arXiv:2404.01714v3 [cs.LG] 11 May 2024

Conjugate-Gradient-like Based Adaptive Moment Estimation Optimization Algorithm for Deep Learning

Jiawu Tian[a,c], Liwei Xu[a,b], Xiaowei Zhang[a,c], Yongqi Li[a]

[a] *School of Mathematical Sciences, University of Electronic Science and Technology of China.*
[b] *Corresponding author.*
[c] *These authors contributed equally to this work.*

Abstract

Training deep neural networks is a challenging task. In order to speed up training and enhance the performance of deep neural networks, we rectify the vanilla conjugate gradient as conjugate-gradient-like and incorporate it into the generic Adam, and thus propose a new optimization algorithm named CG-like-Adam for deep learning. Specifically, both the first-order and the second-order moment estimation of generic Adam are replaced by the conjugate-gradient-like. Convergence analysis handles the cases where the exponential moving average coefficient of the first-order moment estimation is constant and the first-order moment estimation is unbiased. Numerical experiments show the superiority of the proposed algorithm based on the CIFAR10/100 dataset.

Keywords: Deep Learning, Optimization, Conjugate Gradient, Adaptive Moment Estimation

1. Introduction

Deep learning has been used in many aspects, such as recommendation systems [1], natural language processing [2], image recognition [3], reinforcement learning [4], etc. Neural network model is the main research object of deep learning, which includes input layer, hidden layer and output layer. Each layer includes a certain number of neurons, and each neuron is connected with each other in a certain way. The parameters and connection parameters of each neuron determine the performance of the deep learning model. How to optimize these huge number of parameters affects the performance of the deep learning model, attracting researchers to devote their energy to exploration [5].

Stochastic Gradient Descent(for short, SGD) has dominated training of deep neural networks despite it was proposed in the last century. It updates parameters of deep neural network toward the negative gradient direction which would be scaled by a constant called learning rate. Simple as it is but may encounter non-convergence. More than one kind of improvement has been put forward, including momentum [6, 7] and Nesterov's acceleration [8], to accelerate the training process and upgrade optimization. Large number of parameters sharing the same learning rate may be inappropriate since some parameters can be very close to the optimal, which needs to adjust learning rate at that situation. AdaGrad [9] was the practicer of the idea of adaptive

Email addresses: jiawu2015@gmail.com (Jiawu Tian), xul@uestc.edu.cn (Liwei Xu), x.w.zhang@126.com (Xiaowei Zhang), yongqi2023@126.com (Yongqi Li)

learning rate. It scales every coordinate of the gradient by the square roots of sum of the squared past gradients. Since then various of improved optimization algorithm combining the strength of momentum and adaptive learning rate emerge in an endless stream. Some of the most popular ones are AdaDelta [10], RMSProp [11], Adam [12], NAdam [13], etc.

Adam is the most popular one among them as a result of its fast convergence, good performance as well as generality in most of practical applications. Although it has achieved wonderful results in many deep learning tasks, non-convergence issues trouble Adam in some other cases. Reddi etc. constructed several examples to disprove convergence of Adam [14], and proposed AMSGrad fixing the problem in proof of convergence of Adam algorithm given in reference [12]. However, AMSGrad's theoretical proof only handles the case where the objective function is convex, and can not ensure convergence when the exponential moving average coefficient of first-order moment estimation is a constant. Zhou etc. tackled non-convex convergence of adaptive gradient methods [15], but the "constant" case is an open problem until the work of Chen etc. [16]. Unfortunately, when the first-order moment estimation is unbiased, further work needs to be put into the proof. In a word, there is still room for improvement in for Adam-type optimization algorithm.

Conjugate Gradient method was proposed by Hestenes in the 1950s [17] and was generalized to the non-linear optimization by Fletcher and Reeves in 1964 [18]. It is very suitable for solving large-scale unconstrained non-convex optimization problems. The method iteratively moves the current parameters toward negative conjugate gradient direction which can be computed through the current gradient and the previous conjugate gradient multiplied by a conjugate coefficient. Specifically, the conjugate coefficient could be calculated by efficient formulae such as Fletcher-Reeves(FR) [18], Polak-Ribiere-Polyak(PRP) [19, 20], Hestenes-Stiefel(HS) [17], Dai-Yuan(DY) [21] as well as Hager-Zhang(HZ) [22].

Directly replacing the gradient of Adam-type with vanilla conjugate gradient will not bring the satisfactory results, i.e., non-convergence, which was demonstrated in our experiments and the reference [23] whose author incorporated conjugate gradient into optimization algorithm for deep learning. Our work absorbs the idea of conjugate-gradient-like(CG-like) in reference [23, 24]. To be more specific, the conjugate coefficient is scaled by a positive real monotonic decreasing sequence depending on the number of iterations.(see Alg.2 for more details) Such rectification leads to conjugate-gradient-like direction, and we prove the convergence (Th.3.2) of the proposed algorithm for the non-convex. Unlike the work of the reference [24], the conjugate-gradient-like is also used as second-order moment estimation. Experiments show that the proposed algorithm(Alg.2) performs better than CoBA [24].

There are two main perspectives to prove convergence: Regret Bound and Stationary Point. Convergence analysis for convex case usually adopts Regret Bound and Stationary Point for non-convex. Our convergence analysis follows the train of thought in reference [16] and extends the theorem to the case of first-order moment unbiased estimation, obtaining more general convergence theorem. Several experiments were implemented to validate the effectiveness and commendable performance of our optimization algorithm for deep learning.

To sum up, the main contributions can be summarized as follows:

(i) The vanilla conjugate gradient direction is rectified by means of conjugate coefficient multiplied by a positive real monotonic decreasing sequence, which leads to conjugate-gradient-like direction that is incorporated into Adam-type optimization algorithm for deep learning in order to speed up training process, boost performance and enhance generalization ability of deep neural networks. The algorithm, named CG-like-Adam now, combines the advantages of conjugate gradient for solving large-scaled unconstrained non-convex optimization problems and adaptive moment estimation.

(ii) The convergence for non-convex case is analyzed. The convergence analysis tackles two hard situations: a constant coefficient of exponential moving average of first-order moment estimation, and unbiased first-order moment estimation. A great deal of work is based on the objective function convex and yet they cannot deal with the constant coefficient. Our proofs generalize the convergence theorem, making it more practical.

(iii) The numerical experiments both on CIFAR-10 and CIFAR-100 dataset using popular ResNet-34 and VGG-19 network for image classification task are done. Experiment results not only testify the effectiveness of CG-like-Adam but also provide satisfactory performance especially VGG-19 network on CIFAR-10/100 dataset.

2. Preliminaries

Here some necessary knowledge are prepared for better understanding.

2.1. Notation

$\|\cdot\|_2$ which is defined as ℓ_2-norm is replaced by $\|\cdot\|$ for convenience. $\langle\cdot,\cdot\rangle$ denotes inner product. $[d]$, $\mathcal{T}$, S_+^d are both set and denotes $\{1,2,\ldots,d\}$, $\{1,2,\ldots,T\}$, $\{V|V\in\mathbb{R}^{d\times d}, V\succ 0\}$ that is the set of all positive definite $d\times d$ matrices, respectively. For any vector $x_t\in\mathbb{R}^d$, $x_{t,i}$ denotes its i^{th} coordinate where $i\in[d]$ and $\hat{V}_t^{-\frac{1}{2}}x_t$ represents $(\hat{V}_t^{\frac{1}{2}})^{-1}x_t$. Besides, $\sqrt{x_t}$ or $x_t^{\frac{1}{2}}$ represents for element-wise square root, x_t^2 for element-wise square, x_t/y_t or $\frac{x_t}{y_t}$ for element-wise division, $\max(x_t,y_t)$ for element-wise maximum, where $y_t\in\mathbb{R}^d$. $diag(v)\in\mathbb{R}^{d\times d}$ is used to denote a diagonal matrix, in which the diagonal elements d_{ii} are v_i and the other elements are 0, $i\in[d]$. Finally, $O(\cdot), o(\cdot), \Omega(\cdot)$ are used as standard asymptotic notations.

2.2. Stochastic Optimization, Generic Adam and Stationary Point

Stochastic Optimization For any deep learning or machine learning model, it can be analysed by stochastic optimization framework. In the first place, consider the problem in the following form:

$$\min_{x\in\mathcal{X}}\mathbb{E}_\pi[\mathcal{L}(x,\pi)]+\sigma(x), \tag{1}$$

where $\mathcal{X}\subseteq\mathbb{R}^d$ is feasible set. π is a random variable with an unknown distribution, representing randomly selected data sample or random noise. $\sigma(x)$ is a regular term. For any given x, $\mathcal{L}(x,\pi)$ usually represents the loss function on sample π. But for most practical cases, the distribution of π can not be obtained. Hence the expectation $\mathbb{E}_\pi$ can not be computed. Now there is another one to be considered, known as Empirical Risk Minimization Problem(ERM):

$$\min_{x\in\mathcal{X}} f(x)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(x)+\sigma(x), \tag{2}$$

where $\mathcal{L}_i(x)=\mathcal{L}(x,\pi_i)$, $i=1,2,\ldots,N$ and π_i are samples. For the convenience of our discussion below, without loss of generality, the regular term $\sigma(x)$ is ignored.

Generic Adam There are many stochastic optimization algorithms to solve the ERM (see Eq.(2)), such as SGD, AdaDelta[10], RMSProp[11], Adam[12], NAdam[13],etc. All the algorithms listed above are first-order optimization algorithm and can be described by the following Generic Adam paradigm(Alg.1), where $\varphi_t : \mathcal{X} \to S_+^d$, an unspecified "averaging" function, is usually applied to inversely weight the first-order moment estimation. The first-order moment estimation of Generic Adam is biased. Our study will focus on the Generic Adam, yet we will rectify all the moment estimation as unbiased and are going to incorporate conjugate gradient into the Generic Adam as well as prove the convergence based on that.

Algorithm 1 Generic Adam

Require: $x_1 \in \mathcal{X}$, $m_0 := 0$, $(\beta_{1t})_{t\in\mathcal{T}} \subset [0, 1)$.
for $t = 1$ to T **do**
  g_t : noisy gradient
  $m_t := \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$
  $V_t := \varphi_t(g_1, g_2, \dots, g_t)$
  $x_{t+1} := x_t - \frac{\alpha_t}{\sqrt{V_t + \epsilon I}} \cdot m_t$
end for

Stationary Point For any differentiable function $f(x)$, x^* is called a stationary point when $\|\nabla f(x^*)\|^2 = 0$, where ∇f denotes the gradient of f. If f is convex, $x^* \in \mathcal{X}$ will be a global minimizer of f over $\mathcal{X}$. However, f is usually non-convex in practice. When the solution obtained by an optimization algorithm is a stationary point, the gradient of f is almost zero near the solution, resulting in the optimization process appearing stagnant. Although long runs can jump out of this local optimal solution, the time cost can be unacceptable.

2.3. Vanilla Conjugate Gradient

The vanilla conjugate gradient method for solving unconstrained nonlinear optimization problems has been studied for ages[17–22]. It generates update direction d_t by the following manner:

$$d_t := g_t - \gamma_t \cdot d_{t-1}, \tag{3}$$

where γ_t is called conjugate coefficient and can be calculated directly by the following manner:

$$\gamma_t^{\mathrm{HS}} := \frac{\langle g_t, y_t \rangle}{\langle d_{t-1}, y_t \rangle}, \tag{4}$$

$$\gamma_t^{\mathrm{FR}} := \frac{\|g_t\|^2}{\|g_{t-1}\|^2}, \tag{5}$$

$$\gamma_t^{\mathrm{PRP}} := \frac{\langle g_t, y_t \rangle}{\|g_{t-1}\|^2}, \tag{6}$$

$$\gamma_t^{\mathrm{DY}} := \frac{\|g_t\|^2}{\langle d_{t-1}, y_t \rangle}, \tag{7}$$

$$\gamma_t^{\mathrm{HZ}} := \frac{\langle g_t, y_t \rangle}{\langle d_{t-1}, y_t \rangle} - \lambda \frac{\|y_t\|^2}{\langle d_{t-1}, y_t \rangle^2} \langle g_t, d_{t-1} \rangle, \tag{8}$$

where $y_t := g_t - g_{t-1}, \lambda > \frac{1}{4}$. All the above was proposed by Hestenes-Stiefel[17], Fletcher-Reeves[18], Polak-Ribiere-Polyak[19, 20], Dai-Yuan[21], Hanger-Zhang[22], respectively.

3. CG-like-Adam

3.1. Proposed Algorithm

First-order moment estimation In Generic Adam, $m_t := \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$, the biased estimation of the gradient of f, is the update direction. The conjugate gradient is desired to be update direction, so that $m_t := \beta_{1t} m_{t-1} + (1 - \beta_{1t}) d_t$. But this is not over yet. If just use Eq.(3) as the calculation of the conjugate gradient, then the conclusion of algorithm divergence has been verified through both our experiment and the reference [23]. The vanilla conjugate gradient should be modified as conjugate-gradient-like:

$$d_t := g_t - \frac{\gamma_t}{t^a} \cdot d_{t-1}, \tag{9}$$

where $a \in \left[\frac{1}{2}, +\infty\right)$. What needs to emphasize is that our CG-like is different from CoBA[24] in which a very small constant M was used, and therefore the update direction of CoBA is almost the noisy gradient g_t of f.

