boldsymbol{g}_t^2\boldsymbol{\sigma}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}\sqrt{\theta_t\boldsymbol{v}_{t-1}}(\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}})(\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}})}\\
= &\ \hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\left(\frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) + \hat{\boldsymbol{\eta}}_t\boldsymbol{\sigma}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\left(\frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{\sigma}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right).
\end{aligned} \tag{19}$$

Combine Eq. (18) and Eq. (19), we then obtain the desired Eq. (16). The proof is finished. □

Lemma 22. *Let $M_t = \mathbb{E}\left[\langle\boldsymbol{\nabla}f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle + L\left\|\boldsymbol{\Delta}_t\right\|^2\right]$ and $\chi_t = \alpha_t/\sqrt{1-\theta_t}$. Then for any $t \geq 2$, we have*

$$M_t \leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + L\,\mathbb{E}\left[\left\|\boldsymbol{\Delta}_t\right\|^2\right] + C_2G\chi_t\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla}f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \tag{20}$$

and

$$M_1 \leq L\,\mathbb{E}\left[\left\|\boldsymbol{\Delta}_1\right\|^2\right] + C_2G\chi_1\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{v}_1}\right\|^2\right], \tag{21}$$

where $C_2 = 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2$.

Proof. First, for $t \geq 2$ we have

$$\mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle = \underbrace{\frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_{t-1} \rangle}_{\text{(I)}} + \underbrace{\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t - \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \boldsymbol{\Delta}_{t-1} \right\rangle}_{\text{(II)}}. \tag{22}$$

To estimate (I), by Schwartz inequality and the Lipschitz continuity of the gradient of f, we have

$$\begin{aligned} \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_{t-1} \rangle &\leq \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t) - \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle \\ &\leq \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \right\| \left\| \boldsymbol{\Delta}_{t-1} \right\| \\ &= \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{\Delta}_{t-1} \right\|^2. \end{aligned} \tag{23}$$

Hence, we have

$$\text{(I)} \leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{\Delta}_{t-1} \right\|^2 \right] = \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} M_{t-1}. \tag{24}$$

To estimate (II), by Lemma 21, we have

$$\begin{aligned} &\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t - \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \boldsymbol{\Delta}_{t-1} \right\rangle \\ &= -(1-\beta_t)\mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \rangle \underbrace{-\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \boldsymbol{A}_t \right\rangle}_{\text{(III)}} \underbrace{-\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{\sigma}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \boldsymbol{B}_t \right\rangle}_{\text{(IV)}}. \end{aligned} \tag{25}$$

Note that $\hat{\boldsymbol{\eta}}_t$ is independent from $\boldsymbol{g}_t$, and that $\mathbb{E}_t[\boldsymbol{g}_t] = \boldsymbol{\nabla} f(\boldsymbol{x}_t)$. Hence, for the first term in the right hand side of Eq. (25), we have

$$\begin{aligned} -(1-\beta_t)\mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \rangle &= -(1-\beta_t)\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \mathbb{E}_t[\boldsymbol{g}_t] \right\rangle \\ &= -(1-\beta_t)\mathbb{E}\left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t} \\ &\leq -(1-\beta)\mathbb{E}\left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t}. \end{aligned} \tag{26}$$

To estimate (III), we have

$$\text{(III)} \leq \mathbb{E}\left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\sigma}_t}, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\sigma}_t|\boldsymbol{A}_t|(1-\theta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle. \tag{27}$$

Note that $\boldsymbol{\sigma}_t \leq G$. Therefore,

$$\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\sigma}_t = \sqrt{\hat{\boldsymbol{\eta}}_t \boldsymbol{\sigma}_t^2} = \sqrt{\frac{\alpha_t \boldsymbol{\sigma}_t^2}{\sqrt{\hat{\boldsymbol{v}}_t}}} \leq \sqrt{\frac{\alpha_t \boldsymbol{\sigma}_t^2}{\sqrt{(1-\theta_t)\boldsymbol{\sigma}_t^2}}} \leq \sqrt{\frac{G\alpha_t}{\sqrt{1-\theta_t}}} = \sqrt{G\chi_t}. \tag{28}$$

On the other hand,

$$|\boldsymbol{A}_t| = \left| \frac{\beta_t \boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} - \frac{(1-\beta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t} + \sqrt{\hat{\boldsymbol{v}}_t}} \right| \leq \frac{\beta_t |\boldsymbol{m}_{t-1}|}{\sqrt{\theta_t \boldsymbol{v}_{t-1}}} + \frac{(1-\beta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}}. \tag{29}$$

By Lemma 20, we have

$$\frac{|\boldsymbol{m}_{t-1}|}{\sqrt{\boldsymbol{v}_{t-1}}} \leq \frac{1}{\sqrt{C_1(1-\gamma)(1-\theta_t)}}. \tag{30}$$

Meanwhile,

$$\frac{|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \leq \frac{|\boldsymbol{g}_t|}{\sqrt{(1-\theta_t)\boldsymbol{g}_t^2}} = \frac{1}{\sqrt{1-\theta_t}}. \tag{31}$$

