

On the $\mathcal{O}(\frac{\sqrt{d}}{T^{1/4}})$ Convergence Rate of RMSProp and Its Momentum Extension Measured by ℓ_1 Norm

Huan Li ✉

LIHUANSS@NANKAI.EDU.CN

*Institute of Robotics and Automatic Information Systems
College of Artificial Intelligence
Nankai University
Tianjin 300071, China*

Yiming Dong

YMDONG@STU.PKU.EDU.CN

*State Key Lab of General AI
School of Intelligence Science and Technology
Peking University
Beijing 100871, China*

Zhouchen Lin ✉

ZLIN@PKU.EDU.CN

*State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
Institute for Artificial Intelligence, Peking University, Beijing 100871, China
Pazhou Laboratory (Huangpu), Guangzhou 510555, China*

Editor: Prateek Jain

Abstract

Although adaptive gradient methods have been extensively used in deep learning, their convergence rates proved in the literature are all slower than that of SGD, particularly with respect to their dependence on the dimension. This paper considers the classical RMSProp and its momentum extension and establishes the convergence rate of $\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \mathcal{O}(\frac{\sqrt{d}C}{T^{1/4}})$ measured by ℓ_1 norm without the bounded gradient assumption, where d is the dimension of the optimization variable, T is the iteration number, and C is a constant identical to that appeared in the optimal convergence rate of SGD. Our convergence rate matches the lower bound with respect to all the coefficients except the dimension d . Since $\|\mathbf{x}\|_2 \ll \|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$ for problems with extremely large d , our convergence rate can be considered to be analogous to the $\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_2] \leq \mathcal{O}(\frac{C}{T^{1/4}})$ rate of SGD in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$.

Keywords: Adaptive gradient method, RMSProp, Momentum, Convergence rate

1 Introduction

This paper considers adaptive gradient methods for the following nonconvex smooth stochastic optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{P}}[h(\mathbf{x}; \xi)], \quad (1)$$

where ξ is a random variable and \mathcal{P} is the data distribution.

When evaluating the convergence speed of an optimization method, traditional optimization theories primarily focus on the dependence on the iteration number. For example, it

is well known that SGD reaches the precision of $\mathcal{O}(\frac{1}{T^{1/4}})$ after T iterations for nonconvex problem (1), disregarding the constants independent of T within $\mathcal{O}(\cdot)$. However, this measure is inadequate for high-dimensional applications, particularly in deep learning. Consider GPT-3, which possesses 175 billion parameters. In other words, in the training model (1),

$$d = 1.75 \times 10^{11} \text{ in GPT-3.}$$

If a method converges with a rate of $\mathcal{O}(\frac{d}{T^{1/4}})$, it is unrealistic since we rarely train a deep neural network for 10^{44} iterations. Therefore, it is desirable to study the explicit dependence on the dimension d and the constants relying on d in $\mathcal{O}(\cdot)$, and furthermore, to decrease this dependence.

To compound the issue, although adaptive gradient methods, such as AdaGrad (Duchi et al., 2011; McMahan and Streeter, 2010), RMSProp (Tieleman and Hinton, 2012), and Adam (Kingma and Ba, 2015), have become dominant in training deep neural networks, their convergence rates have not been thoroughly investigated, particularly with regard to their dependence on the dimension. Current analyses of convergence rates indicate that these methods often exhibit a strong dependence on the dimension. For example, recently, Hong and Lin (2024a) (see Section 3 for the detailed literature reviews) proved the following state-of-the-art convergence rate for AdaGrad with high probability

$$\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_2 \leq \mathcal{O} \left(\frac{\sqrt{d \ln T}}{T^{1/4}} \left(\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} + \sigma_s \right) \right) \quad (2)$$

under assumption $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2 \leq \sigma_s^2$, where \mathbf{g}^k represents the stochastic gradient at \mathbf{x}^k . In contrast, the convergence rate of SGD (Bottou et al., 2018) can be as fast as

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_2] \leq \mathcal{O} \left(\frac{\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)}}{T^{1/4}} \right) \quad (3)$$

with weaker assumption $\mathbb{E} [\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2] \leq \sigma_s^2$. We observe that the convergence rate of SGD is $\sqrt{d \ln T}$ times faster than (2). It remains an open problem of how to establish the convergence rate of adaptive gradient methods in a manner analogous to that of SGD, in order to bridge the gap between their rapid convergence observed in practice and their theoretically slower convergence rate compared to SGD.