The update direction m_t should be unbiased, so the following amendment is made and it is taken as the update direction of our algorithm:

$$\hat{m}_t := \frac{m_t}{1 - \beta_{11}^t}. \tag{10}$$

Such amendment will bring us considerable trouble to the convergence proof, but we have managed to solve it.

Second-order moment estimation Theoretically, there are many ways to instantiate φ_t in Generic Adam, but the commonly used one is the exponential moving average of the square of the past gradient of f until the current step t:

$$v_t := \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{11}$$

$$V_t := diag(v_t). \tag{12}$$

This momentum adaptively adjusts the learning rate α_t. In other words, the update stepsize of each dimension of the solution toward the $-\hat{m}_t$ direction will be different. Since CG-like is used in our algorithm, the second-order moment estimation should be as follows and also unbiased:

$$v_t := \beta_2 v_{t-1} + (1 - \beta_2) d_t^2, \tag{13}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}. \tag{14}$$

In order to further ensure the algorithm convergence, the maximum of all $\hat{v}_t$ until the current time step t is also maintained:

$$\hat{v}_t := max\{\hat{v}_{t-1}, \hat{v}_t\}, \tag{15}$$

$$\hat{V}_t := diag(\hat{v}_t). \tag{16}$$

Finally we get the CG-like-Adam(Alg.2).

Algorithm 2 CG-like-Adam Algorithm

Require: $x_1 \in \mathcal{X}$, $(\beta_{1t})_{t\in\mathcal{T}} \subset [0,1)$, $\beta_2 \in (0,1)$, $a \in \left[\frac{1}{2},+\infty\right)$, $\gamma_1 = 0$, $\epsilon > 0$.
$m_0 := 0$, $v_0 := 0$, $d_0 := 0$.
for $t = 1$ to T **do**
  g_t : noisy gradient
  $d_t := g_t - \frac{\gamma_t}{t^a} \cdot d_{t-1}$
  $m_t := \beta_{1t} m_{t-1} + (1-\beta_{1t}) d_t$
  $\hat{m}_t := \frac{m_t}{1-\beta_{11}^t}$
  $v_t := \beta_2 v_{t-1} + (1-\beta_2) d_t^2$
  $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
  $\hat{v}_t \triangleq max\left(\hat{v}_t, \hat{v}_{t-1}\right)$
  $\hat{V}_t := diag(\hat{v}_t)$
  $x_{t+1} := x_t - \frac{\alpha_t}{\sqrt{\hat{V}_t + \epsilon I}} \cdot \hat{m}_t$
end for

3.2. Assumptions and Convergence Analysis

Based on the demanding of our convergence analysis, the assumptions are directly listed as follows.

A 3.1. *f is differentiable and has L-Lipschitz gradient: $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x-y\|$ holds for all $x, y \in \mathcal{X}$.*

A 3.2. *f is lower bounded: $f(x^*) > -\infty$, where x^* is an optimal solution.*

A 3.3. *The noisy gradient g_t is unbiased and the noise is independent: $g_t = \nabla f(x_t) + \zeta_t$, $\mathbb{E}[\zeta_t] = 0$, and ζ_i is independent of ζ_j if $i \neq j$.*

A 3.4. *There exists a constant H, for all $t \in \mathcal{T}$, $\|\nabla f(x_t)\| \leqslant H$, $\|g_t\| \leqslant H$.*

Theorem 3.1. *Suppose that the assumptions A3.1-A3.4 are satisfied. $\beta_{1t} \in [0,1)$, $\beta_{1t} \leq \beta_{1(t+1)}$, $\beta_{1(t+1)} \leq \beta_{1t} h(t)$ (or $\beta_{1t} h(t) \leq \beta_{1(t+1)}$) hold for all $t \in \mathcal{T}$, in which $h(t) = \frac{(1-\beta_{11}^{t-1})(1-\beta_{11}^{t+1})}{(1-\beta_{11}^t)^2}$. And $\exists G \in \mathbb{R}^+, \forall t \in \mathcal{T}, \|\alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{m}_t\| \leq G$. Then the CG-like-Adam(Alg.2) satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(x_t\right), \alpha_t \hat{V}_t^{-\frac{1}{2}} \nabla f(x_t)\right\rangle\right] \\
&\leq C_1 \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] \\
&\quad + C_2 \mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t \hat{V}_t^{-\frac{1}{2}} d_t\right\|^2\right] + C_3 + C_4 \sum_{t=1}^{T}\frac{\alpha_t\left|\gamma_t\right|}{t^a},
\end{aligned}
\tag{17}
$$

where C_1,C_2,C_3 and C_4 are both constant independent of T, $\mu_t = \frac{\alpha_t(1-\beta_{1t})}{\xi_t}$, $\xi_t = \left(1-\beta_{11}^t\right) - \beta_{1t}\left(1-\beta_{11}^{t-1}\right)$. The expectation $\mathbb{E}$ is taken with respect to all the randomness corresponding to g_t.

Proof. See Appendix Appendix B. □

Theorem 3.2. *Suppose that the assumptions A3.1-A3.4 and the conditions of Th.3.1 are satisfied. When $T \to +\infty$, there is*

$$\min_{t\in\mathcal{T}} \left[\mathbb{E}\left\|\nabla f\left(x_t\right)\right\|^2\right] = O\left(\frac{S_1(T)}{S_2(T)}\right), \tag{18}$$

where when $T \to +\infty$,

$$\begin{aligned} O(S_1(T)) = C_1\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t\hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] \\ + C_2\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + C_3 + C_4\sum_{t=1}^{T}\frac{\alpha_t\left|\gamma_t\right|}{t^a}, \end{aligned} \tag{19}$$

$\Omega\left(S_2(T)\right) = \sum_{t=1}^{T}\tau_t$ in which $\tau_t := \underset{k\in[d]}{min}\left\{\underset{\{g_i\}_{i=1}^{t}}{min}\frac{\alpha_t}{\sqrt{\hat{v}_{t,k}}}\right\}$ denotes the minimum possible value of effective stepsize at time t over all possible coordinate and the past noisy gradients $\{g_i\}_{i=1}^{t}$. And $S_1(T)$, $S_2(T)$ are functions that are all unrelated to random variables.

Proof. See Appendix Appendix C. □

The theorem 3.2 is the direct result of theorem 3.1, which provides a sufficient condition that guarantees convergence of our CG-like-Adam algorithm: the Right Hand Side(RHS) of Eq.(17) increases much slower than the sum of the minimum possible value of effective stepsize as $T \to +\infty$. Equivalently, when $T \to +\infty$, $S_2(T)$ grows slower than $S_1(T)$: $S_2(T) = o(S_1(T))$.

The conclusions of theorem 3.1 and 3.2 are similar to those of [16], but our theorem is extended to the case where the moment estimation is unbiased, which makes the conclusion more universal.

If the learning rate α_t is specified as $\frac{\alpha}{t^b}$ and β_{1t} is a constant for all $t \in \mathcal{T}$, then the following corollary describes the convergence rate of CG-like-Adam(Alg.2):

Corollary 3.1. *Suppose that the assumptions A3.1-A3.4 and the conditions of Th.3.1 are satisfied. $\forall t \in \mathcal{T}, \beta_{1t} = \beta_{1(t+1)}, \alpha_t = \frac{\alpha}{t^b}, b \in [\frac{1}{2}, 1)$, and $\exists c \in \mathbb{R}^+, \forall i \in [d], |g_{1,i}| \geqslant c$, then the following holds:*

$$\min_{t\in\mathcal{T}} \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leqslant \frac{Q_1}{T^{1-b}}\left(Q_2 \ln T + Q_3\right), \tag{20}$$

where Q_1, Q_2 and Q_3 are both constant independent of T.

Proof. See Appendix Appendix D. □

The corollary 3.1 indicates that the best convergence rate of CG-like-Adam(Alg.2) with the learning rate in the form of $\alpha_t = \frac{\alpha}{t^b}$ will be $\frac{\ln T}{\sqrt{T}}$ when $b = \frac{1}{2}$. The derived best convergence rate still falls short of the fastest rate of first-order methods($\frac{1}{\sqrt{T}}$) due to an additional factor $\log T$. $\alpha_t = \alpha(\forall t \in \mathcal{T})$ is usually adopted in practice to mitigate the slowdown[16], yet it is still an open problem of convergence rate analysis in this case.

In our theoretical analysis, the positive definite matrix εI in the algorithm did not be taken into account for the reason of convenience. Practical implementations may require adding εI for numerical stability. However, that does not lead to the loss of generality of our analysis since εI can be very easy converted and incorporated into $\hat{v}_t$. Therefore, the assumption $|g_{1,i}| \geqslant c$ $(\forall i \in [d])$ in the corollary 3.1 is a mild assumption which means it could be easy to hold in practice.

4. Experiments

In this section, we conduct several experiments of image classification task on the benchmark dataset CIFAR-10[25] and CIFAR-100[25] using popular network VGG-19[26] and ResNet-34[27] to numerically compare CG-like-Adam(Alg.2) with CoBA[24] and Adam[12]. Specifically, VGG-19 was trained on CIFAR-10 and ResNet-34 was trained on CIFAR-100.

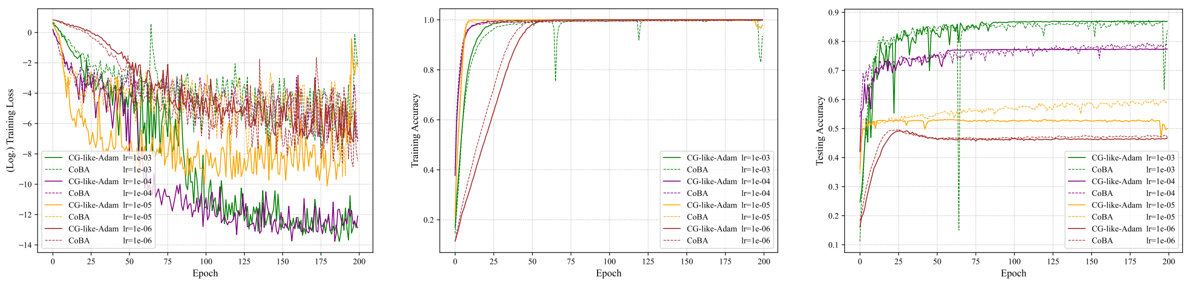
The CIFAR-10 and CIFAR-100 dataset both consist of 60000 32×32 colour images with 10 and 100 classes respectively. Both two datasets are divided into training and testing datasets, which includes 50000 training images and 10000 test images respectively. The test batch contains exactly 100 or 1000 randomly-selected images from each class.

VGG-19, proposed by Karen Simonyan[26] in 2014, consisting of 16 convolution layers and 3 fully-connected layers, is a concise network structure made up of 5 vgg-blocks with consecutive 3×3 convolutional kernels in each vgg-block. The final fully-connected layer is 1000-way with a softmax function. The ResNet-34 network, proposed by He[27], incorporates residual unit through a short-cut connection, enhancing the network's learning ability. ResNet-34 is organized as a 7×7 convolution layer, four convolution-blocks including total 32 convolution layers with 3×3 convolutional kernels and finally a 1000-way-fully-connected layer with a softmax function. VGG-19 and ResNet-34 have a strong ability to extract image features. All networks was trained for 200 epochs on NVIDIA Tesla V100 GPU.

$\beta_{1t} = 0.9(\forall t \in \mathcal{T})$ and $\beta_2 = 0.999$ are set as default values for all optimizers, Adam, CoBA and CG-like-Adam(ours, Alg.2). $\lambda = 2$ of HZ(see Eq.(8)) and $a = 1 + 10^{-5}$ are set as default values for CoBA and CG-like-Adam. $M = 10^{-4}$ is the default value for CoBA. The cross entropy is used as the loss function for all experiments for the reason of commonly used strategy in image classification. Besides, batch size is set 512 as default.

4.1. Compare CG-like-Adam with CoBA

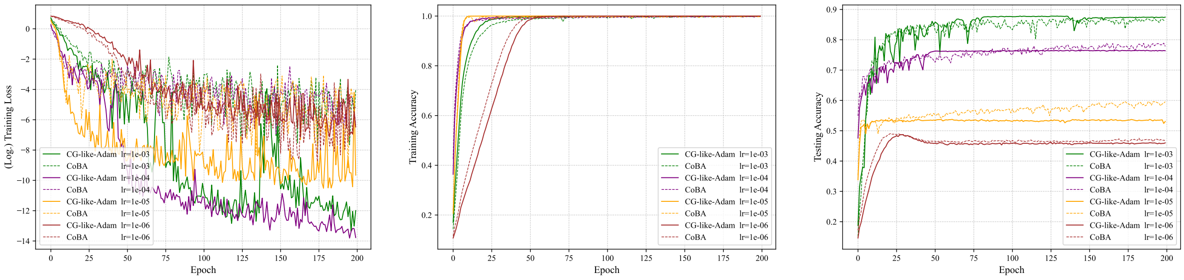
We firstly compare CG-like-Adam with CoBA. All the algorithms are run under different learning rates $\alpha_t \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}, \forall t \in \mathcal{T}$. Figure 1-5, 6-10 show the results of the experiments of VGG-19 on CIFAR-10, and ResNet-34 on CIFAR-100, respectively.