Hence, we have

$$\begin{aligned} |\boldsymbol{A}_t| &\leq \frac{\beta_t}{\sqrt{C_1(1-\gamma)(1-\theta_t)\theta_t}} + \frac{1-\beta_t}{\sqrt{1-\theta_t}} \leq \left(\frac{\beta_t/(1-\beta_t)}{\sqrt{C_1(1-\gamma)\theta_t}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} \\ &\leq \left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} := \frac{C_2'(1-\beta_t)}{\sqrt{1-\theta_t}}, \end{aligned} \tag{32}$$

where $C_2' = \left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)$. The last inequality is due to $\beta_t/(1-\beta_t) \leq \beta/(1-\beta)$ as $\beta_t \leq \beta$. Therefore, we have

$$\begin{aligned} &\left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\sigma}_t}, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\sigma}_t|\boldsymbol{A}_t|(1-\theta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\sigma}_t}, \sqrt{G\chi_t}C_2'(1-\beta_t)\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\frac{1-\beta_t}{4}\left\|\frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\sigma}_t}\right\|^2 + C_2'^2 G(1-\beta_t)\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 \\ \leq\ &\frac{1-\beta_t}{4}\left\|\frac{\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2|\boldsymbol{g}_t|^2}{\boldsymbol{\sigma}_t^2}\right\|_1 + C_2'^2 G\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \end{aligned} \tag{33}$$

Note that $\boldsymbol{\sigma}_t^2 = \mathbb{E}_t[\boldsymbol{g}_t^2]$. Hence,

$$\mathbb{E}_t\left\|\frac{\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2|\boldsymbol{g}_t|^2}{\boldsymbol{\sigma}_t^2}\right\|_1 = \left\|\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2\right\|_1 = \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2. \tag{34}$$

Combine Eq. (27), Eq. (33) and Eq. (34), we get

$$\text{(III)} \leq \frac{1-\beta_t}{4}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] + C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \tag{35}$$

The term (IV) is estimated similarly as term (III). First, we have

$$\begin{aligned} |\boldsymbol{B}_t| &\leq \left(\frac{\beta_t|\boldsymbol{m}_{t-1}|}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{\sigma}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) + \frac{(1-\beta_t)\boldsymbol{\sigma}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}} \\ &\leq \left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} = \frac{C_2'(1-\beta_t)}{\sqrt{1-\theta_t}}, \end{aligned} \tag{36}$$

where C_2' is the constant defined above. We have

$$\begin{aligned} \text{(IV)} \leq\ &\mathbb{E}\left\langle \sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\sigma}_t|\boldsymbol{B}_t|(1-\theta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\mathbb{E}\left\langle \sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|, \sqrt{G\chi_t}C_2'(1-\beta_t)\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\frac{1-\beta_t}{4}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] + C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \end{aligned} \tag{37}$$

Combine Eq. (22), Eq. (23), Eq. (25), Eq. (26), Eq. (35) and Eq. (37), we obtain that

$$\begin{aligned} \mathbb{E}\langle\boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle &\leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + 2C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 - \frac{1-\beta_t}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \\ &\leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + 2C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]. \end{aligned} \tag{38}$$

Let C_2 denote the constant $2(C_2')^2$. Then

$$C_2 = 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2.$$

We obtain Eq. (20) by adding the term $L\mathbb{E}\left[\left\|\boldsymbol{\Delta}_t\right\|^2\right]$ to both sides of Eq. (38).

When $t = 1$, we have

$$M_1 = \mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1 \boldsymbol{m}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle + L\left\|\boldsymbol{\Delta}_1\right\|^2\right] = \mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1(1-\beta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle + L\left\|\boldsymbol{\Delta}_1\right\|^2\right]. \tag{39}$$

The same as what we did for term (I) in Lemma 21, we have

$$\frac{(1-\beta_1)\alpha_1\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_t}} = (1-\beta_1)\hat{\boldsymbol{\eta}}_1\boldsymbol{g}_1 + \hat{\boldsymbol{\eta}}_1\boldsymbol{\sigma}_1\frac{(1-\theta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\frac{(1-\beta_1)\boldsymbol{\sigma}_1}{\sqrt{\boldsymbol{v}_1}+\sqrt{\hat{\boldsymbol{v}}_1}} - \hat{\boldsymbol{\eta}}_1\boldsymbol{g}_1\frac{(1-\theta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\frac{(1-\beta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}+\sqrt{\hat{\boldsymbol{v}}_1}}. \tag{40}$$

Then the similar argument as Eq. (33) implies that

$$\begin{aligned}\mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1\boldsymbol{m}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle\right] &\le C_2 G\chi_1\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\|^2\right] - \frac{1-\beta_1}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_1)\right\|^2_{\hat{\boldsymbol{\eta}}_1}\right] \\ &\le C_2 G\chi_1\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\|^2\right].\end{aligned} \tag{41}$$

Combine Eq. (39) and Eq. (41), and adding both sides by $L\mathbb{E}\left[\left\|\boldsymbol{\Delta}\right\|_1^2\right]$, we obtain Eq. (21). This finishes the proof. □

Lemma 23. *The following estimate holds*

$$\sum_{t=1}^{T}\left\|\boldsymbol{\Delta}_t\right\|^2 \le \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})^2}\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \tag{42}$$

Proof. Note that $\boldsymbol{v}_t \ge \theta_t\boldsymbol{v}_{t-1}$, hence we have $\boldsymbol{v}_t \ge \left(\prod_{j=i+1}^{t}\theta_j\right)\boldsymbol{v}_i = \Theta_{(t,i)}\boldsymbol{v}_i$. By Lemma 18, this follows that $\boldsymbol{v}_t \ge C_1(\theta')^{t-i}\boldsymbol{v}_i$ for all $i \le t$. On the other hand,

$$|\boldsymbol{m}_t| \le \sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}\beta_j\right)(1-\beta_i)|\boldsymbol{g}_i| \le \sum_{i=1}^{t}\beta^{t-i}|\boldsymbol{g}_i|.$$