Algorithm 1 RMSProp

Initialize $\mathbf{x}^1, \mathbf{v}_i^0$
for $k = 1, 2, \dots, T$ **do**
 $\mathbf{v}^k = \beta \mathbf{v}^{k-1} + (1 - \beta)(\mathbf{g}^k)^{\odot 2}$
 $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\eta}{\sqrt{\mathbf{v}^k}} \odot \mathbf{g}^k$
end for

Algorithm 2 RMSProp with Momentum

Initialize $\mathbf{x}^1, \mathbf{m}_i^0 = 0, \mathbf{v}_i^0$
for $k = 1, 2, \dots, T$ **do**
 $\mathbf{v}^k = \beta \mathbf{v}^{k-1} + (1 - \beta)(\mathbf{g}^k)^{\odot 2}$
 $\mathbf{m}^k = \theta \mathbf{m}^{k-1} + (1 - \theta) \frac{1}{\sqrt{\mathbf{v}^k}} \odot \mathbf{g}^k$
 $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \mathbf{m}^k$
end for

1.1 Contribution

In this paper, we consider the classical RMSProp and its momentum extension (Tieleman and Hinton, 2012), which are presented in Algorithms 1 and 2, respectively. Specifically, for both methods, we prove the convergence rate of

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \tilde{\mathcal{O}} \left(\frac{\sqrt{d}}{T^{1/4}} \sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} + \frac{\sqrt{d}}{\sqrt{T}} \sqrt{L(f(\mathbf{x}^1) - f^*)} \right)$$

measured by ℓ_1 norm under the assumption of coordinate-wise bounded noise variance, which does not require the boundedness of the gradient or stochastic gradient. Our convergence rate matches the lower bound established in (Arjevani et al., 2023) with respect to T , L , $f(\mathbf{x}^1) - f^*$, and σ_s . Since $L(f(\mathbf{x}^1) - f^*) \geq \frac{1}{2} \|\nabla f(\mathbf{x}^1)\|^2 = \frac{1}{2} \sum_{i=1}^d |\nabla_i f(\mathbf{x}^1)|^2$ and $\sigma_s^2 \geq \sum_{i=1}^d \mathbb{E} [|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)|^2]$, they could assume large values in high-dimensional settings¹. So it is significant to achieve the optimal dependence on $L(f(\mathbf{x}^1) - f^*)$ and σ_s . The only coefficient left unclear whether it is tight measured by ℓ_1 norm is the dimension d .

Note that $\|\mathbf{x}\|_2 \ll \|\mathbf{x}\|_1 \leq \sqrt{d} \|\mathbf{x}\|_2$ for any $\mathbf{x} \in \mathbb{R}^d$ with extremely large d , and additionally, $\|\mathbf{x}\|_1 = \Theta(\sqrt{d}) \|\mathbf{x}\|_2$ when \mathbf{x} is generated from uniform or Gaussian distribution. Therefore, our convergence rate can be considered to be analogous to (3) of SGD in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d}) \|\nabla f(\mathbf{x})\|_2$. Fortunately, as demonstrated in Figure 1, we have empirically observed that in real deep neural networks, the relationship $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d}) \|\nabla f(\mathbf{x})\|_2$ holds true.