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

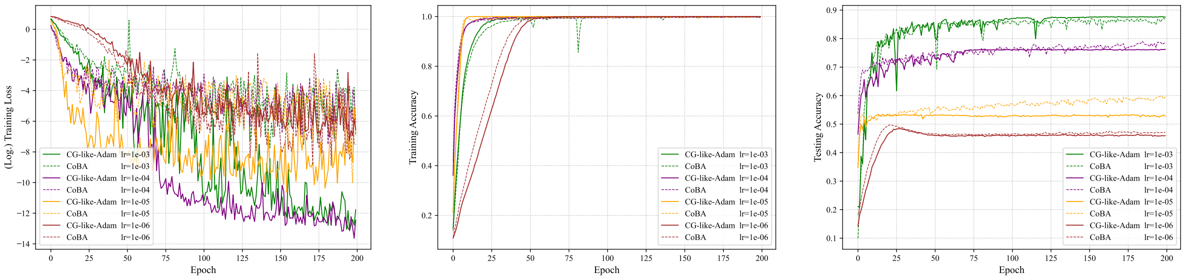
Figure 1: CG-like-Adam V.S. CoBA under different learning rates. (VGG-19, CIFAR-10, HS(4))

From figure 1-5, as you can see, no matter what type of conjugate coefficient calculation method is employed, CG-like-Adam keeps the training loss of VGG-19 to the minimum, except the learning rate $\alpha_t = 10^{-6}$. The training accuracy finally hits 100% through the optimization of CG-like-Adam and CoBA, however, CG-like-Adam achieves this goal faster and more stable. More importantly, our algorithm performs better than CoBA on test dataset when the learning rate $\alpha_t = 10^{-3}$,



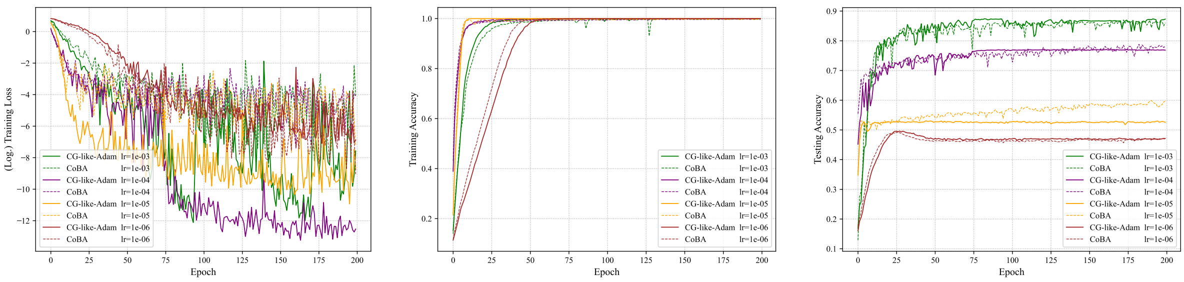
(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 2: CG-like-Adam V.S. CoBA under different learning rates. (VGG-19, CIFAR-10, FR(5))



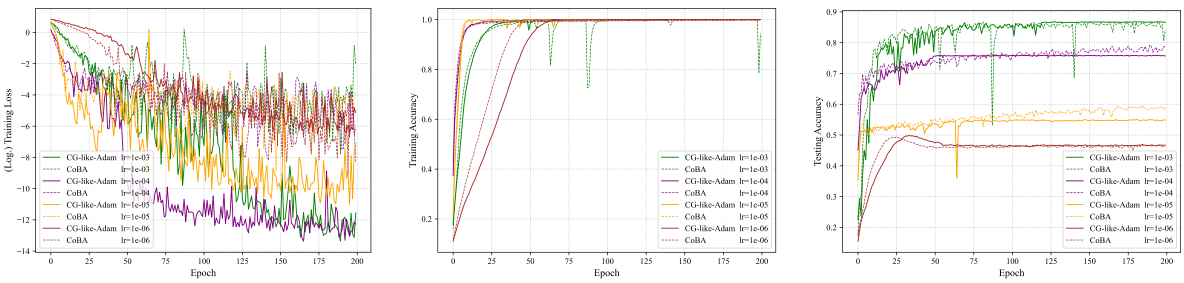
(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 3: CG-like-Adam V.S. CoBA under different learning rates. (VGG-19, CIFAR-10, PRP(6))



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

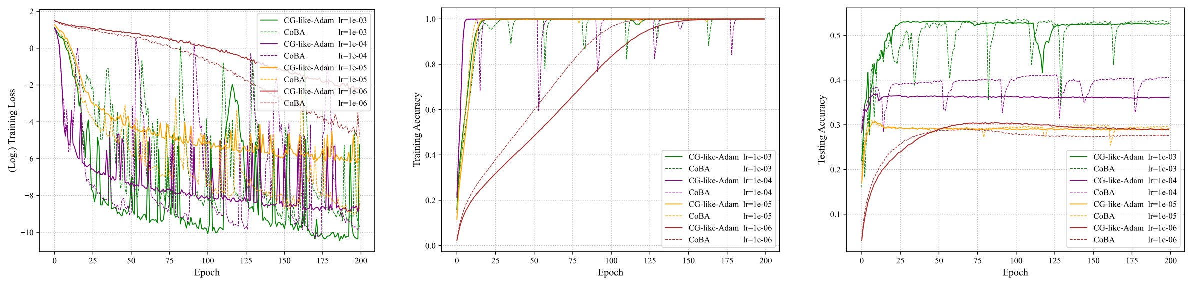
Figure 4: CG-like-Adam V.S. CoBA under different learning rates. (VGG-19, CIFAR-10, DY(7))



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

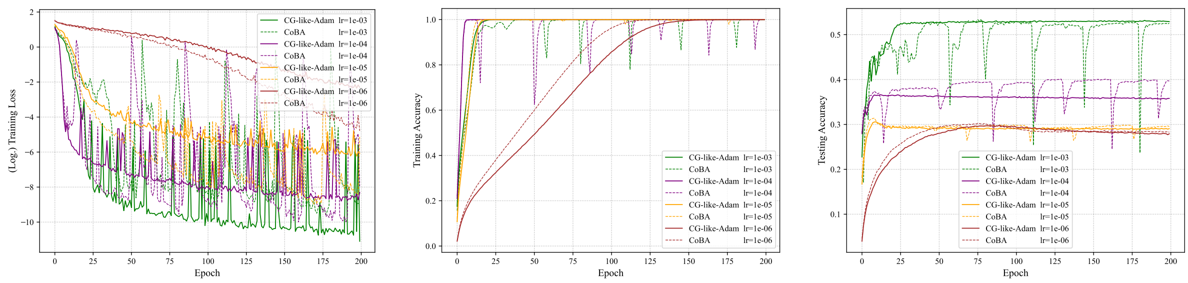
Figure 5: CG-like-Adam V.S. CoBA under different learning rates. (VGG-19, CIFAR-10, HZ(8))

although at other learning rates, the testing accuracy of CG-like-Adam is similar to or a little worse than that of CoBA.



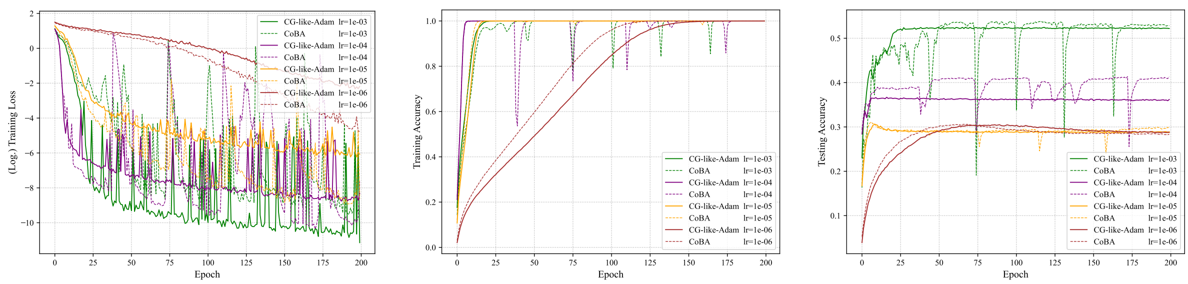
(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 6: CG-like-Adam V.S. CoBA under different learning rates. (ResNet-34, CIFAR-100, HS(4))



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

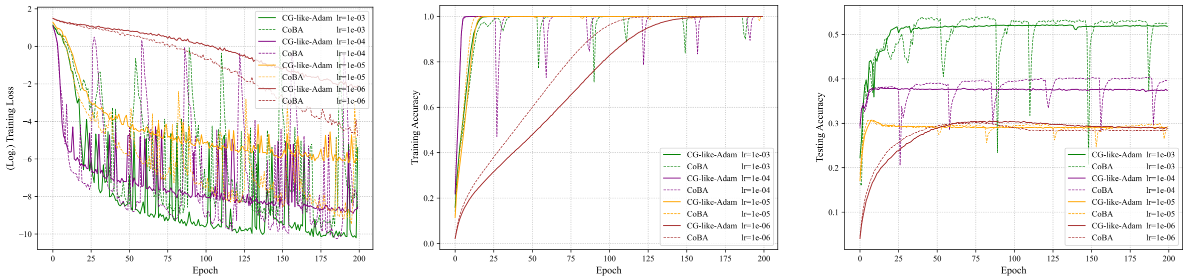
Figure 7: CG-like-Adam V.S. CoBA under different learning rates. (ResNet-34, CIFAR-100, FR(5))



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

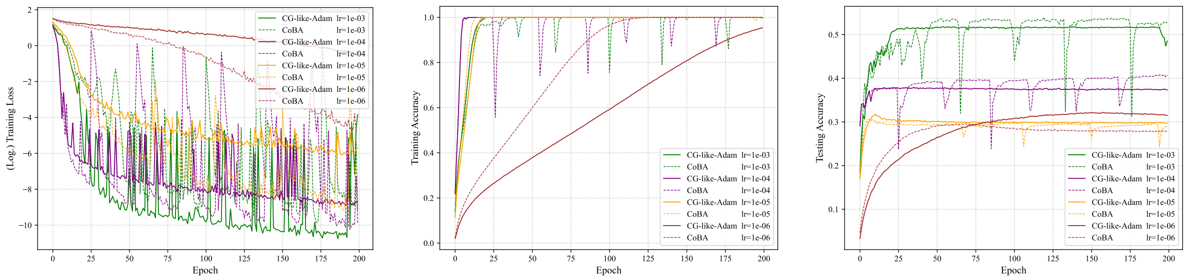
Figure 8: CG-like-Adam V.S. CoBA under different learning rates. (ResNet-34, CIFAR-100, PRP(6))

Figure 6-10 show the results of the experiments of ResNet-34 on CIFAR-100. Although training loss failed to reach the minimum unless the learning rate $\alpha_t = 10^{-3}$, CG-like-Adam obtained 100% training accuracy in less than 10 epochs when the learning rate is not too small, which, in our opinion, leads to overfitting and thus the testing accuracy is inferior(or similar) to CoBA.



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 9: CG-like-Adam V.S. CoBA under different learning rates. (ResNet-34, CIFAR-100, DY(7))

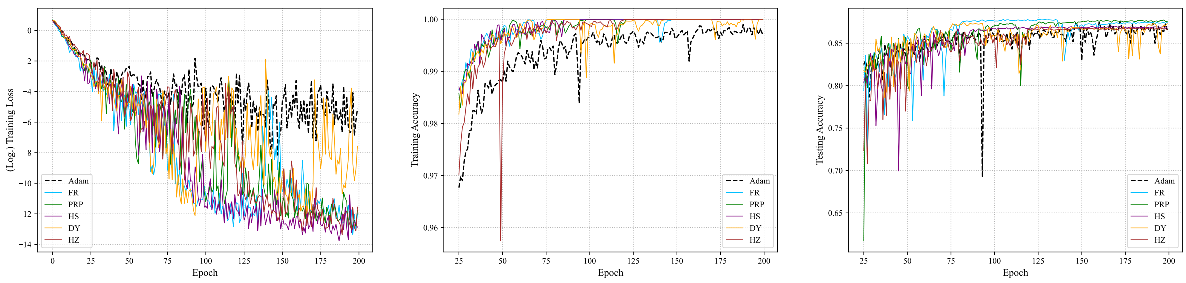


(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 10: CG-like-Adam V.S. CoBA under different learning rates. (ResNet-34, CIFAR-100, HZ(8))

4.2. Compare CG-like-Adam with Adam

$\alpha_t = 10^{-3} (\forall t \in \mathcal{T})$ is set as default value for this experiment. Figure 11 and figure 12 show the results of the experiments of VGG-19 on CIFAR-10, ResNet-34 on CIFAR-100, respectively. From

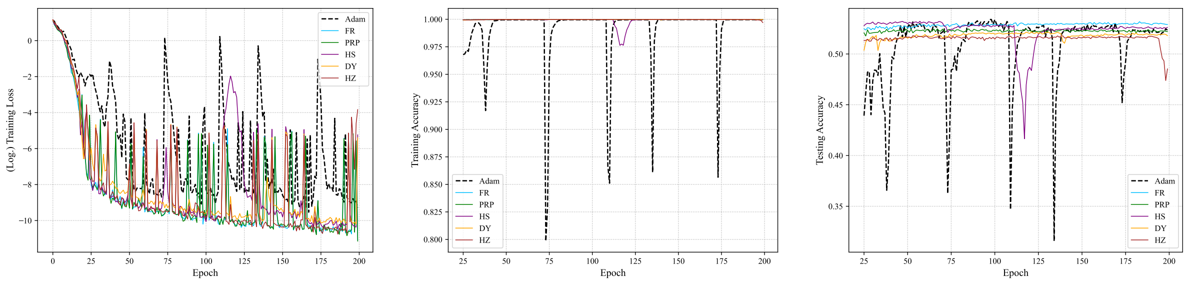


(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 11: Adam V.S. CG-like-Adam with different conjugate coefficient. Train VGG-19 on CIFAR-10.