It follows that

$$\frac{|\boldsymbol{m}_t|}{\sqrt{\boldsymbol{v}_t}} \le \sum_{i=1}^{t}\frac{\beta^{t-i}|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_t}} \le \frac{1}{\sqrt{C_1}}\sum_{i=1}^{t}\left(\frac{\beta}{\sqrt{\theta'}}\right)^{t-i}\frac{|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}} = \frac{1}{\sqrt{C_1}}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\frac{|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}}. \tag{43}$$

Since $\alpha_t = \chi_t\sqrt{1-\theta_t} \le \chi_t\sqrt{1-\theta_i}$ for $i \le t$, it follows that

$$\begin{aligned}\left\|\boldsymbol{\Delta}_t\right\|^2 = \left\|\frac{\alpha_t\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 &\le \frac{\chi_t^2}{C_1}\left\|\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\frac{\sqrt{1-\theta_i}|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \le \frac{\chi_t^2}{C_1}\left(\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\right)\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\ &\le \frac{\chi_t^2}{C_1(1-\sqrt{\gamma})}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2.\end{aligned} \tag{44}$$

By Lemma 17,

$$\chi_t \le C_0\chi_i, \forall i \le t.$$

Hence,

$$\left\|\boldsymbol{\Delta}_t\right\|^2 = \left\|\frac{\alpha_t \boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 \leq \frac{C_0^2 \chi_1}{C_1(1-\sqrt{\gamma})} \sum_{i=1}^{t} \sqrt{\gamma}^{t-i} \chi_i \left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2. \tag{45}$$

It follows that

$$\begin{aligned}
\sum_{t=1}^{T} \left\|\boldsymbol{\Delta}_t\right\|^2 \leq\ & \frac{C_0^2 \chi_1}{C_1(1-\sqrt{\gamma})} \sum_{t=1}^{T} \sum_{i=1}^{t} \sqrt{\gamma}^{t-i} \chi_i \left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\
=\ & \frac{C_0^2 \chi_1}{C_1(1-\sqrt{\gamma})} \sum_{i=1}^{T} \left(\sum_{t=i}^{T} \sqrt{\gamma}^{t-i}\right) \chi_i \left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\
\leq\ & \frac{C_0^2 \chi_1}{C_1(1-\sqrt{\gamma})^2} \sum_{i=1}^{T} \chi_i \left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2.
\end{aligned} \tag{46}$$

The proof is finished. □

Lemma 24. *Let* $M_t = \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle + L\left\|\boldsymbol{\Delta}_t\right\|^2\right]$. *For* $T \geq 1$ *we have*

$$\sum_{t=1}^{T} M_t \leq C_3 \mathbb{E}\left[\sum_{t=1}^{T} \chi_t \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] - \frac{1-\beta}{2} \mathbb{E}\left[\sum_{t=1}^{T} \left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right]. \tag{47}$$

where the constant C_3 *is given by*

$$C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2 G\right).$$

Proof. Let $N_t = L\mathbb{E}\left[\left\|\boldsymbol{\Delta}_t\right\|^2\right] + C_2 G \chi_t \mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right]$. By Lemma 22, we have $M_1 \leq N_1$ and

$$M_t \leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} M_{t-1} + N_t - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} M_{t-1} + N_t. \tag{48}$$

It is straightforward to get by induction that

$$\begin{aligned}
M_t &\leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \frac{\beta_{t-1}\alpha_{t-1}}{\sqrt{\theta_{t-1}}\alpha_{t-2}} M_{t-2} + \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} N_{t-1} + N_t - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \\
&\ \ \vdots \\
&\leq \frac{\alpha_t B_{(t,1)}}{\alpha_1\sqrt{\Theta_{(t,1)}}} M_1 + \sum_{i=2}^{t} \frac{\alpha_t B_{(t,i)}}{\alpha_i\sqrt{\Theta_{(t,i)}}} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \\
&\leq \sum_{i=1}^{t} \frac{\alpha_t B_{(t,i)}}{\alpha_i\sqrt{\Theta_{(t,i)}}} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right].
\end{aligned} \tag{49}$$

By Lemma 17, $\alpha_t \leq C_0\alpha_i$ for any $i \leq t$. By Lemma 18, $\Theta_{(t,i)} \geq C_1(\theta')^{t-i}$. In addition, $B_{(t,i)} \leq \beta^{t-i}$. Hence,

$$\begin{aligned}
M_t \leq\ & \frac{C_0}{\sqrt{C_1}} \sum_{i=1}^{t} \left(\frac{\beta}{\sqrt{\theta'}}\right)^{t-i} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right] \\
=\ & \frac{C_0}{\sqrt{C_1}} \sum_{i=1}^{t} \sqrt{\gamma}^{t-i} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\left\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\right\|_{\hat{\boldsymbol{\eta}}_t}^2\right].
\end{aligned} \tag{50}$$

Hence,

$$\begin{aligned}\sum_{t=1}^{T} M_t &\le \frac{C_0}{\sqrt{C_1}}\sum_{t=1}^{T}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}N_i - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]\\ &= \frac{C_0}{\sqrt{C_1}}\sum_{i=1}^{T}\left(\sum_{t=i}^{T}\sqrt{\gamma}^{t-i}\right)N_i - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]\\ &= \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\sum_{t=1}^{T}N_t - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right].\end{aligned} \tag{51}$$