1.2 Notations and Assumptions

Denote \mathbf{x}_i and $\nabla_i f(\mathbf{x})$ as the i th element of vectors \mathbf{x} and $\nabla f(\mathbf{x})$, respectively. Let \mathbf{x}^k represent the value at iteration k . For scalars, such as v , we use v_k instead of v^k to denote its value at iteration k , while the latter represents its k th power. Denote $\|\cdot\|$, or $\|\cdot\|_2$ if emphasis is required, as the ℓ_2 Euclidean norm and $\|\cdot\|_1$ as the ℓ_1 norm for vectors, respectively. Denote $f^* = \inf f(\mathbf{x})$. Denote \odot to stand for the Hadamard product between vectors. Denote $\mathcal{F}_k = \sigma(\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^k)$ to be the sigma field of the stochastic gradients up to k . Let $\mathbb{E}_{\mathcal{F}_k}[\cdot]$ denote the expectation with respect to \mathcal{F}_k and $\mathbb{E}_k[\cdot|\mathcal{F}_{k-1}]$ the conditional expectation with respect to \mathbf{g}^k conditioned on \mathcal{F}_{k-1} . We use $f = \mathcal{O}(g)$, $f = \Omega(g)$, and $f = \Theta(g)$ to denote $f \leq c_1 g$, $f \geq c_2 g$, and $c_2 g \leq f \leq c_1 g$ for some constants c_1 and c_2 , respectively, and $\tilde{\mathcal{O}}$ to hide polylogarithmic factors. The base of natural logarithms is denoted by e .

Throughout this paper, we make the following assumptions:

1. Smoothness: $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L \|\mathbf{y} - \mathbf{x}\|$,
2. Unbiased estimator: $\mathbb{E}_k [\mathbf{g}_i^k | \mathcal{F}_{k-1}] = \nabla_i f(\mathbf{x}^k)$,
3. Coordinate-wise bounded noise variance: $\mathbb{E}_k [|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)|^2 | \mathcal{F}_{k-1}] \leq \sigma_i^2$.

1. However, empirical observations in deep learning training indicate that each element of $\nabla f(\mathbf{x})$ tends to be very small, with both $\|\nabla f(\mathbf{x}^1)\|$ and $f(\mathbf{x}^1) - f^*$ typically being of order $\mathcal{O}(1)$.