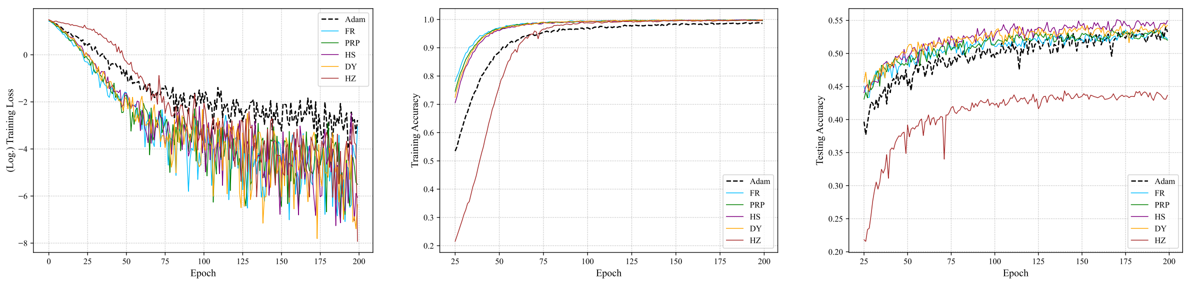
figure 11, CG-like-Adam is superior to Adam in the criterion of training loss, training accuracy and testing accuracy. Figure 12 tells that it also attains the minimum of training loss although it has some vibrations which, however, is more stable than Adam. In addition, CG-like-Adam defeats Adam in terms of training accuracy and testing accuracy and performs more consistently.

We trained VGG-19 on CIFAR-100 as well. Conclusion can be drawn from the results showed as figure 13 that CG-like-Adam performed better than Adam once again except the conjugate coefficient HZ(Eq.(8)), which may be upgraded by adjusting its λ.



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 12: Adam V.S. CG-like-Adam with different conjugate coefficient. Train ResNet-34 on CIFAR-100.



(a) (Log.) Training Loss (b) Training Accuracy (c) Testing Accuracy

Figure 13: Adam V.S. CG-like-Adam with different conjugate coefficient. Train VGG-19 on CIFAR-100.

5. Conclusion

In this paper, for the purpose of accelerating deep neural networks training and helping the network find more optimal parameters, the conjugate-gradient-like is incorporated into the generic Adam, which is named CG-like-Adam. The conjugate-gradient-like is modified from vanilla conjugate gradient, via scaling the conjugate coefficient by using a decreasing sequence over time step. The first-order and the second-order moment estimation of CG-like-Adam are both adopting conjugate-gradient-like. We theoretically prove the convergence of our algorithm and manage to not only provide convergence for the non-convex but also deal with the cases where the exponential moving average coefficient of the first-order moment estimation is constant and the first-order moment estimation is unbiased. Numerical experiments of training VGG-19/ResNet-34 on CIFAR-10/100 for image classification demonstrate effectiveness and better performance. Our algorithm performs more stable and arrives at 100% training accuracy faster than Adam. Higher testing accuracy provides strong evidence that our algorithm obtains more optimal parameters of deep neural networks. More future work includes conducting various experiments for deep learning task, applying variance reduction technique for the conjugate-gradient-like, exploring line-search for finding a suitable stepsize, etc.

References

[1] R. Xie, C. Ling, Y. Wang, R. Wang, F. Xia, L. Lin, Deep feedback network for recommendation, Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (2021) 2519–2525.

[2] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE transactions on neural networks and learning systems 28 (10) (2016) 2222–2232.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[4] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, Advances in Neural Information Processing Systems 33 (2020) 3008–3021.

[5] D. Soydaner, A comparison of optimization algorithms for deep learning, International Journal of Pattern Recognition and Artificial Intelligence 34 (13) (2020) 2052013.

[6] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR computational mathematics and mathematical physics 4 (5) (1964) 1–17.

[7] N. Qian, On the momentum term in gradient descent learning algorithms, Neural networks 12 (1) (1999) 145–151.

[8] Y. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$, Dokl. Akad. Nauk. SSSR 269 (1983) 543–547.

[9] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, Journal of machine learning research 12 (61) (2011) 2121–2159.

[10] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701 (2012).

[11] T. Tieleman, G. R. Hinton, Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning (2012) 4(2):26–31.

[12] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[13] T. Dozat, Incorporating nesterov momentum into adam, Proceedings of the 4th International Conference on Learning Representations (2016) 1–4.

[14] S. J. Reddi, S. Kale, S. Kumar, On the convergence of adam and beyond, arXiv preprint arXiv:1904.09237 (2019).

[15] D. Zhou, J. Chen, Y. Cao, Y. Tang, Z. Yang, Q. Gu, On the convergence of adaptive gradient methods for nonconvex optimization, arXiv preprint arXiv:1808.05671 (2018).

[16] X. Chen, S. Liu, R. Sun, M. Hong, On the convergence of a class of adam-type algorithms for non-convex optimization, arXiv preprint arXiv:1808.02941 (2018).

[17] M. R. Hestenes, E. Stiefel, et al., Methods of conjugate gradients for solving linear systems, Journal of research of the National Bureau of Standards 49 (6) (1952) 409–436.

[18] R. Fletcher, C. M. Reeves, Function minimization by conjugate gradients, The computer journal 7 (2) (1964) 149–154.

[19] E. Polak, G. Ribiere, Note on convergence of conjugate direction methods, Revue Francaise D Informatique De Recherche Operationnelle 3 (16) (1969) 35–43.

[20] B. T. Polyak, The conjugate gradient method in extremal problems, USSR Computational Mathematics and Mathematical Physics 9 (4) (1969) 94–112.

[21] Y.-H. Dai, Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, SIAM Journal on optimization 10 (1) (1999) 177–182.

[22] W. W. Hager, H. Zhang, A new conjugate gradient method with guaranteed descent and an efficient line search, SIAM Journal on optimization 16 (1) (2005) 170–192.

[23] H. Iiduka, Y. Kobayashi, Training deep neural networks using conjugate gradient-like methods, Electronics 9 (11) (2020) 1809.

[24] Y. Kobayashi, H. Iiduka, Conjugate-gradient-based adam for stochastic optimization and its application to deep learning, arXiv preprint arXiv:2003.00231 (2020).

[25] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Tech. rep., University of Toronto, Toronto, ON, Canada (2009).

[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.

Appendix A. Proof of Some Lemmas

Before beginning the proof of theorem 3.1, 3.2 and corollary 3.1, all the necessary lemmas as well as the proof of them are firstly provided at once. For the convenience of the following proof and without the loss of generality, let $\left(\frac{\alpha_1}{\sqrt{\hat{V}_1}}-\frac{\alpha_0}{\sqrt{\hat{V}_0}}\right)\hat{m}_0=0$.

Lemma 1. *Suppose* $\forall t\in\mathcal{T}$, $\beta_{1t}\in[0,1)$, $\beta_{1(t+1)}\leqslant\beta_{1t}$. *Defining* $\eta_t=\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{\xi_t}$ *and* $\xi_t=\left(1-\beta_{11}^t\right)\left[1-\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right]$, *then* $\forall t\in\mathcal{T}$, *the following holds:*

$$\beta_{1t}<\frac{1-\beta_{11}^t}{1-\beta_{11}^{t-1}},\ \left(1-\beta_{11}\right)^2\leqslant\xi_t\leqslant1,\ 0\leqslant\eta_t\leqslant\frac{1}{1-\beta_{11}}. \tag{A.1}$$

Proof. It is obvious that $\frac{1-\beta_{11}^t}{1-\beta_{11}^{t-1}}\geqslant\frac{1-\beta_{11}^{t-1}}{1-\beta_{11}^{t-1}}=1$. Because of $\beta_{1t}\in[0,1)$, hence $\forall t\in\mathcal{T}$, $\beta_{1t}<\frac{1-\beta_{11}^t}{1-\beta_{11}^{t-1}}$, which is equivalent to $\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}<1$. So then $\xi_t\leqslant1-\beta_{11}^t\leqslant1$.

If $\forall t\in\mathcal{T}$, $\beta_{1t}=0$, obviously, $\xi_t=1$, $\eta_t=0$.

If $\beta_{1t}\in(0,1)$,

$$\begin{aligned}\xi_t&=\left(1-\beta_{11}^t\right)\left[1-\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right]\geqslant\left(1-\beta_{11}^{t-1}\right)-\beta_{1t}\left(1-\beta_{11}^{t-1}\right)\\&=\left(1-\beta_{1t}\right)\left(1-\beta_{11}^{t-1}\right)\geqslant\left(1-\beta_{11}\right)^2,\end{aligned}$$

$$0<\eta_t\leqslant\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{\left(1-\beta_{1t}\right)\left(1-\beta_{11}^{t-1}\right)}=\frac{\beta_{1t}}{1-\beta_{1t}}\leqslant\frac{1}{1-\beta_{11}}.$$

The proof is over. □

Lemma 2. *Suppose* $\forall t\in\mathcal{T}$, $\beta_{1t}\in[0,1)$, $\beta_{1(t+1)}\leqslant\beta_{1t}$. *Define* $h(t)=\frac{\left(1-\beta_{11}^{t-1}\right)\left(1-\beta_{11}^{t+1}\right)}{\left(1-\beta_{11}^t\right)^2}$. $\forall t\in\mathcal{T}$, *if* $\beta_{1(t+1)}\leqslant(\geqslant)\beta_{1t}h(t)$, *then* $\eta_{t+1}\leqslant(\geqslant)\eta_t$ *holds.*

Proof. $\forall t\in\mathcal{T}$, $h(t)>0$,

$$\left(1-\beta_{11}^{t-1}\right)\left(1-\beta_{11}^{t+1}\right)-\left(1-\beta_{11}^t\right)^2=\beta_{11}^{t-1}\left(2\beta_{11}-1-\beta_{11}^2\right).$$

Because of $\beta_{11}\in[0,1)$, there is $-1\leqslant2\beta_{11}-1-\beta_{11}^2<0$. Such that

$$\left(1-\beta_{11}^{t-1}\right)\left(1-\beta_{11}^{t+1}\right)\leqslant\left(1-\beta_{11}^t\right)^2,$$

which is equivalent to $h(t)\leqslant1$.

From the lemma 1, $\forall t\in\mathcal{T}$, $\xi_t>0$. Let $\delta_t=\beta_{1(t+1)}\left(1-\beta_{11}^t\right)\xi_t-\beta_{1t}\left(1-\beta_{11}^{t-1}\right)\xi_{t+1}$, then

$$\eta_{t+1}-\eta_t=\frac{\delta_t}{\xi_t\xi_{t+1}},$$

$$\delta_t=\beta_{1(t+1)}\left(1-\beta_{11}^t\right)^2-\beta_{1t}\left(1-\beta_{11}^{t-1}\right)\left(1-\beta_{11}^{t+1}\right).$$

Then $\forall t \in \mathcal{T}$, different situations will be discussed below:

(1) If $\beta_{1t} = 0$, then $\delta_t = 0$, $\eta_{t+1} - \eta_t = 0$.

(2) If $\beta_{1t} = \beta_{1(t+1)} \neq 0$, then $h(t) \leqslant 1 = \frac{\beta_{1(t+1)}}{\beta_{1t}}$, $\delta_t \geqslant 0$, hence $\eta_{t+1} \geqslant \eta_t$.

(3) If $\beta_{1t} \in (0,1)$, because of $\beta_{1(t+1)} \leqslant (\geqslant) \beta_{1t} h(t)$, such that $\delta_t \leqslant (\geqslant) 0$, which leads to $\eta_{t+1} \leqslant (\geqslant) \eta_t$.

The proof is over. □

Lemma 3. *Suppose the assumption 3.4 is satisfied and* $\exists \bar{\gamma} \in \mathbb{R}^+$, $\forall t \in \mathcal{T}$, $|\gamma_t| \leqslant \bar{\gamma}$. *Further suppose* $\exists t_0 \in \mathcal{T}$, $\exists \bar{H} \in \mathbb{R}^+$, *such that* $\bar{H} = \max \left\{ 2H, \underset{t \in \{1,\cdots,t_0-1\}}{max} \|d_t\| \right\}$. *Then* $\forall t \in \mathcal{T}$, *the following holds:*

$$\|d_t\| \leqslant \bar{H}. \tag{A.2}$$

Proof. Due to $\underset{t \to +\infty}{lim} \frac{|\gamma_t|}{t^a} = 0$, thus $\exists t_0 \in \mathcal{T}$, $\forall t \geqslant t_0$, the following holds:

$$\frac{|\gamma_t|}{t^a} \leqslant \frac{1}{2}.$$

By the definition of $\bar{H}$, it is apparent that $\forall t < t_0$, $\|d_t\| \leqslant \bar{H}$.