Finally, by Lemma 23, we have

$$\begin{aligned}\sum_{t=1}^{T} N_i &= \mathbb{E}\left[L\sum_{t=1}^{T}\|\boldsymbol{\Delta}_t\|^2 + C_2 G\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right]\\ &\le \left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2}+C_2 G\right)\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right].\end{aligned} \tag{52}$$

Let

$$\begin{aligned}C_3 &= \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2}+C_2 G\right)\\ &= \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2}+2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}}+1\right)^2 G\right).\end{aligned}$$

Combine Eq. (51) and Eq. (52), we then obtain the desired estimate Eq. (47). The proof is finished. □

Lemma 25. *The following estimate holds*

$$\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \le d\left[\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{i=1}^{t}\log(\theta_i^{-1})\right]. \tag{53}$$

Proof. Let $W_0 = 1$ and $W_t = \prod_{i=1}^{T}\theta_i^{-1}$. Let $w_t = W_t - W_{t-1} = (1-\theta_t)\prod_{i=1}^{t}\theta_i^{-1} = (1-\theta_t)W_t$. We therefore have

$$\frac{w_t}{W_t} = 1-\theta_t, \quad \frac{W_{t-1}}{W_t} = \theta_t.$$

Note that $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$ and $\boldsymbol{v}_t = \theta_t\boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t$, hence $W_0\boldsymbol{v}_0 = \boldsymbol{\epsilon}$ and $W_t\boldsymbol{v}_t = W_{t-1}\boldsymbol{v}_{t-1} + w_t\boldsymbol{g}_t^2$. Hence, $W_t\boldsymbol{v}_t = W_0\boldsymbol{v}_0 + \sum_{i=1}^{t} w_i\boldsymbol{g}_i^2 = \boldsymbol{\epsilon} + \sum_{i=1}^{t} w_i\boldsymbol{g}_i^2$. It follows that

$$\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 = \sum_{i=1}^{t}\left\|\frac{(1-\theta_i)\boldsymbol{g}_t^2}{\boldsymbol{v}_i}\right\|_1 = \sum_{i=1}^{t}\left\|\frac{w_i\boldsymbol{g}_i^2}{W_i\boldsymbol{v}_i}\right\|_1 = \sum_{i=1}^{t}\left\|\frac{w_i\boldsymbol{g}_i^2}{\boldsymbol{\epsilon}+\sum_{\ell=1}^{i} w_\ell\boldsymbol{g}_\ell^2}\right\|_1. \tag{54}$$

Write the norm in terms of coordinates, we get

$$\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 = \sum_{i=1}^{t}\sum_{k=1}^{d}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2} = \sum_{k=1}^{d}\sum_{i=1}^{t}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2}. \tag{55}$$

By Lemma 18, for each $k = 1, 2, \dots, d$,

$$\sum_{i=1}^{t}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2} \le \log\left(\epsilon+\sum_{\ell=1}^{t} w_\ell g_{\ell,k}^2\right) - \log(\epsilon) = \log\left(1+\frac{1}{\epsilon}\sum_{\ell=1}^{t} w_\ell g_{\ell,k}^2\right). \tag{56}$$

Hence,

$$\begin{aligned}\sum_{i=1}^{t} \left\| \frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}} \right\|^2 \leq\ & \sum_{k=1}^{d} \log\left(1 + \frac{1}{\epsilon}\sum_{i=1}^{t} w_i g_{i,k}^2\right) \\ \leq\ & d\log\left(\frac{1}{d}\sum_{k=1}^{d}\left(1 + \frac{1}{\epsilon}\sum_{i=1}^{t} w_i g_{i,k}^2\right)\right) = d\log\left(1 + \frac{1}{\epsilon d}\sum_{i=1}^{t} w_i \left\|\boldsymbol{g}_i\right\|^2\right).\end{aligned} \tag{57}$$

The second inequality is due to the convex inequality $\frac{1}{d}\sum_{k=1}^{d}\log(z_i) \leq \log\left(\frac{1}{d}\sum_{k=1}^{d} z_i\right)$. Indeed, we have the more general convex inequality that

$$\mathbb{E}[\log(X)] \leq \log\mathbb{E}[X] \tag{58}$$

for any positive random variable X. Taking X to be $1 + \frac{1}{\epsilon d}\sum_{i=1}^{t} w_i \left\|\boldsymbol{g}_i\right\|^2$ in the right hand side of Eq. (57), we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \leq d\,\mathbb{E}\left[\log\left(1 + \frac{1}{\epsilon d}\sum_{i=1}^{t} w_i\left\|\boldsymbol{g}_i\right\|^2\right)\right] \leq d\log\left(1 + \frac{1}{\epsilon d}\sum_{i=1}^{t} w_i \mathbb{E}\left[\left\|\boldsymbol{g}_i\right\|^2\right]\right) \\ \leq\ & d\log\left(1 + \frac{G^2}{\epsilon d}\sum_{i=1}^{t} w_i\right) = d\log\left(1 + \frac{G^2}{\epsilon d}(W_t - W_0)\right) = d\log\left(1 + \frac{G^2}{\epsilon d}\left(\prod_{i=1}^{t}\theta_i^{-1} - 1\right)\right) \\ \leq\ & d\left[\log\left(1 + \frac{G^2}{\epsilon d}\right) + \log\left(\prod_{i=1}^{t}\theta_i^{-1}\right)\right].\end{aligned} \tag{59}$$