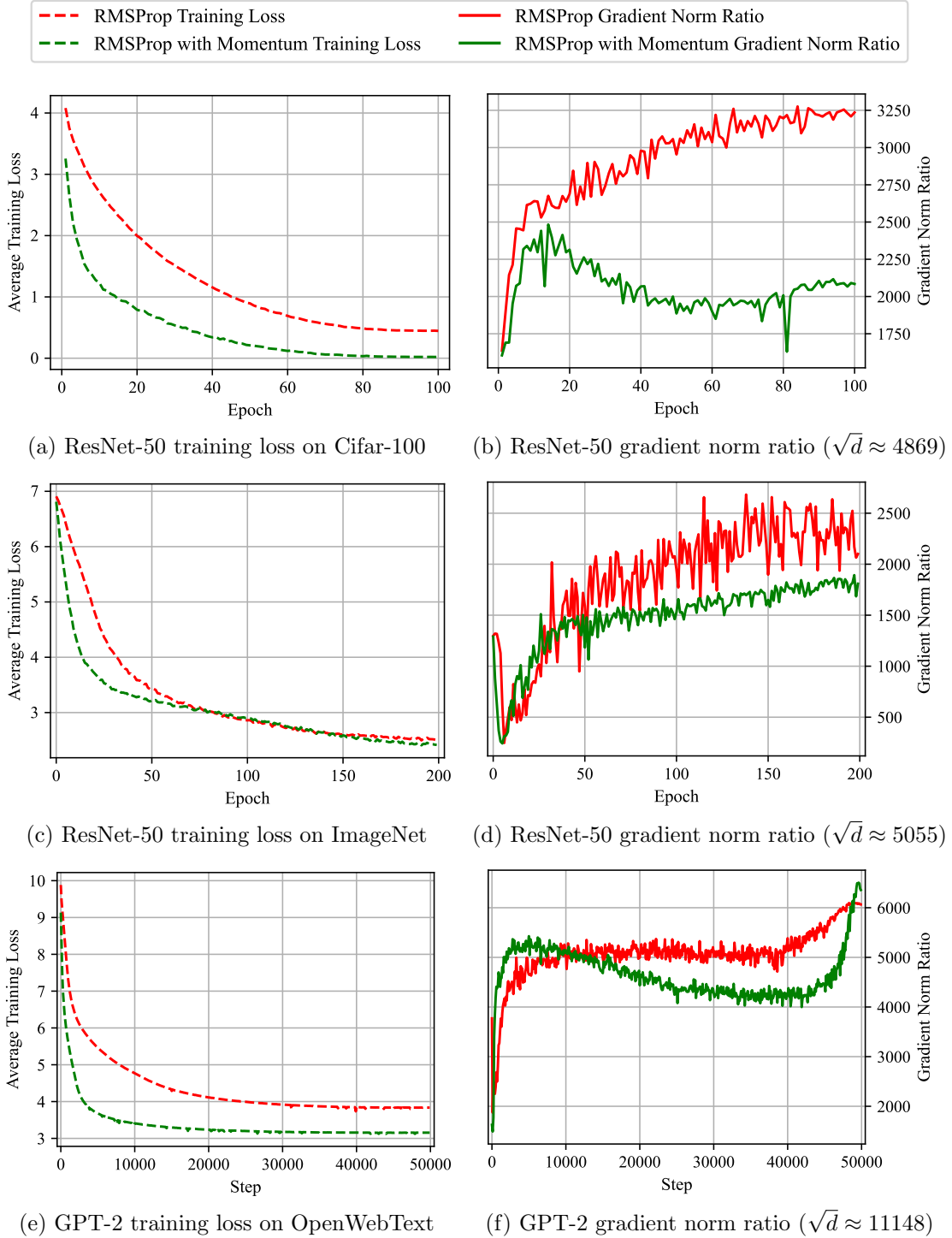


Figure 1: Illustration of the relationship $\|\nabla f(\mathbf{x}^k)\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x}^k)\|_2$. We use RMSProp and RMSProp with momentum to train ResNet50 on CIFAR-100 and ImageNet, and train GPT2 on the OpenWebText dataset. The gradient norm ratio shows $\frac{\|\nabla f(\mathbf{x}^k)\|_1}{\|\nabla f(\mathbf{x}^k)\|_2}$ and the average training loss shows the average loss over training samples.

Denoting $\boldsymbol{\sigma} = [\sigma_1, \dots, \sigma_d]$ and $\sigma_s = \|\boldsymbol{\sigma}\|_2 = \sqrt{\sum_i \sigma_i^2}$, we have the standard bounded noise variance assumption

$$\mathbb{E}_k \left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2 | \mathcal{F}_{k-1} \right] \leq \sigma_s^2,$$

which is used in the analysis of SGD. Note that we do not assume the boundedness of $\nabla f(\mathbf{x}^k)$ or \mathbf{g}^k .

2 Convergence Rates of RMSProp and Its Momentum Extension

In this section, we prove the convergence rates of the classical RMSProp and its momentum extension. Both methods are implemented in PyTorch by the following API with broad applications in deep learning:

`torch.optim.RMSprop(lr, ..., momentum, ...),`

where `momentum` and `lr` equal θ and $(1 - \theta)\eta$ in Algorithm 2, respectively. Specially, If we set `momentum=0` in default, it reduces to RMSProp.

We establish the convergence rate of RMSProp with momentum in the following theorem. Additionally, if we set $\theta = 0$, Theorem 1 also provides the convergence rate of RMSProp. For brevity, we omit the details.

Theorem 1 *Suppose that Assumptions 1-3 hold. Let $\eta = \frac{\gamma}{\sqrt{dT}}$, $\beta = 1 - \frac{1}{T}$, $\mathbf{v}_i^0 = \lambda \max \{ \sigma_i^2, \frac{1}{dT} \}$, $\forall i$, and $T \geq \frac{e^2}{\lambda}$, where $\theta \in [0, 1)$, $\lambda \leq 1$, and γ can be any constants serving as hyper-parameters for tuning performance in practice. Then for Algorithm 2, we have*