When $t = t_0$, $\|d_{t_0-1}\| \leqslant \bar{H}$. From the update rule of d_{t_0} and triangular inequality, $\|d_{t_0}\|$ is bounded as follows:

$$\begin{aligned}
\|d_{t_0}\| &\leqslant \|g_{t_0}\| + \frac{|\gamma_{t_0}|}{t_0^a} \|d_{t_0-1}\| \\
&\leqslant \|g_{t_0}\| + \frac{1}{2} \|d_{t_0-1}\| \\
&\leqslant \bar{H}.
\end{aligned}$$

Suppose $\exists j > t_0$, $\|d_{j-1}\| \leqslant \bar{H}$. $\|d_j\|$ is bounded as follows:

$$\begin{aligned}
\|d_j\| &\leqslant \|g_j\| + \frac{|\gamma_j|}{j^a} \|d_{j-1}\| \\
&\leqslant \|g_j\| + \frac{1}{2} \|d_{j-1}\| \\
&\leqslant \bar{H}.
\end{aligned}$$

By the mathematical induction, it completes the proof. □

Lemma 4. *Suppose* $\forall t \in \mathcal{T}$, $a_t \geqslant 0$ *and* $\beta_{11} \in [0,1)$. *Let* $b_t = \sum_{i=1}^{t} \beta_{11}^{t-i} \sum_{l=i+1}^{t} a_l$. *Then the following inequality holds:*

$$\sum_{t=1}^{T-1} b_t^2 \leqslant \frac{1}{(1-\beta_{11})^4} \sum_{t=2}^{T-1} a_t^2. \tag{A.3}$$

Proof.

$$
\begin{aligned}
\sum_{t=1}^{T-1} b_t^2 &= \sum_{t=1}^{T-1}\left(\sum_{i=1}^{t}\beta_{11}^{t-i}\sum_{l=i+1}^{t} a_l\right)^2 = \sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\sum_{i=1}^{l}\beta_{11}^{t-i}a_l\right)^2 \\
&= \sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\sum_{i=1}^{l}\beta_{11}^{t-l}\beta_{11}^{l-i}a_l\right)^2 = \sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\beta_{11}^{t-l}a_l\sum_{i=1}^{l}\beta_{11}^{l-i}\right)^2 \\
&\leqslant \frac{1}{(1-\beta_{11})^2}\sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\beta_{11}^{t-l}a_l\right)^2 \\
&= \frac{1}{(1-\beta_{11})^2}\sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\beta_{11}^{t-l}a_l\right)\left(\sum_{m=2}^{t}\beta_{11}^{t-m}a_m\right) \\
&= \frac{1}{(1-\beta_{11})^2}\sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\sum_{m=2}^{t}\beta_{11}^{t-l}a_l\beta_{11}^{t-m}a_m\right) \\
&\leqslant \frac{1}{(1-\beta_{11})^2}\sum_{t=1}^{T-1}\left[\sum_{l=2}^{t}\sum_{m=2}^{t}\beta_{11}^{t-l}\beta_{11}^{t-m}\cdot\frac{1}{2}\left(a_l^2+a_m^2\right)\right] \\
&= \frac{1}{(1-\beta_{11})^2}\sum_{t=1}^{T-1}\left(\sum_{l=2}^{t}\sum_{m=2}^{t}\beta_{11}^{t-l}\beta_{11}^{t-m}a_l^2\right) \\
&\leqslant \frac{1}{(1-\beta_{11})^3}\sum_{t=1}^{T-1}\sum_{l=2}^{t}\beta_{11}^{t-l}a_l^2 = \frac{1}{(1-\beta_{11})^3}\sum_{l=2}^{T-1}\sum_{t=l}^{T-1}\beta_{11}^{t-l}a_l^2 \\
&\leqslant \frac{1}{(1-\beta_{11})^4}\sum_{l=2}^{T-1}a_l^2.
\end{aligned}
$$

The first, the third and the last inequality sign are both due to $\sum_{k=0}^{K}\beta_{11}^k \leqslant \frac{1}{1-\beta_{11}}$, and the second inequality sign is due to $ab \leqslant \frac{1}{2}(a^2+b^2)$. The last equal sign is because of the symmetry of t and l in the summation.

The proof is over. □

Lemma 5. *Let $x_0 \triangleq x_1$, $\beta_{11} \neq \frac{1}{2}$ in the CG-like-Adam(Alg.2). Consider the sequence $z_t = x_t + \eta_t\left(x_t - x_{t-1}\right), \forall t \in \mathcal{T}$, then the following holds:*

$$
\begin{aligned}
z_{t+1} - z_t = &-\left(\eta_{t+1}-\eta_t\right)\alpha_t\hat{V}_t^{-\frac{1}{2}}\hat{m}_t \\
&-\eta_t\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1} \\
&-\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t
\end{aligned}
\tag{A.4}
$$

and

$$
z_2 - z_1 = -\left(\eta_2-\eta_1\right)\alpha_1\hat{V}_1^{-\frac{1}{2}}\hat{m}_1 - \alpha_1\hat{V}_1^{-\frac{1}{2}}d_1,
$$

where $\eta_t = \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{\xi_t}$, $\xi_t = \left(1-\beta_{11}^t\right)\left[1-\frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right]$.

Proof. From the update rule of the algorithm, for all $t > 1$, there is

$$\begin{aligned}
x_{t+1} - x_t &= -\alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{m}_t = -\frac{\alpha_t}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} m_t \\
&= -\frac{\alpha_t}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} \left[\beta_{1t} m_{t-1} + (1-\beta_{1t})\, d_t\right] \\
&= \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} (x_t - x_{t-1}) - \frac{\alpha_t (1-\beta_{1t})}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} d_t \\
&\quad + \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} \left(\frac{\alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{V}_{t-1}^{\frac{1}{2}}}{\alpha_{t-1}} - I \right) (x_t - x_{t-1}) \\
&= \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} (x_t - x_{t-1}) - \frac{\alpha_t (1-\beta_{1t})}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} d_t \\
&\quad - \frac{\beta_{1t}\alpha_{t-1}}{1-\beta_{11}^t} \left(\frac{\alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{V}_{t-1}^{\frac{1}{2}}}{\alpha_{t-1}} - I \right) \hat{V}_{t-1}^{-\frac{1}{2}} m_{t-1} \\
&= \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} (x_t - x_{t-1}) - \frac{\alpha_t (1-\beta_{1t})}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} d_t \\
&\quad - \frac{\beta_{1t}}{1-\beta_{11}^t} \left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}} \right) m_{t-1}.
\end{aligned} \tag{A.5}$$

Since $\left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t+1}\right)}{1-\beta_{11}^t}\right] x_{t+1} + \frac{\beta_{1t}\left(1-\beta_{11}^{t+1}\right)}{1-\beta_{11}^t} (x_{t+1} - x_t) = \left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t+1}\right)}{1-\beta_{11}^t}\right] x_t + (x_{t+1} - x_t)$, combining with Eq.(A.5) we have

$$\begin{aligned}
&\left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right] x_{t+1} + \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} (x_{t+1} - x_t) \\
&= \left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right] x_t + \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} (x_t - x_{t-1}) \\
&\quad - \frac{\beta_{1t}}{1-\beta_{11}^t} \left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}} \right) m_{t-1} - \frac{\alpha_t (1-\beta_{1t})}{1-\beta_{11}^t} \hat{V}_t^{-\frac{1}{2}} d_t.
\end{aligned} \tag{A.6}$$

From the lemma 1, $\forall t \in \mathcal{T}$, $\beta_{1t} < \frac{1-\beta_{11}^t}{1-\beta_{11}^{t-1}}$. So as long as $\beta_{11} \neq \frac{1}{2}$, there is $\forall t \in \mathcal{T}$, $1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t} \neq 0$. Divide both sides of Eq.(A.6) by $\left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right]$, the following holds:

$$\begin{aligned}
&x_{t+1} + \eta_t (x_{t+1} - x_t) \\
&= x_t + \eta_t (x_t - x_{t-1}) - \frac{\beta_{1t}}{\xi_t} \left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}} \right) m_{t-1} - \frac{\alpha_t (1-\beta_{1t})}{\xi_t} \hat{V}_t^{-\frac{1}{2}} d_t.
\end{aligned} \tag{A.7}$$

Defining sequence $z_t = x_t + \eta_t (x_t - x_{t-1})$, $\eta_t = \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{\xi_t}$ in which

$$\xi_t = \left(1-\beta_{11}^t\right) \left[1 - \frac{\beta_{1t}\left(1-\beta_{11}^{t-1}\right)}{1-\beta_{11}^t}\right].$$

Then the above Eq.(A.7) can be converted into

$$
\begin{aligned}
& x_{t+1}+\eta_t\left(x_{t+1}-x_t\right)+\eta_{t+1}\left(x_{t+1}-x_t\right)-\eta_{t+1}\left(x_{t+1}-x_t\right) \\
& =z_{t+1}+\left(\eta_t-\eta_{t+1}\right)\left(x_{t+1}-x_t\right) \\
& =z_t-\frac{\eta_t}{1-\beta_{11}^{t-1}}\left(\alpha_t \hat{V}_t^{-\frac{1}{2}}-\alpha_{t-1} \hat{V}_{t-1}^{\frac{1}{2}}\right) m_{t-1} \\
& \quad-\frac{\alpha_t\left(1-\beta_{1 t}\right)}{\xi_t} \hat{V}_t^{-\frac{1}{2}} d_t .
\end{aligned} \tag{A.8}
$$

The Eq.(A.8) can be written as

$$
\begin{aligned}
z_{t+1}-z_t= & -\left(\eta_{t+1}-\eta_t\right) \alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{m}_t-\frac{\alpha_t\left(1-\beta_{1 t}\right)}{\xi_t} \hat{V}_t^{-\frac{1}{2}} d_t \\
& -\eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}}-\alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}}\right) \hat{m}_{t-1} .
\end{aligned}
$$

When $t=1$, $z_1=x_1+\eta_1\left(x_1-x_0\right)=x_1$, there is

$$
\begin{aligned}
z_2-z_1 & =x_2+\eta_2\left(x_2-x_1\right)-x_1 \\
& =\left(\eta_2-\eta_1\right)\left(x_2-x_1\right)+\left(1+\eta_1\right)\left(x_2-x_1\right) \\
& =-\left(\eta_2-\eta_1\right) \alpha_1 \hat{V}_1^{-\frac{1}{2}} \hat{m}_1-\alpha_1 \hat{V}_1^{-\frac{1}{2}} d_1 .
\end{aligned}
$$

The proof is over. □

Lemma 6. *Suppose the conditions in theorem 3.1 are satisfied, then*

$$
\mathbb{E}\left[f\left(z_{T+1}\right)-f\left(z_1\right)\right] \leqslant \sum_{i=1}^6 T_i, \tag{A.9}
$$

where

$$
T_1=-\mathbb{E}\left[\sum_{t=1}^T\left\langle\nabla f\left(z_t\right), \eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}}-\alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}}\right) \hat{m}_{t-1}\right\rangle\right], \tag{A.10}
$$

$$
T_2=-\mathbb{E}\left[\sum_{t=1}^T\left\langle\nabla f\left(z_t\right), \frac{\alpha_t\left(1-\beta_{1 t}\right)}{\xi_t} \hat{V}_t^{-\frac{1}{2}} d_t\right\rangle\right], \tag{A.11}
$$

$$
T_3=-\mathbb{E}\left[\sum_{t=1}^T\left\langle\nabla f\left(z_t\right),\left(\eta_{t+1}-\eta_t\right) \alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{m}_t\right\rangle\right], \tag{A.12}
$$

$$
T_4=\frac{3 L}{2} \mathbb{E}\left[\sum_{t=1}^T\left\|\left(\eta_{t+1}-\eta_t\right) \alpha_t \hat{V}_t^{-\frac{1}{2}} \hat{m}_t\right\|^2\right], \tag{A.13}
$$

$$
T_5=\frac{3 L}{2} \mathbb{E}\left[\sum_{t=1}^T\left\|\eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}}-\alpha_{t-1} \hat{V}_{t-1}^{-\frac{1}{2}}\right) \hat{m}_t\right\|^2\right], \tag{A.14}
$$

$$
T_6=\frac{3 L}{2} \mathbb{E}\left[\sum_{t=1}^T\left\|\frac{\alpha_t\left(1-\beta_{1 t}\right)}{\xi_t} \hat{V}_t^{-\frac{1}{2}} d_t\right\|^2\right]. \tag{A.15}
$$

Proof. Since ∇f is Lipschitz smooth, then

$$f\left(z_{t+1}\right) \leqslant f\left(z_t\right) + \left\langle \nabla f\left(z_t\right), r_t\right\rangle + \frac{L}{2}\left\|r_t\right\|^2, \tag{A.16}$$

where $r_t = z_{t+1} - z_t$. By the lemma 5,

$$\begin{aligned} r_t &= z_{t+1} - z_t \\ &= -\left(\eta_{t+1} - \eta_t\right)\alpha_t \hat{V}_t^{-\frac{1}{2}}\hat{m}_t - \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t \\ &\quad - \eta_t\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}. \end{aligned} \tag{A.17}$$

Combining Eq.(A.16) and Eq.(A.17) gets

$$\begin{aligned} &\mathbb{E}\left[f\left(z_{T+1}\right) - f\left(z_1\right)\right] \\ &\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(z_t\right), r_t\right\rangle\right] + \frac{L}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|r_t\right\|^2\right] \\ &= -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(z_t\right), \eta_t\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\rangle\right] \\ &\quad - \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(z_t\right), \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\rangle\right] \\ &\quad - \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(z_t\right), \left(\eta_{t+1}-\eta_t\right)\alpha_t\hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\rangle\right] \\ &\quad + \frac{L}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|r_t\right\|^2\right] \\ &= T_1 + T_2 + T_3 + \frac{L}{2}\mathbb{E}\left[\sum_{i=1}^{T}\left\|r_t\right\|^2\right]. \end{aligned} \tag{A.18}$$

Further more, by using the inequality $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, the last term of RHS of Eq.(A.18) can be bounded as follows:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T}\left\|r_t\right\|^2\right] \leq\ & 3\mathbb{E}\left[\sum_{t=1}^{T}\left\|\left(\eta_{t+1}-\eta_t\right)\alpha_t\hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\|^2\right] \\ &+ 3\mathbb{E}\left[\sum_{t=1}^{T}\left\|\eta_t\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\|^2\right] \\ &+ 3\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right]. \end{aligned} \tag{A.19}$$

Combining Eq.(A.18) and Eq.(A.19) leads to Eq.(A.9), which completes the proof. □

The following lemmas 7-11 separately bound the terms Eq.(A.10)-Eq.(A.14).