The last inequality is due to the following trivial inequality

$$\log(1+ab) \leq \log(1+a+b+ab) = \log(1+a) + \log(1+b)$$

for non-negative a and b. It then follows that

$$\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \leq d\left[\log\left(1 + \frac{G^2}{\epsilon d}\right) + \sum_{i=1}^{t}\log(\theta_i^{-1})\right]. \tag{60}$$

The proof is finished. □

Lemma 26. *We have the following estimate*

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \leq C_0 d\left[\chi_1\log\left(1 + \frac{G^2}{\epsilon d}\right) + \frac{1}{\theta_1}\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right]. \tag{61}$$

Proof. For simplicity of notations, let $\omega_t := \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2$, and $\Omega_t := \sum_{i=1}^{t}\omega_i$. Note that $\chi_t \leq C_0 a_t$. Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \leq C_0\,\mathbb{E}\left[\sum_{t=1}^{T}a_t\omega_t\right]. \tag{62}$$

By Lemma 16, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}a_t\omega_t\right] = \mathbb{E}\left[\sum_{t=1}^{T-1}(a_t - a_{t+1})\Omega_t + a_T\Omega_T\right] \tag{63}$$

Let $S_t := \log\left(1 + \frac{G^2}{\epsilon d}\right) + \sum_{i=1}^{t}\log(\theta_i^{-1})$. By Lemma 25, we have

$$\mathbb{E}[\Omega_t] \leq dS_t. \tag{64}$$

Since $\{a_t\}$ is a non-increasing sequence, we have $a_t - a_{t+1} \geq 0$. By Eq. (63), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T-1}(a_t - a_{t+1})\Omega_t + a_T\Omega_T\right] \leq d\left(\sum_{t=1}^{T-1}(a_t - a_{t+1})S_t + a_T S_T\right) \\
=&d\left(a_1 S_0 + \sum_{t=1}^{T} a_t(S_t - S_{t-1})\right) = d\left[a_1 \log\left(1 + \frac{G^2}{\epsilon d}\right) + \sum_{t=1}^{T} a_t \log(\theta_t^{-1})\right]
\end{aligned} \tag{65}$$

Note that $a_t \leq \chi_t$. Combining Eq. (62), Eq. (63) and Eq. (65), we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \leq\ & C_0 d\left[\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right) + \sum_{t=1}^{T}\chi_t \log(\theta_t^{-1})\right] \\
=\ & C_0 d\left[\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right) + \sum_{t=1}^{T}\chi_t \log(\theta_t^{-1})\right].
\end{aligned} \tag{66}$$

Note that $\log(1+x) \leq x$ for all $x > -1$, it follows that

$$\log(\theta_t^{-1}) = \log(1 + (\theta_t^{-1} - 1)) \leq \theta_t^{-1} - 1 \leq \frac{1-\theta_t}{\theta_1}.$$

Note that $\chi_t = \alpha_t/\sqrt{1-\theta_t}$. Hence, by Eq. (62) and Eq. (65), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \leq C_0 d\left[\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right) - \frac{1}{\theta_1}\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right]. \tag{67}$$

The proof is finished. □

Lemma 27. *Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. We have the following estimate*

$$\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^{4/3}\right]\right)^{3/2} \leq \frac{C_0\sqrt{G^2 + \epsilon d}}{T\alpha_T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right]. \tag{68}$$

Proof. For any two random variables X and Y, by the Hölder's inequality, we have

$$\mathbb{E}[|XY|] \leq \mathbb{E}\left[|X|^p\right]^{1/p}\mathbb{E}\left[|Y|^q\right]^{1/q}. \tag{69}$$

Let $X = \left(\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}}\right)^{2/3}$, $Y = \|\hat{\boldsymbol{v}}_t\|_1^{1/3}$, and let $p = 3/2$, $q = 3$. By Eq. (69), we have

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^{4/3}\right] \leq \mathbb{E}\left[\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}}\right]^{2/3}\mathbb{E}\left[\|\hat{\boldsymbol{v}}_t\|_1\right]^{1/3}. \tag{70}$$

On the one hand, we have

$$\begin{aligned}
\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}} = \sum_{k=1}^{d}\frac{|\nabla_k f(\boldsymbol{x}_t)|^2}{\sqrt{\sum_{k=1}^{d}\hat{v}_{t,k}}} \leq\ & \alpha_t^{-1}\sum_{k=1}^{d}\frac{\alpha_t}{\sqrt{\hat{v}_{t,k}}}|\nabla_k f(\boldsymbol{x}_t)|^2 \\
=\ & \alpha_t^{-1}\sum_{k=1}^{d}\hat{\eta}_{t,k}|\nabla_k f(\boldsymbol{x}_t)|^2 = \alpha_t^{-1}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}.
\end{aligned} \tag{71}$$

On the other hand, since $\hat{\boldsymbol{v}}_t = \theta_t\boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{\sigma}_t^2$, and all entries are non-negative, we have

$$\|\hat{\boldsymbol{v}}_t\|_1 = \theta_t\|\boldsymbol{v}_{t-1}\|_1 + (1-\theta_t)\|\boldsymbol{\sigma}_t\|^2.$$

Notice that $\boldsymbol{v}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t^2$, and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$ and $\mathbb{E}_t\left[\boldsymbol{g}_t^2\right] \leq G^2$, it is straightforward to prove by induction that $\mathbb{E}[\|\boldsymbol{v}_t\|_1] \leq G^2 + \epsilon d$. Hence,

$$\mathbb{E}[\|\hat{\boldsymbol{v}}_t\|_1] \leq G^2 + \epsilon d. \tag{72}$$

By Eq. (70), Eq. (71) and Eq. (72), we obtain

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^{4/3}\right] \leq \left(\alpha_t^{-1}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right]\right)^{2/3}(G^2+\epsilon d)^{1/3}. \tag{73}$$