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E} \left[\|\nabla f(\mathbf{x}^k)\|_1 \right] \leq \frac{d^{1/4}}{T^{1/4}} \sqrt{\frac{2F\|\boldsymbol{\sigma}\|_1}{\gamma}} + \frac{\sqrt{d}}{\sqrt{T}} \frac{4F}{\gamma},$$

where

$$\begin{aligned} \frac{F}{\gamma} = \max \left\{ 1, \quad & 3(2L\gamma + 3) \ln(2L\gamma + 3), \quad \frac{3(f(\mathbf{x}^1) - f^*)}{\gamma}, \right. \\ & 3 \left(\frac{6e\sigma_s}{\sqrt{\lambda T}} + \frac{3L\gamma}{(1-\theta)^{1.5}} + 3 \right) \ln \left(\frac{6e\sigma_s}{\sqrt{\lambda T}} + \frac{3L\gamma}{(1-\theta)^{1.5}} + 3 \right), \\ & \left. \left(\frac{6e\sigma_s}{\sqrt{\lambda T}} + \frac{3L\gamma}{(1-\theta)^{1.5}} \right) \ln \left(\frac{4L\gamma e^2}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(1 + \frac{\theta^2}{2T(1-\theta)^2} \right) + \frac{12}{\lambda} \right) \right\}. \end{aligned} \quad (4)$$

Since θ is a constant independent of T and d , we can simplify Theorem 1 in the following corollary.

Corollary 2 *Under the settings of Theorem 1, letting $\gamma = \sqrt{\frac{f(\mathbf{x}^1) - f^*}{L}}$ and $T \geq \frac{\sigma_s^2}{\lambda L(f(\mathbf{x}^1) - f^*)}$, we have $\frac{F}{\gamma} = \tilde{O} \left(\sqrt{L(f(\mathbf{x}^1) - f^*)} \right)$ and*

$$\begin{aligned} \frac{1}{T} \sum_{k=1}^T \mathbb{E} \left[\|\nabla f(\mathbf{x}^k)\|_1 \right] &\leq \tilde{O} \left(\frac{d^{1/4}}{T^{1/4}} \sqrt{\|\boldsymbol{\sigma}\|_1^2 L(f(\mathbf{x}^1) - f^*)} + \frac{\sqrt{d}}{\sqrt{T}} \sqrt{L(f(\mathbf{x}^1) - f^*)} \right) \\ &\leq \tilde{O} \left(\frac{\sqrt{d}}{T^{1/4}} \sqrt{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} + \frac{\sqrt{d}}{\sqrt{T}} \sqrt{L(f(\mathbf{x}^1) - f^*)} \right), \end{aligned}$$

where $\tilde{\mathcal{O}}$ hides $\ln(L(f(\mathbf{x}^1) - f^*))$ and $\ln\left(\frac{\sqrt{L(f(\mathbf{x}^1) - f^*)}}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} + \frac{1}{\lambda}\right)$. If $T \geq \frac{dL(f(\mathbf{x}^1) - f^*)}{\|\boldsymbol{\sigma}\|_1^2}$, the first term dominates and the convergence rate becomes

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \tilde{\mathcal{O}} \left(\frac{d^{1/4}}{T^{1/4}} \sqrt{\|\boldsymbol{\sigma}\|_1^2 L(f(\mathbf{x}^1) - f^*)} \right) \leq \tilde{\mathcal{O}} \left(\frac{\sqrt{d}}{T^{1/4}} \sqrt{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} \right). \quad (5)$$

On the other hand, if the noise variance is small, that is, $\|\boldsymbol{\sigma}\|_1^2 \leq \frac{dL(f(\mathbf{x}^1) - f^*)}{T}$, the second term dominates and the convergence rate becomes

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \tilde{\mathcal{O}} \left(\frac{\sqrt{d}}{\sqrt{T}} \sqrt{L(f(\mathbf{x}^1) - f^*)} \right).$$

In deep learning with extremely large d , it can be expected that $d \min_i \sigma_i^2 > \frac{1}{T}$, making it unlikely for $\ln T$ to appear in $\tilde{\mathcal{O}}$.

Tightness with respect to the coefficients. Arjevani et al. (2023) established the lower bound of stochastic optimization methods under the assumptions of smoothness and bounded noise variance. The convergence rate of SGD in (3) matches this lower bound. By comparing our convergence rate (5) with (3), we observe that our convergence rate is also tight up to logarithmic factors with respect to the smoothness coefficient L , the initial function value gap $f(\mathbf{x}^1) - f^*$, the noise variance σ_s , and the iteration number T . The only coefficient left unclear whether it is tight measured by ℓ_1 norm is the dimension d .