Lemma 7. *Suppose the conditions in theorem 3.1 are satisfied, then the term T_1(see Eq.(A.10)) in the lemma 6 satisfies the following inequality:*

$$\begin{aligned} T_1 &= -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(z_t\right), \eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\rangle\right] \\ &\leqslant \frac{\alpha_1 H\bar{M}}{1-\beta_{11}} E\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-\frac{1}{2}}\right], \end{aligned} \tag{A.20}$$

where $\bar{M}$ is a constant independent of T.

Proof. By the lemma 3, $\forall t \in \mathcal{T}$, $\|d_t\| \leqslant \bar{H}$. Letting $M \triangleq \max\left\{H, \bar{H}\right\} = \bar{H}$ and supposing $\|m_{t-1}\| \leqslant M$, then by the update rule of the algorithm,

$$\begin{gathered} \|m_t\| = \|\beta_{1t} m_{t-1} + (1-\beta_{1t})\, d_t\|, \\ \max\left\{\|d_t\|, \|m_{t-1}\|\right\} \leqslant M. \end{gathered}$$

By $m_0 = 0$, $\|m_0\| = 0 \leqslant M$ and mathematical induction, there is $\forall t \in \mathcal{T}$, $\|m_t\| \leqslant M$. Further more, $\|\hat{m}_t\|$ can be bounded as follows:

$$\|\hat{m}_t\| = \left\|\frac{m_t}{1-\beta_{11}^t}\right\| = \frac{\|m_t\|}{1-\beta_{11}^t} \leqslant \frac{\|m_t\|}{1-\beta_{11}} \leqslant \frac{M}{1-\beta_{11}} \triangleq \bar{M}.$$

Using Cauchy-Schwarz inequality and assumption $\left(\frac{\alpha_1}{\sqrt{\hat{V}_1}} - \frac{\alpha_0}{\sqrt{\hat{V}_0}}\right)\hat{m}_0 = 0$, we further have

$$\begin{aligned} T_1 &= -\mathbb{E}\left[\sum_{t=2}^{T}\left\langle \nabla f\left(x_t\right), \eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\rangle\right] \\ &\leqslant \mathbb{E}\left[\sum_{t=2}^{T}\|\nabla f\left(x_t\right)\| \cdot \left\|\eta_t\left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\|\right] \\ &\leqslant \frac{1}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\|\nabla f\left(x_t\right)\| \cdot \left\|\left(\alpha_t \hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_{t-1}\right\|\right] \\ &\leqslant \frac{1}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\|\nabla f\left(x_t\right)\| \cdot \|\hat{m}_{t-1}\| \right. \\ &\qquad\qquad \left. \cdot \sqrt{\sum_{i=1}^{d}\left(\alpha_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}} - \alpha_t \hat{v}_{t,i}^{-\frac{1}{2}}\right)^2}\right] \\ &\leqslant \frac{H\bar{M}}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{d}\left(\alpha_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}} - \alpha_t \hat{v}_{t,i}^{-\frac{1}{2}}\right)\right] \\ &= \frac{H\bar{M}}{1-\beta_{11}}\mathbb{E}\left[\sum_{i=1}^{d}\left(\alpha_1 \hat{v}_{1,i}^{-\frac{1}{2}} - \alpha_T \hat{v}_{T,i}^{-\frac{1}{2}}\right)\right] \leqslant \frac{\alpha_1 H\bar{M}}{1-\beta_{11}}\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-\frac{1}{2}}\right]. \end{aligned} \tag{A.21}$$

The second inequality sign is because of the lemma 1. The fourth inequality sign is due to the fact that $\forall t \in \mathcal{T}$, $\forall i \in [d]$, $\alpha_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}} - \alpha_t \hat{v}_{t,i}^{-\frac{1}{2}} \geq 0$ and the fact that when $a \geqslant 0$ and $b \geqslant 0$, $\left(a^2+b^2\right) \leqslant (a+b)^2$.

The proof is over. □

Lemma 8. *Suppose the conditions in theorem 3.1 are satisfied, then the term T_3(see Eq.(A.12)) in the lemma 6 satisfies the following inequality:*

$$T_3 = -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(z_t\right), \left(\eta_{t+1}-\eta_t\right)\alpha_t \hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\rangle\right] \leq \frac{1}{2}\left(H^2+G^2\right)\left|\eta_T-\eta_1\right|. \tag{A.22}$$

Proof. By the triangular inequality, there is

$$\begin{aligned} T_3 &\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\left|\eta_{t+1}-\eta_t\right| \cdot \frac{1}{2}\left(\left\|\nabla f\left(z_t\right)\right\|^2 + \left\|\alpha_t \hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\|^2\right)\right] \\ &\leqslant \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left|\eta_{t+1}-\eta_t\right|\left(H^2+G^2\right)\right] \\ &= \frac{1}{2}\left(H^2+G^2\right)\mathbb{E}\left[\sum_{t=1}^{T}\left|\eta_{t+1}-\eta_t\right|\right]. \end{aligned} \tag{A.23}$$

From the lemma 2, Eq.(A.23) can be further bounded as follows:

$$T_3 \leqslant \begin{cases} \frac{1}{2}\left(H^2+G^2\right)\left(\eta_T-\eta_1\right), & \forall t \in \mathcal{T}, \beta_{1(t+1)} \leqslant h(t)\beta_{1t}, \\ \frac{1}{2}\left(H^2+G^2\right)\left(\eta_1-\eta_T\right), & \forall t \in \mathcal{T}, \beta_{1(t+1)} \geqslant h(t)\beta_{1t}. \end{cases} \tag{A.24}$$

Eq.(A.24) is equivalent to $T_3 \leqslant \frac{1}{2}\left(H^2+G^2\right)\left|\eta_T-\eta_1\right|$.

The proof is over. □

What needs to remind readers is that, if $\exists T_0 \in \mathcal{T}$, $\forall t \in [T_0]$, $\beta_{1(t+1)} \leqslant (\geqslant) h(t)\beta_{1t}$, and $\forall t \in \mathcal{T} \setminus [T_0]$, $\beta_{1(t+1)} \geqslant (\leqslant) h(t)\beta_{1t}$, then it is obvious that

$$T_3 \leqslant \frac{1}{2}\left(H^2+G^2\right)\left|2\eta_{T_0}-\eta_T-\eta_1\right| = \frac{1}{2}\left(H^2+G^2\right)\left|-2\eta_{T_0}+\eta_1+\eta_T\right|.$$

Lemma 9. *Suppose the conditions in theorem 3.1 are satisfied, then the term T_4(see Eq.(A.13)) in the lemma 6 satisfies the following inequality:*

$$T_4 = \frac{3L}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\left(\eta_{t+1}-\eta_t\right)\alpha_t \hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\|^2\right] \leqslant \frac{3LG^2}{1-\beta_{11}}\left|\eta_T-\eta_1\right|. \tag{A.25}$$

Proof.

$$\begin{aligned} T_4 &= \frac{3L}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\eta_{t+1}-\eta_t\right)^2\left\|\alpha_t \hat{V}_t^{-\frac{1}{2}}\hat{m}_t\right\|^2\right] \leqslant \frac{3LG^2}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\eta_{t+1}-\eta_t\right)^2\right] \\ &\leqslant \frac{3LG^2}{2}\cdot\frac{2}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=1}^{T}\left|\eta_{t+1}-\eta_t\right|\right] \leqslant \frac{3LG^2}{1-\beta_{11}}\left|\eta_T-\eta_1\right|, \end{aligned} \tag{A.26}$$

where the penultimate inequality sign is because of the lemma 1 and $|\eta_{t+1}-\eta_t| \leqslant \eta_{t+1}+\eta_t$, while the last one is similar to the proof of lemma 8.

The proof is over. □

Lemma 10. *Suppose the conditions in theorem 3.1 are satisfied, then the term T_5(see Eq.(A.14)) in the lemma 6 satisfies the following inequality:*

$$T_5 = \frac{3L}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\eta_t\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_t\right\|^2\right] \leqslant \frac{3\alpha_1^2 L\bar{M}^2}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-1}\right]. \tag{A.27}$$

Proof.

$$\begin{aligned}
T_5 &\leqslant \frac{3L}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\left(\alpha_t\hat{V}_t^{-\frac{1}{2}} - \alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\hat{m}_t\right\|^2\right] \\
&\leqslant \frac{3L}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=2}^{T}\left(\sum_{i=1}^{d}\left(\alpha_t\hat{v}_{t,i}^{-\frac{1}{2}} - \alpha_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}}\right)^2\right)\left\|\hat{m}_t\right\|^2\right] \\
&\leqslant \frac{3L\bar{M}^2}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{d}\left(\alpha_t\hat{v}_{t,i}^{-\frac{1}{2}} - \alpha_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}}\right)^2\right] \\
&\leqslant \frac{3L\bar{M}^2}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{d}\left(\alpha_{t-1}^2\hat{v}_{t-1,i}^{-1} - \alpha_t^2\hat{v}_{t,i}^{-1}\right)\right] \\
&= \frac{3L\bar{M}^2}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{i=1}^{d}\left(\alpha_1^2\hat{v}_{1,i}^{-1} - \alpha_T^2\hat{v}_{T,i}^{-1}\right)\right] \leqslant \frac{3\alpha_1^2 L\bar{M}^2}{2\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-1}\right],
\end{aligned} \tag{A.28}$$

where the first inequality sign is because of the lemma 1.

The proof is over. □

Lemma 11. *Suppose the conditions in theorem 3.1 are satisfied, then the term T_2(see Eq.(A.11)) in the lemma 6 satisfies the following inequality:*

$$\begin{aligned}
T_2 &= -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(z_t\right), \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\rangle\right] \\
&\leqslant \frac{L^2}{\left(1-\beta_{11}\right)^6}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + \frac{\alpha_1^2L^2\bar{H}^2}{\left(1-\beta_{11}\right)^8}\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-1}\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + 2\mu_1H^2\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-\frac{1}{2}}\right] \\
&\quad + 2H^2\mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{d}\left|\mu_t\hat{v}_{t,i}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,i}^{-\frac{1}{2}}\right|\right] \\
&\quad - E\left[\sum_{t=1}^{T}\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle\nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}}c_t\right\rangle\right],
\end{aligned} \tag{A.29}$$

where $\mu_t = \frac{\alpha_t(1-\beta_{1t})}{\xi_t}$, $c_t = \nabla f\left(x_t\right) - \frac{\gamma_t}{t^a}d_{t-1}$.

Proof. Let

$$T_{21} = -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(x_t\right), \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\rangle\right], \tag{A.30}$$

$$T_{22} = -\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(z_t\right) - \nabla f\left(x_t\right), \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\rangle\right]. \tag{A.31}$$

Then T_2 can be written as

$$T_2 = T_{21} + T_{22}. \tag{A.32}$$

The following is firstly to bound the term T_{22}(see Eq.(A.31)). Since $z_1 = x_1$, $z_t - x_t = \eta_t\left(x_t - x_{t-1}\right) = -\eta_t\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\hat{m}_{t-1}$ and the triangular inequality, we get

$$\begin{aligned}
T_{22} &\leqslant \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left(\left\|\nabla f\left(z_t\right) - \nabla f\left(x_t\right)\right\|^2 + \left\|\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right)\right] \\
&\leqslant \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}L^2\left\|z_t - x_t\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] \\
&= \frac{L^2}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\eta_t\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\hat{m}_{t-1}\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right].
\end{aligned} \tag{A.33}$$

The second inequality sign is due to the assumption 3.1. We now define

$$T_7 = \mathbb{E}\left[\sum_{t=2}^{T}\left\|\eta_t\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\hat{m}_{t-1}\right\|^2\right]$$

and bound it as follows:

$$\begin{aligned}
T_7 &\leqslant \frac{1}{\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\hat{m}_{t-1}\right\|^2\right] \\
&= \frac{1}{\left(1-\beta_{11}\right)^2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\frac{\left(\beta_{1(t-1)}m_{t-2} + \left(1-\beta_{1(t-1)}\right)d_{t-1}\right)}{1-\beta_{11}^{t-1}}\right\|^2\right] \\
&\leqslant \frac{1}{\left(1-\beta_{11}\right)^4}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)\left(1-\beta_{1i}\right)d_i\right)\right\|^2\right] \\
&\leqslant \frac{2}{\left(1-\beta_{11}\right)^4}\left(T_{71} + T_{72}\right),
\end{aligned} \tag{A.34}$$

where

$$T_{71} = \mathbb{E}\left[\sum_{t=2}^{T}\left\|\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)\left(1-\beta_{1i}\right)\alpha_i\hat{V}_i^{-\frac{1}{2}}d_i\right\|^2\right], \tag{A.35}$$

$$T_{72} = \mathbb{E}\left[\sum_{t=2}^{T}\left\|\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\left(\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}} - \alpha_i\hat{V}_i^{-\frac{1}{2}}\right)d_i\right\|^2\right]. \tag{A.36}$$

The first inequality sign is because of the lemma 1 and the last is due to $a^2 = (b+a-b)^2 \leqslant 2b^2 + 2(a-b)^2$.