By Lemma 17, $\alpha_T \leq C_0\alpha_t$ for any $t \leq T$, hence $\alpha_t^{-1} \leq C_0\alpha_T^{-1}$. Hence,

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^{4/3}\right]^{3/2} \leq \frac{C_0\sqrt{G^2+\epsilon d}}{\alpha_T}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right], \ \forall t \leq T. \tag{74}$$

The lemma is followed by

$$\begin{aligned}\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^{4/3}\right]\right)^{3/2} &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^{4/3}\right]\right)^{3/2} \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^{4/3}\right]\right)^{3/2} \leq \frac{C_0\sqrt{G^2+\epsilon d}}{T\alpha_T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right].\end{aligned} \tag{75}$$

The proof is finished. □

B. Proof of the main results

In this section, we provide the detailed proof of propositions, theorems and corollaries in the main body.

B.1. Proof of Proposition 3

Proposition. *Algorithm 1 and Algorithm 2 are equivalent.*

Proof. It suffices to show that Algorithm 1 can be realized as Algorithm 2 with a particular choice of parameters, and vice versa. Note that for Algorithm 1, it holds that

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\alpha_t \boldsymbol{m}_t}{\sqrt{\left(\prod_{i=1}^{t}\theta_i\right)\boldsymbol{\epsilon} + \sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}\theta_j(1-\theta_i)\right)\boldsymbol{g}_i^2}}. \tag{76}$$

While for Algorithm 2, we have

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\alpha_t \boldsymbol{m}_t}{\sqrt{\frac{1}{W_t}\boldsymbol{\epsilon} + \sum_{i=1}^{t}\frac{w_i}{W_t}\boldsymbol{g}_t^2}}. \tag{77}$$

Hence, given the parameters θ_t in Algorithm 1, we take $w_t = (1-\theta_t)\prod_{i=1}^{t}\theta_i^{-1}$. Then it holds that

$$W_t = 1 + \sum_{i=1}^{t} w_i = \prod_{i=1}^{t}\theta_i^{-1}.$$

It follows that Eq. (77) becomes Eq. (76). Conversely, given the parameters w_t of Algorithm 2, we take $\theta_t = W_{t-1}/W_t$. Then Eq. (76) becomes Eq. (77). The proof is completed. □

B.2. Proof of Theorem 4

Theorem. *Let $\{\boldsymbol{x}_t\}$ be a sequence generated by Generic Adam for initial values $\boldsymbol{x}_1$, $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$. Assume that f and stochastic gradients $\boldsymbol{g}_t$ satisfy assumptions (A1)-(A4). Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. We have the following estimate*

$$\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^{4/3}\right]\right)^{3/2} \leq \frac{C + C' \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t}}{T\alpha_T}, \tag{78}$$

where the constant C and C' are given by

$$\begin{aligned} C &= \frac{2C_0\sqrt{G^2+\epsilon d}}{1-\beta}\left(f(x_1) - f^* + C_3 C_0 d\,\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right)\right), \\ C' &= \frac{2C_0^2 C_3 d\sqrt{G^2+\epsilon d}}{(1-\beta)\theta_1}. \end{aligned}$$

Proof. By the L-Lipschitz continuity of the gradient of f and the descent lemma, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + \frac{L}{2}\|\boldsymbol{\Delta}_t\|^2. \tag{79}$$

Let $M_t := \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L\|\boldsymbol{\Delta}_t\|^2\right]$. In particular, we have $\mathbb{E}[f(\boldsymbol{x}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{x}_t)] + M_t$. Taking sum for $t = 1, 2, \ldots, T$, we obtain that

$$\mathbb{E}\left[f(\boldsymbol{x}_{T+1})\right] \leq f(\boldsymbol{x}_1) + \sum_{t=1}^{T} M_t. \tag{80}$$

Note that $f(x)$ is bounded from below by f^*, hence, $\mathbb{E}[f(\boldsymbol{x}_{T+1})] \geq f^*$. Applying the estimate of Lemma 24, we have

$$f^* \leq f(\boldsymbol{x}_1) + C_3 \mathbb{E}\left[\sum_{t=1}^{T} \chi_t \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right], \tag{81}$$

where constant C_3 is the constant given in Lemma 24. It follows by applying the estimates in Lemma 25 and Lemma 27 for the second and third terms in the right hand side of Eq. (81), and appropriately rearranging the terms. Then we get

$$\begin{aligned} &\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau^T)\|^{4/3}\right]\right)^{3/2} \leq \frac{C_0\sqrt{G^2+\epsilon d}}{T\alpha_T}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\ \leq\ & \frac{2C_0\sqrt{G^2+\epsilon d}}{(1-\beta)T\alpha_T}\left(f(\boldsymbol{x}_1) - f^* + C_3\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|\right]\right) \\ \leq\ & \frac{2C_0\sqrt{G^2+\epsilon d}}{(1-\beta)T\alpha_T}\left[f(\boldsymbol{x}_1) - f^* + C_3 C_0 d\,\chi_1 \log\left(1+\frac{G^2}{\epsilon d}\right) - \frac{C_3 C_0 d}{\theta_1}\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right] \\ =\ & \frac{C + C'\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}}{T\alpha_T} \end{aligned} \tag{82}$$

where

$$\begin{aligned} C &= \frac{2C_0\sqrt{G^2+\epsilon d}}{1-\beta}\left(f(x_1) - f^* + C_3 C_0 d\,\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right)\right) \\ C' &= \frac{2C_0^2 C_3 d\sqrt{G^2+\epsilon d}}{(1-\beta)\theta_1}. \end{aligned}$$

The proof is finished. □

B.3. Proof of Main Theorem 5

In this section we give a complete proof of the main Theorem 5. For readers' convenience we restate the theorem here.