Comparison to SGD. Our convergence rate (5) can be considered to be analogous to (3) of SGD in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$. Fortunately, we have empirically observed that the relationship $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$ holds true in common deep neural networks, as shown in Figures 1. On the other hand, our theory relies on slightly stronger assumption of coordinate-wise bounded noise variance compared to SGD. Consequently, the constant $\sigma_s = \sqrt{\sum_i \sigma_i^2}$ is larger than the one used in SGD. Nonetheless, if each $\mathbb{E}_k [\|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)\|^2 | \mathcal{F}_{k-1}]$ does not oscillate intensely during iterations, we may expect the two constants not to differ greatly.

Two key points in our proof. To establish the tight dependence on $L(f(\mathbf{x}^1) - f^*)$, we should upper bound $\sum_{i=1}^d \sum_{k=1}^T \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right]$ (see the definition in (10)) by σ_s (or $\|\boldsymbol{\sigma}\|_1$) predominantly instead of $L(f(\mathbf{x}^1) - f^*)$. To address this issue, we provide a simple proof in Lemma 8 to bound $\sum_{i=1}^d \sum_{k=1}^T \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right]$ by $\mathcal{O} \left(T\|\boldsymbol{\sigma}\|_1 + \frac{F}{\gamma} \sqrt{dT} \right)$, with the first term dominating. Additionally, to ensure the tight dependence on σ_s , we give a sharper upper bound for the error term in Lemma 4, such that $\frac{F}{\gamma}$ in (4) includes $\frac{\sigma_s}{\sqrt{\lambda T}}$ instead of just σ_s , where the former can be relaxed as $\frac{\sigma_s}{\sqrt{\lambda T}} \leq \sqrt{L(f(\mathbf{x}^1) - f^*)}$ by setting $T \geq \frac{\sigma_s^2}{\lambda L(f(\mathbf{x}^1) - f^*)}$.

ℓ_1 norm or ℓ_2 norm. The choice to utilize the ℓ_1 norm is motivated by SignSGD (Bernstein et al., 2018), which is closely related to Adam (Balles and Hennig, 2018). Technically, if we were to use the ℓ_2 norm as conventionally done, we would need to make the right hand side of the convergence rate criteria independent of d while remaining the other coefficients tight, as shown in (3), to make it no slower than SGD. However, achieving this target presents

a challenge with current techniques. Instead, by using the ℓ_1 norm, we can maintain the term \sqrt{d} on the right hand side, as demonstrated in (5), since $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$ in the ideal case.

Relation to AdaGrad. From the parameter settings in Corollary 2, we have $\eta = \frac{\gamma}{\sqrt{dT}}$, $1 - \beta = \frac{1}{T}$, $\mathbf{v}_i^0 \geq \lambda \sigma_i^2 \geq \Omega(\frac{\sigma_i^2}{T})$, and $\frac{1}{e^2} \leq \beta^t \leq 1$ for any $t \leq T$ from (19). So for the update direction of \mathbf{x}_i^k , we have

$$\eta \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} = \eta \frac{\mathbf{g}_i^k}{\sqrt{\beta^k \mathbf{v}_i^0 + (1 - \beta) \sum_{t=1}^k \beta^{k-t} |\mathbf{g}_i^t|^2}} \approx \frac{\gamma}{\sqrt{dT}} \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^0 + \frac{\sum_{t=1}^k |\mathbf{g}_i^t|^2}{T}}} \approx \frac{\gamma}{\sqrt{d}} \frac{\mathbf{g}_i^k}{\sqrt{\sigma_i^2 + \sum_{t=1}^k |\mathbf{g}_i^t|^2}}.$$

On the other hand, the update direction of \mathbf{x}_i^k in AdaGrad is $\eta \frac{\mathbf{g}_i^k}{\sqrt{\sum_{t=1}^k |\mathbf{g}_i^t|^2 + \varepsilon}}$. So in our parameter settings, RMSProp can be regarded as a refined variant of AdaGrad.