The following is next to bound the term T_{71}(see Eq.(A.35)).

$$\begin{aligned}
T_{71} &= \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\,\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)\right. \\
&\qquad\qquad \left.\cdot\left(\sum_{l=1}^{t-1}\left(\prod_{p=l+1}^{t-1}\beta_{1p}\right)(1-\beta_{1l})\,\alpha_l\hat{v}_{l,k}^{-\frac{1}{2}}d_{l,k}\right)\right] \\
&= \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\sum_{i=1}^{t-1}\sum_{l=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)\right. \\
&\qquad\qquad \left.\cdot\left(\prod_{p=l+1}^{t-1}\beta_{1p}\right)(1-\beta_{1l})\left(\alpha_l\hat{v}_{l,k}^{-\frac{1}{2}}d_{l,k}\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\sum_{i=1}^{t-1}\sum_{l=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\left(\prod_{p=l+1}^{t-1}\beta_{1p}\right)\right. \\
&\qquad\qquad \left.\cdot(1-\beta_{1l})\frac{1}{2}\left(\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)^2 + \left(\alpha_l\hat{v}_{l,k}^{-\frac{1}{2}}d_{l,k}\right)^2\right)\right] \qquad \text{(A.37)} \\
&\leqslant \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\sum_{i=1}^{t-1}\beta_{11}^{t-i-1}\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)^2\sum_{l=1}^{t-1}\beta_{11}^{t-l-1}\right] \\
&\leqslant \frac{1}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\sum_{i=1}^{t-1}\beta_{11}^{t-i-1}\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)^2\right] \\
&= \frac{1}{1-\beta_{11}}\mathbb{E}\left[\sum_{i=1}^{T-1}\sum_{k=1}^{d}\sum_{t=i+1}^{T}\beta_{11}^{t-i-1}\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)^2\right] \\
&\leqslant \frac{1}{(1-\beta_{11})^2}\mathbb{E}\left[\sum_{i=1}^{T-1}\sum_{k=1}^{d}\left(\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}d_{i,k}\right)^2\right] \\
&= \frac{1}{(1-\beta_{11})^2}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right].
\end{aligned}$$

The first, second, third inequality sign are because of $ab \leqslant \frac{1}{2}(a^2+b^2)$; $\forall t \in \mathcal{T}$, $\beta_{1t} \leqslant \beta_{11}$; $\sum_{l=1}^{t-1}\beta_{11}^{t-l-1} \leqslant \frac{1}{1-\beta_{11}}$, respectively. The third, fourth equal sign are due to the symmetry of i and l in the summation; exchanging order of summation, respectively.

The next is to bound the term T_{72}(see Eq.(A.36)).

$$\begin{aligned}
T_{72} &= \mathbb{E}\left[\sum_{t=2}^{T}\left\|\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\left(\alpha_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}-\alpha_i\hat{V}_i^{-\frac{1}{2}}\right)d_i\right\|^2\right] \\
&\leqslant \bar{H}^2\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}\beta_{1j}\right)(1-\beta_{1i})\left|\alpha_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}-\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}\right|\right)^2\right] \\
&\leqslant \bar{H}^2\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{k=1}^{d}\left(\sum_{i=1}^{t}\beta_{11}^{t-i}\left|\alpha_t\hat{v}_{t,k}^{-\frac{1}{2}}-\alpha_i\hat{v}_{i,k}^{-\frac{1}{2}}\right|\right)^2\right] \\
&\leqslant \bar{H}^2\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{k=1}^{d}\left(\sum_{i=1}^{t}\beta_{11}^{t-i}\sum_{l=i+1}^{t}\left|\alpha_l\hat{v}_{l,k}^{-\frac{1}{2}}-\alpha_{l-1}\hat{v}_{l-1,k}^{-\frac{1}{2}}\right|\right)^2\right].
\end{aligned} \tag{A.38}$$

The first and the last inequality signs are both because of triangular inequality. By defining $a_l=\left|\alpha_l\hat{v}_{l,k}^{-\frac{1}{2}}-\alpha_{l-1}\hat{v}_{l-1,k}^{-\frac{1}{2}}\right|$ and using the lemma 4, we have

$$\begin{aligned}
T_{72} &\leqslant \frac{\bar{H}^2}{(1-\beta_{11})^4}\mathbb{E}\left[\sum_{t=2}^{T-1}\sum_{k=1}^{d}\left(\alpha_t\hat{v}_{t,k}^{-\frac{1}{2}}-\alpha_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right)^2\right] \\
&\leqslant \frac{\bar{H}^2}{(1-\beta_{11})^4}\mathbb{E}\left[\sum_{t=2}^{T-1}\sum_{k=1}^{d}\left(\alpha_{t-1}^2\hat{v}_{t-1,k}^{-1}-\alpha_t^2\hat{v}_{t,k}^{-1}\right)\right] \\
&= \frac{\bar{H}^2}{(1-\beta_{11})^4}\mathbb{E}\left[\sum_{k=1}^{d}\left(\alpha_1^2\hat{v}_{1,k}^{-1}-\alpha_T^2\hat{v}_{T,k}^{-1}\right)\right] \leqslant \frac{\alpha_1^2\bar{H}^2}{(1-\beta_{11})^4}\mathbb{E}\left[\sum_{k=1}^{d}\hat{v}_{1,k}^{-1}\right].
\end{aligned} \tag{A.39}$$

Combining Eq.(A.33)-Eq.(A.39), the term T_{22}(see Eq.(A.31)) can be bounded as follows:

$$\begin{aligned}
T_{22} &\leqslant \frac{L^2}{2}T_7+\frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t(1-\beta_{1t})}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] \\
&\leqslant \frac{L^2}{(1-\beta_{11})^4}(T_{71}+T_{72})+\frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t(1-\beta_{1t})}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] \\
&\leqslant \frac{1}{2}\mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_t(1-\beta_{1t})}{\xi_t}\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right]+\frac{L^2}{(1-\beta_{11})^6}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}a_t\right\|^2\right] \\
&\quad+\frac{\alpha_1^2L^2\bar{H}^2}{(1-\beta_{11})^8}\mathbb{E}\left[\sum_{k=1}^{d}v_{1,k}^{-1}\right].
\end{aligned} \tag{A.40}$$

It is now bounding the term T_{21}(see Eq.(A.30)). From the assumption 3.3, we can get $d_t=g_t-\frac{\gamma_t}{t^a}d_{t-1}=\nabla f(x_t)+\zeta_t-\frac{\gamma_t}{t^a}d_{t-1}=\nabla f(x_t)-\frac{\gamma_t}{t^a}d_{t-1}+\zeta_t=c_t+\zeta_t$, in which $c_t=\nabla f(x_t)-\frac{\gamma_t}{t^a}d_{t-1}$,

$\mathbb{E}[\zeta_t] = 0$. Then the following holds:

$$\begin{aligned}
T_{21} &= -\mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle \nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}}\left(c_t+\zeta_t\right)\right\rangle\right] \\
&= -\mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle \nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}} c_t\right\rangle\right] \\
&\quad - \mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle \nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}} \zeta_t\right\rangle\right].
\end{aligned} \tag{A.41}$$

Let $\mu_t = \frac{\alpha_t(1-\beta_{1t})}{\xi_t}$, then

$$\begin{aligned}
&- \mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle \nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}} \zeta_t\right\rangle\right] \\
&= -\mathbb{E}\left[\sum_{t=2}^{T} \left\langle \nabla f\left(x_t\right), \left(\mu_t \hat{V}_t^{-\frac{1}{2}} - \mu_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\zeta_t\right\rangle\right] \\
&\quad - \mathbb{E}\left[\sum_{t=2}^{T} \mu_{t-1}\left\langle \nabla f\left(x_t\right), \hat{V}_{t-1}^{-\frac{1}{2}} \zeta_t\right\rangle\right] - \mathbb{E}\left[\mu_1\left\langle \nabla f(x_t), \hat{V}_1^{-\frac{1}{2}}\zeta_1\right\rangle\right] \\
&\leqslant -\mathbb{E}\left[\sum_{t=2}^{T} \left\langle \nabla f\left(x_t\right), \left(\mu_t \hat{V}_t^{-\frac{1}{2}} - \mu_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\zeta_t\right\rangle\right] + 2\mu_1 H^2 \mathbb{E}\left[\sum_{k=1}^{d} \hat{v}_{1,k}^{-\frac{1}{2}}\right].
\end{aligned} \tag{A.42}$$

The last inequality sign is due to $\forall t \in \mathcal{T}\backslash\{1\}$, $\mathbb{E}\left[\hat{V}_{t-1}^{-\frac{1}{2}}\zeta_t \mid x_t, \hat{V}_{t-1}\right] = 0$, and the fact that $\|\zeta_t\| - \|\nabla f\left(x_t\right)\| \leqslant \|\nabla f\left(x_t\right) + \zeta_t\| = \|g_t\| \leqslant H$, $\|\zeta_t\| \leqslant H + \|\nabla f\left(x_t\right)\| \leqslant 2H$. Further more,

$$\begin{aligned}
&- \mathbb{E}\left[\sum_{t=2}^{T} \left\langle \nabla f\left(x_t\right), \left(\mu_t \hat{V}_t^{-\frac{1}{2}} - \mu_{t-1}\hat{V}_{t-1}^{-\frac{1}{2}}\right)\zeta_t\right\rangle\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d} \left|\left(\nabla f\left(x_t\right)\right)_k\right| \cdot \left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right| \cdot \left|\zeta_{t,k}\right|\right] \\
&\leqslant 2H^2 \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d} \left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right].
\end{aligned} \tag{A.43}$$

Therefore, combining Eq.(A.42) and Eq.(A.43), term T_{21}(see Eq.(A.41)) can be bounded as follows:

$$\begin{aligned}
T_{21} &\leqslant -\mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left(1-\beta_{1t}\right)}{\xi_t}\left\langle \nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}} c_t\right\rangle\right] \\
&\quad + 2H^2 \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d} \left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] + 2\mu_1 H^2 \mathbb{E}\left[\sum_{k=1}^{d} \hat{v}_{1,k}^{-\frac{1}{2}}\right].
\end{aligned} \tag{A.44}$$

Finally, the results of Eq.(A.32), Eq.(A.40) and Eq.(A.44) ensure Eq.(A.29).
The proof is over. □

Appendix B. Proof of Theorem 3.1

Proof. By the lemma 6, it is obvious that

$$\mathbb{E}\left[f\left(z_{t+1}\right)-f\left(z_{1}\right)\right] \leqslant \sum_{i=1}^{6} T_{i}. \tag{B.1}$$

From the lemma 7-11, Eq.(B.1) can be further bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[f\left(z_{t+1}\right)-f\left(z_{1}\right)\right] \\
&\leqslant\left[\frac{\alpha_{1} H \bar{M}}{1-\beta_{11}}+2 \mu_{1} H^{2}\right] E\left[\sum_{i=1}^{d} \hat{v}_{1, i}^{-\frac{1}{2}}\right] \\
&\quad+\left[\frac{H^{2}+G^{2}}{2}+\frac{3 L G^{2}}{1-\beta_{11}}\right]\left|\eta_{T}-\eta_{1}\right| \\
&\quad+\left[\frac{3 \alpha_{1}^{2} L \bar{M}^{2}}{2\left(1-\beta_{11}\right)^{2}}+\frac{\alpha_{1}^{2} L^{2} \bar{H}^{2}}{\left(1-\beta_{11}\right)^{8}}\right] \mathbb{E}\left[\sum_{i=1}^{d} \hat{v}_{1, i}^{-1}\right] \\
&\quad+\frac{L^{2}}{\left(1-\beta_{11}\right)^{6}} \mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_{t} \hat{V}_{t}^{-\frac{1}{2}} d_{t}\right\|^{2}\right] \\
&\quad+\frac{1}{2} \mathbb{E}\left[\sum_{t=2}^{T}\left\|\frac{\alpha_{t}\left(1-\beta_{1 t}\right)}{\xi_{t}} \hat{V}_{t}^{-\frac{1}{2}} d_{t}\right\|^{2}\right] \\
&\quad+2 H^{2} \mathbb{E}\left[\sum_{t=2}^{T} \sum_{i=1}^{d}\left|\mu_{t} \hat{v}_{t, i}^{-\frac{1}{2}}-\mu_{t-1} \hat{v}_{t-1, i}^{-\frac{1}{2}}\right|\right] \\
&\quad-E\left[\sum_{t=1}^{T} \frac{\alpha_{t}\left(1-\beta_{1 t}\right)}{\xi_{t}}\left\langle\nabla f\left(x_{t}\right), \hat{V}_{t}^{-\frac{1}{2}} c_{t}\right\rangle\right] .
\end{aligned} \tag{B.2}$$

Rearranging Eq.(B.2) and uniting like terms, we can get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_{t}\left(1-\beta_{1 t}\right)}{\xi_{t}}\left\langle\nabla f\left(x_{t}\right), \hat{V}_{t}^{-\frac{1}{2}} c_{t}\right\rangle\right] \\
&\leqslant\left(\frac{\alpha_{1} H \bar{M}}{1-\beta_{11}}+2 \mu_{1} H^{2}\right) \mathbb{E}\left[\sum_{i=1}^{d} \hat{v}_{1, i}^{-\frac{1}{2}}\right] \\
&\quad+\left(\frac{H^{2}+G^{2}}{2}+\frac{3 L G^{2}}{1-\beta_{11}}\right) \frac{2}{1-\beta_{11}} \\
&\quad+\left[\frac{3 \alpha_{1}^{2} L \bar{M}^{2}}{2\left(1-\beta_{11}\right)^{2}}+\frac{\alpha_{1}^{2} L^{2} \bar{H}^{2}}{\left(1-\beta_{11}\right)^{8}}\right] \mathbb{E}\left[\sum_{i=1}^{d} \hat{v}_{1, i}^{-1}\right] \\
&\quad+\left[\frac{L^{2}}{\left(1-\beta_{11}\right)^{6}}+\frac{1}{2(1-\beta_{11})^{4}}\right] \mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_{t} \hat{V}_{t}^{-\frac{1}{2}} d_{t}\right\|^{2}\right] \\
&\quad+2 H^{2} \mathbb{E}\left[\sum_{t=2}^{T} \sum_{i=1}^{d}\left|\mu_{t} \hat{v}_{t, i}^{-\frac{1}{2}}-\mu_{t-1} \hat{v}_{t-1, i}^{-\frac{1}{2}}\right|\right]+\mathbb{E}\left[f\left(z_{1}\right)-f\left(z^{*}\right)\right],
\end{aligned} \tag{B.3}$$

where $z^* = \underset{z\in X}{arg\,min} f(z)$.