Theorem. *Let $\{\boldsymbol{x}_t\}$ be a sequence generated by Generic Adam for initial values $\boldsymbol{x}_1$, $\boldsymbol{m}_0 = \mathbf{0}$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$. Assume that f and stochastic gradients $\boldsymbol{g}_t$ satisfy assumptions (A1)-(A4). Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. Then for any $\delta > 0$, the following bound holds with probability at least $1 - \delta^{2/3}$:*

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^2 \leq \frac{C + C' \sum_{t=1}^T \alpha_t \sqrt{1-\theta_t}}{\delta T \alpha_T} := Bound(T), \tag{83}$$

where the constants C and C' are given by

$$C = \frac{2C_0\sqrt{G^2+\epsilon d}}{1-\beta}\left(f(x_1) - f^* + C_3 C_0 d\, \chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right)\right),$$
$$C' = \frac{2C_0^2 C_3 d\sqrt{G^2+\epsilon d}}{(1-\beta)\theta_1},$$

in which the constant C_3 is given by

$$C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2 G\right).$$

Proof of the Theorem. Denote the right hand side of Eq. (78) as $C(T)$. Let $\zeta = \|\nabla f(x_\tau)\|^2$. By Theorem 4 we have $\mathbb{E}\left[|\zeta|^{2/3}\right] \leq C(T)^{2/3}$. Let $\mathcal{P}$ denote the probability measure. By Chebyshev's inequality, we have

$$\mathcal{P}\left(|\zeta|^{2/3} > \frac{C(T)^{2/3}}{\delta^{2/3}}\right) \leq \frac{\mathbb{E}\left[|\zeta|^{2/3}\right]}{\frac{C(T)^{2/3}}{\delta^{2/3}}} \leq \delta^{2/3}. \tag{84}$$

Namely, $\mathcal{P}\left(|\zeta| > \frac{C(T)}{\delta}\right) \leq \delta^{2/3}$. Therefore, $\mathcal{P}\left(|\zeta| \leq \frac{C(T)}{\delta}\right) \geq 1 - \delta^{2/3}$. This finishes the proof. □

B.4. Proof of Corollary 7

Corollary. *Take $\alpha_t = \eta/t^s$ with $0 \leq s < 1$. Suppose $\lim_{t\to\infty} \theta_t = \theta < 1$, then the $Bound(T)$ in Theorem 5 is bounded from below by constants*

$$Bound(T) \geq \frac{C'\sqrt{1-\theta}}{\delta}. \tag{85}$$

In particular, when $\theta_t = \theta < 1$, we have the following more subtle estimate on lower and upper-bounds for $Bound(T)$

$$\frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta} \leq Bound(T) \leq \frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta(1-s)}.$$

Proof. Since $\lim_{t\to\infty} \theta_t = \theta$, and θ_t is non-decreasing, we have $(1-\theta_t) \geq 1 - \theta$. Hence, by Theorem 5, it holds

$$\begin{aligned} Bound(T) &\geq \frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta}\left(\frac{\sum_{t=1}^T t^{-s}}{T^{1-s}}\right) \\ &\geq \frac{C'\sqrt{1-\theta}}{\delta}. \end{aligned} \tag{86}$$

If in particular, $\theta_t = \theta < 1$, then by Theorem 5 we have

$$Bound(T) = \frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta}\left(\frac{\sum_{t=1}^T t^{-s}}{T^{1-s}}\right). \tag{87}$$

Note that

$$1 \leq \frac{\sum_{t=1}^T t^{-s}}{T^{1-s}} = \sum_{t=1}^T \left(\frac{t}{T}\right)^{-s} \frac{1}{T} \leq \int_0^1 x^{-s} dx = \frac{1}{1-s}. \tag{88}$$

Combining Eqs. (87)-(88), we obtain the desired result. □

B.5. Proof of Corollary 10

Corollary. *Generic Adam with the above family of parameters converges as long as* $0 < r \leq 2s < 2$*, and its non-asymptotic convergence rate is given by*

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^2 \leq \begin{cases} \mathcal{O}(T^{-r/2}), & r/2+s<1 \\ \mathcal{O}(\log(T)/T^{1-s}), & r/2+s=1 \\ \mathcal{O}(1/T^{1-s}), & r/2+s>1 \end{cases}.$$

Proof. It is not hard to verify that the following equalities hold:

$$\begin{aligned} \textstyle\sum_{t=K}^T \alpha_t \sqrt{1-\theta_t} &= \eta\sqrt{\alpha} \textstyle\sum_{t=K}^T t^{-(r/2+s)} \\ &= \begin{cases} \mathcal{O}(T^{1-(r/2+s)}), & r/2+s<1 \\ \mathcal{O}(\log(T)), & r/2+s=1 \\ \mathcal{O}(1), & r/2+s>1 \end{cases}. \end{aligned}$$

In this case, $T\alpha_T = \eta T^{1-s}$. Therefore, by Theorem 5 the non-asymptotic convergence rate is given by

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^2 \leq \begin{cases} \mathcal{O}(T^{-r/2}), & r/2+s<1 \\ \mathcal{O}(\log(T)/T^{1-s}), & r/2+s=1 \\ \mathcal{O}(1/T^{1-s}), & r/2+s>1 \end{cases}.$$

To guarantee convergence, then $0 < r \leq 2s < 2$. □

B.6. Proof of Corollary 12

Corollary. *Suppose in Weighted AdaEMA the weights* $w_t = t^r$ *for* $r \geq 0$*, and* $\alpha_t = \eta/\sqrt{t}$*. Then Weighted AdaEMA has the* $\mathcal{O}(\log(T)/\sqrt{T})$ *non-asymptotic convergence rate.*

Proof. By the proof procedures of Theorem 3, the equivalent Generic Adam has the parameters $\theta_t = W_{t-1}/W_t$, where $W_t = 1 + \sum_{i=1}^t w_i$. Hence, it holds that

$$1 - \theta_t = \frac{w_t}{W_t} = \frac{t^r}{1+\sum_{i=1}^t i^r} = \mathcal{O}(1/t).$$

We have $\lim_{t\to\infty} \theta_t = 1 > \beta$ and θ_t is increasing. In addition, we have that $\chi_t = \alpha_t/\sqrt{1-\theta_t}$ is bounded, and hence "almost" non-increasing (by taking $a_t = 1$ in (R3)). The restrictions (R1)-(R3) are all satisfied. Hence, we can apply Theorem 5 in this case. It follows that its convergence rate is given by

$$\mathcal{O}\big(\frac{\sum_{i=1}^T \alpha_t\sqrt{1-\theta_t}}{T\alpha_T}\big) = \mathcal{O}\big(\frac{\sum_{t=1}^T 1/t}{\sqrt{T}}\big) = \mathcal{O}\big(\frac{\log(T)}{\sqrt{T}}\big).$$

The proof is completed. □

C. Experimental Implementation

In this section, we describe the statistics of the training and validation datasets of MNIST[3] and CIFAR-100[4], the architectures of LeNet and ResNet-18, and detailed implementations.

C.1. Datasets

MNIST [17] is composed with ten classes of digits among $\{0, 1, 2, \ldots, 9\}$, which includes 60000 training examples and 10000 validation examples. The dimension of each example is 28×28.

CIFAR-100 [17] is composed with 100 classes of 32×32 color images. Each class includes 6000 images. In addition, these images are devided into 50000 training examples and 10000 validation examples.

[3] http://yann.lecun.com/exdb/mnist/

[4] https://www.cs.toronto.edu/ kriz/cifar.html

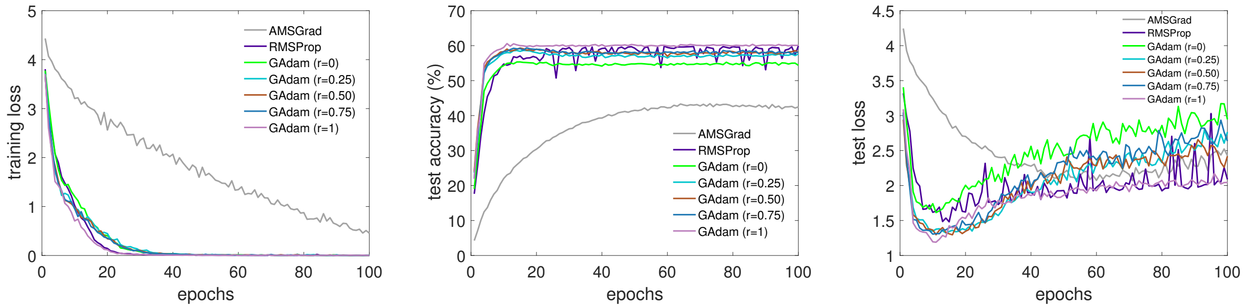
C.2. Architectures of Neural Networks

LetNet [16] used in the experiments is a five-layer convolutional neural network with ReLU activation function whose detailed architecture is described in [16]. The batch size is set as 64. The training stage lasts for 100 epochs in total. No $L2$ regularization on the weights is used.

ResNet-18 [10] is a ResNet model containing 18 convolutional layers for CIFAR-100 classification [10]. Input images are down-scaled to $1/8$ of their original sizes after the 18 convolutional layers, and then fed into a fully-connected layer for the 100-class classification. The output channel numbers of 1-3 conv layers, 4-8 conv layers, 9-13 conv layers and 14-18 conv layers are 64, 128, 256 and 512, respectively. The batch size is 64. The training stage lasts for 100 epochs in total. No $L2$ regularization on the weights is used.

C.3. Additional Experiments of ResNet-18 on CIFAR-100

We further illustrate Generic Adam with different $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad with an alternative base learning rate $\alpha = 0.01$ on ResNet-18. We do cut-off by taking $\alpha_t = 0.001$ if $t < 2500$. Note that α_t is still non-increasing. The motivation is that at the very beginning the learning rate $\alpha_t = \frac{0.01}{\sqrt{t}}$ could be large which would deteriorate the performance. The performance profiles are also exactly in accordance with the analysis in theory, *i.e.*, larger r leads to a faster training process.



(a)

(b)

(c)

Figure 4. Performance profiles of Generic Adam with $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad for training ResNet on the CIFAR-100 dataset. Figures (a), (b), and (c) illustrate training loss vs. epochs, test accuracy vs. epochs, and test loss vs. epochs, respectively.