Let

$$
\begin{aligned}
&C_1^{'} = 2H^2,\ C_2^{'} = \frac{L^2}{(1-\beta_{11})^6} + \frac{1}{2(1-\beta_{11})^4},\\
&C_3^{'} = \mathbb{E}\left[f(z_1) - f(z^*)\right]\\
&\quad + \left(\frac{\alpha_1 H\bar{M}}{1-\beta_{11}} + 2\mu_1 H^2\right)\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-\frac{1}{2}}\right]\\
&\quad + \left(\frac{H^2+G^2}{2} + \frac{3LG}{1-\beta_{11}}\right)\frac{2}{1-\beta_{11}}\\
&\quad + \left[\frac{3\alpha_1^2 L\bar{M}}{2(1-\beta_{11})^2} + \frac{\alpha_1^2 L^2\bar{H}^2}{(1-\beta_{11})^8}\right]\mathbb{E}\left[\sum_{i=1}^{d}\hat{v}_{1,i}^{-1}\right].
\end{aligned}
$$

Hence the following holds:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\frac{\alpha_t(1-\beta_{1t})}{\xi_t}\left\langle\nabla f(x_t), \hat{V}_t^{-\frac{1}{2}}c_t\right\rangle\right]\\
&\leqslant C_1^{'}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t\hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] + C_2^{'}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + C_3^{'}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\left\langle\nabla f(x_t), \hat{V}_t^{-\frac{1}{2}}c_t\right\rangle\right]\\
&\leqslant \frac{\xi_t C_1^{'}}{1-\beta_{1t}}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t\hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] + \frac{\xi_t C_2^{'}}{1-\beta_{1t}}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + \frac{\xi_t C_3^{'}}{1-\beta_{1t}}\\
&\leqslant \frac{C_1^{'}}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t\hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] + \frac{C_2^{'}}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + \frac{C_3^{'}}{1-\beta_{11}}.
\end{aligned}
\tag{B.4}
$$

The last inequality sign is due to the lemma 1 and $1-\beta_{1t} \geqslant 1-\beta_{11}$.

Since $c_t = \nabla f(x_t) - \frac{\gamma_t}{t^a}d_{t-1}$, then there is

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\left\langle\nabla f(x_t), \hat{V}_t^{-\frac{1}{2}}c_t\right\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f(x_t), \alpha_t\hat{V}_t^{-\frac{1}{2}}\nabla f(x_t)\right\rangle\right]\\
&\quad - \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f(x_t), \frac{\alpha_t\gamma_t}{t^a}\hat{V}_t^{-\frac{1}{2}}d_{t-1}\right\rangle\right].
\end{aligned}
\tag{B.5}
$$

Let

$$
\begin{aligned}
R_1 &= \frac{C_1^{'}}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t\hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right]\\
&\quad + \frac{C_2^{'}}{1-\beta_{11}}\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t\hat{V}_t^{-\frac{1}{2}}d_t\right\|^2\right] + \frac{C_3^{'}}{1-\beta_{11}}.
\end{aligned}
\tag{B.6}
$$

Combining Eq.(B.4), Eq.(B.5) and Eq.(B.6) comes to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(x_t\right), \alpha_t \hat{V}_t^{-\frac{1}{2}} \nabla f\left(x_t\right)\right\rangle\right] \\
&\leqslant R_1+\mathbb{E}\left[\sum_{t=1}^{T}\left\langle\nabla f\left(x_t\right), \frac{\alpha_t \gamma_t}{t^a} \hat{V}_t^{-\frac{1}{2}} d_{t-1}\right\rangle\right] \\
&\leqslant R_1+\mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left|\gamma_t\right|}{t^a}\left\|\nabla f\left(x_t\right)\right\| \cdot\left\|\hat{V}_t^{-\frac{1}{2}} d_{t-1}\right\|\right] \\
&\leqslant R_1+H \mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left|\gamma_t\right|}{t^a} \sqrt{\sum_{i=1}^{d} \hat{v}_{t, i}^{-1} d_{t-1, i}^2}\right] \\
&\leqslant R_1+H \bar{H} \sqrt{\sum_{i=1}^{d} \hat{v}_{1, i}^{-1}} \mathbb{E}\left[\sum_{t=1}^{T} \frac{\alpha_t\left|\gamma_t\right|}{t^a}\right] .
\end{aligned} \tag{B.7}$$

The last inequality sgin is due to the lemma 3 and $\forall t \in \mathcal{T}$, $\hat{v}_{t-1,i} \leqslant \hat{v}_{t,i}$.

Let

$$\begin{aligned}
&C_1=\frac{C_1^{'}}{1-\beta_{11}}, \ C_2=\frac{C_2^{'}}{1-\beta_{11}}, \\
&C_3=\frac{C_3^{'}}{1-\beta_{11}}, \ C_4=H \bar{H} \sqrt{\sum_{i=1}^{d} \hat{v}_{1, i}^{-1}} .
\end{aligned} \tag{B.8}$$

Combining Eq.(B.6), Eq.(B.7) and Eq.(B.8) completes the proof. □

Appendix C. Proof of Theorem 3.2

Proof. Let

$$R_2 = C_1 \mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}} - \mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right] + C_2 \mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t \hat{V}_t^{-\frac{1}{2}} d_t\right\|^2\right] + C_3 + C_4 \sum_{t=1}^{T}\frac{\alpha_t\left|\gamma_t\right|}{t^a}. \tag{C.1}$$

On one hand, from the definition of $O(\cdot)$, $\Omega(\cdot)$, obviously $\exists K_1, K_2 \in \mathcal{R}^+$, $\exists T_0 \in \mathcal{T}$, $\forall T \geqslant T_0$,

$$R_2 \leqslant K_1 S_1(T), \quad \sum_{t=1}^{T}\tau_t \geqslant K_2 S_2(T) > 0. \tag{C.2}$$

On the other hand, if $T \geqslant T_0$, then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \nabla f\left(x_t\right), \alpha_t \hat{V}_t^{-\frac{1}{2}} \nabla f\left(x_t\right)\right\rangle\right] \\
&\geqslant \mathbb{E}\left[\sum_{t=1}^{T}\tau_t\left\|\nabla f\left(x_t\right)\right\|^2\right] \\
&= \sum_{t=1}^{T}\tau_t \cdot \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \\
&\geqslant \sum_{t=1}^{T}\tau_t \cdot \underset{t\in T}{min}\ \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \\
&\geqslant K_2 S_2(T) \cdot \underset{t\in T}{min}\ \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right].
\end{aligned} \tag{C.3}$$

Combining the theorem 3.1, Eq.(C.1), Eq.(C.2) and Eq.(C.3), when $T \geqslant T_0$, we have

$$K_2 S_2(T) \cdot \underset{t\in T}{min}\ \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leqslant R_2 \leqslant K_1 S_1(T). \tag{C.4}$$

It is equivalent to when $T \to +\infty$,

$$\begin{aligned}
\underset{t\in T}{min}\ \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] &\leqslant \frac{K_1}{K_2}\frac{S_1(T)}{S_2(T)}, \\
\underset{t\in T}{min}\ \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] &= O\left(\frac{S_1(T)}{S_2(T)}\right).
\end{aligned} \tag{C.5}$$

The proof is over. $\square$

Appendix D. Proof of Corollary 3.1

Proof. Firstly proof the following:

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T-1}\left\|\alpha_t \hat{V}_t^{-\frac{1}{2}} d_t\right\|^2\right] \leqslant \mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_t}{c} d_t\right\|^2\right] \\ &\leqslant \frac{\alpha^2 \bar{H}^2}{c^2}\sum_{t=1}^{T}\frac{1}{t^{2b}} \leqslant \frac{\alpha^2 \bar{H}^2}{c^2}\sum_{t=1}^{T}\frac{1}{t} \leqslant \frac{\alpha^2 \bar{H}^2}{c^2}(1+\ln T) \end{aligned} \tag{D.1}$$

and

$$\sum_{t=1}^{T}\frac{\alpha_t\left|\gamma_t\right|}{t^b} \leqslant \alpha\bar{\gamma}\sum_{t=1}^{T}\frac{1}{t^{a+b}} \leqslant \alpha\bar{\gamma}\sum_{t=1}^{T}\frac{1}{t} \leqslant \alpha\bar{\gamma}(1+\ln T). \tag{D.2}$$

Since β_{1t} is a constant,namely $\beta_{1t}=\beta_{11}$, so $\mu_t=\frac{\alpha_t(1-\beta_{1t})}{\xi_t}=\frac{\alpha_t(1-\beta_{11})}{1-\beta_{11}^t-\beta_{11}(1-\beta_{11}^{t-1})}=\alpha_t$, we have

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left|\mu_t \hat{v}_{t,k}^{-\frac{1}{2}}-\mu_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}\right|\right]=\mathbb{E}\left[\sum_{t=2}^{T}\sum_{k=1}^{d}\left(\alpha_{t-1}\hat{v}_{t-1,k}^{-\frac{1}{2}}-\alpha_t \hat{v}_{t,k}^{-\frac{1}{2}}\right)\right] \\ &=\mathbb{E}\left[\sum_{k=1}^{d}\left(\alpha_1 \hat{v}_{1,k}^{-\frac{1}{2}}-\alpha_T \hat{v}_{T,k}^{-\frac{1}{2}}\right)\right] \leqslant \mathbb{E}\left[\sum_{k=1}^{d}\alpha_1 \hat{v}_{1,k}^{-\frac{1}{2}}\right] \leqslant \frac{\alpha d}{c}. \end{aligned} \tag{D.3}$$

Therefore the term R_2 (Eq.(C.1)) can be bounded as follows:

$$R_2 \leqslant \left(\frac{\alpha^2\bar{H}^2 C_2}{c^2}+C_4\alpha\bar{\gamma}\right)(1+\ln T)+\frac{\alpha d C_1}{c}+C_3. \tag{D.4}$$

Besides, because of $\hat{v}_t=\frac{1}{1-\beta_2^t}\left[\beta_2 v_{t-1}+(1-\beta_2)\,d_t^2\right]$ and the lemma 3, then

$$\frac{\alpha_t}{\sqrt{\hat{v}_{t,k}}} \geqslant \frac{(1-\beta_2)\,\alpha_t}{\bar{H}}=\frac{1-\beta_2}{\bar{H}}\frac{\alpha}{t^b} \geqslant \frac{\alpha\,(1-\beta_2)}{\bar{H}}\frac{1}{T^b}. \tag{D.5}$$

From Eq.(D.5), the following holds apparently:

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\left\langle\nabla f\left(x_t\right), \hat{V}_t^{-\frac{1}{2}}\nabla f\left(x_t\right)\right\rangle\right] \\ &\geqslant \frac{\alpha\,(1-\beta_2)}{\bar{H}}\frac{1}{T^b}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f\left(x_t\right)\right\|^2\right] \\ &\geqslant \frac{\alpha\,(1-\beta_2)}{\bar{H}}T^{1-b}\cdot\min_{t\in\mathcal{T}}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right]. \end{aligned} \tag{D.6}$$

The theorem 3.1, the Eq.(D.4) and Eq.(D.6) lead to

$$\frac{\alpha\,(1-\beta_2)}{\bar{H}}T^{1-b}\cdot\min_{t\in\mathcal{T}}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leqslant \left(\frac{\alpha^2\bar{H}^2 C_2}{c^2}+C_4\alpha\bar{\gamma}\right)(1+\ln T)+\frac{\alpha d C_1}{c}+C_3. \tag{D.7}$$

Let

$$Q_1 = \frac{\bar{H}}{\alpha\left(1-\beta_2\right)}, \quad Q_2 = \frac{\alpha^2 \bar{H}^2 C_2}{c^2} + C_4 \alpha \bar{\gamma}, \quad Q_3 = Q_2 + \frac{\alpha d C_1}{c} + C_3. \tag{D.8}$$

Combining Eq.(D.7) and Eq.(D.8) obtains

$$\min_{t \in \mathcal{T}} \mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leqslant \frac{Q_1}{T^{1-b}}\left(Q_2 \ln T + Q_3\right). \tag{D.9}$$

The proof is over. □