3 Literature Comparisons

It is not easy to compare convergence rates in the literature due to variations in assumptions. Furthermore, most literature does not state explicit dependence on the dimension in their theorems, and instead hides it within the proofs. In this section, we attempt to compare our convergence rate with the representative ones in the literature. Particularly, we primarily compare with the ones without the bounded gradient assumption.

3.1 Convergence Rate of AdaGrad in (Hong and Lin, 2024a)

Hong and Lin (2024a, Corollary 1) studied AdaGrad under the relaxed noise assumption of $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2 \leq A(f(\mathbf{x}^k) - f^*) + B\|\nabla f(\mathbf{x}^k)\|^2 + C$ with probability 1, and their result can be extended to the sub-Gaussian assumption where $\mathbb{E} \left[\exp\left(\frac{\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2}{A(f(\mathbf{x}^k) - f^*) + B\|\nabla f(\mathbf{x}^k)\|^2 + C}\right) \right] \leq e$. We compare with their convergence rate by setting $C = \sigma_s^2$ and $A = B = 0$. They proved the following convergence rate with high probability,

$$\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_2^2 \leq \mathcal{O} \left(\Delta_1 \left(\frac{\Delta_1 + \sqrt{L\eta\Delta_1}}{T} + \frac{\sigma_s}{\sqrt{T}} \right) \right) = \mathcal{O} \left(\frac{\sigma_s \Delta_1}{\sqrt{T}} \right),$$

where $\Delta_1 = \mathcal{O} \left(\frac{f(\mathbf{x}^1) - f^*}{\eta} + d\sigma_s \ln T + L\eta d^2 \ln^2 T \right)$.

Δ_1 is minimized to be $\tilde{\mathcal{O}} \left(\left(\sqrt{L(f(\mathbf{x}^1) - f^*)} + \sigma_s \right) d \ln T \right)$ by letting $\eta = \sqrt{\frac{f(\mathbf{x}^1) - f^*}{L}} \frac{1}{d \ln T}$. Accordingly, their convergence rate is

$$\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_2 \leq \mathcal{O} \left(\frac{\sqrt{d \ln T}}{T^{1/4}} \left(\sqrt{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} + \sigma_s \right) \right),$$

which is $\sqrt{d \ln T}$ times slower than (3) of SGD. It is also inferior to our convergence rate (5) due to $\|\mathbf{x}\|_2 \ll \|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$. Additionally, their dependence on σ_s is not optimal, especially when $\sigma_s \geq \sqrt{L(f(\mathbf{x}^1) - f^*)}$. Hong and Lin also studied Adam in (Hong and Lin, 2024b, 2023),

but the convergence rate is not superior to that of AdaGrad. It should be noted that Hong and Lin (2024a) established this result based on the assumption that $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2 \leq \sigma_s^2$ with probability 1, or the sub-Gaussian assumption of $\mathbb{E} \left[\exp\left(\frac{\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2}{\sigma_s^2}\right) \right] \leq e$. In contrast, our assumption is $\mathbb{E}_k [\|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)\|^2 | \mathcal{F}_{k-1}] \leq \sigma_i^2$ in the coordinate-wise manner. Determining which assumption is stronger is difficult.

3.2 Convergence Rate of AdaGrad in (Liu et al., 2023)

Liu et al. (2023, Theorem 4.6) studied AdaGrad under the coordinate-wise sub-Gaussian assumption of $\mathbb{E}[\exp(\lambda^2 |\mathbf{g}_i - \nabla_i f(\mathbf{x})|^2)] \leq \exp(\lambda^2 \sigma_i^2), \forall |\lambda| \leq \frac{1}{\sigma_i}$. Liu et al. (2023) also used ℓ_1 norm to measure the convergence rate. Specifically, from Theorem 4.6 and the corresponding proof on page 42 in (Liu et al., 2023), they proved the following convergence rate with probability at least $1 - \delta$,

$$\begin{aligned} \frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_1^2 &\leq g(\delta) \mathcal{O} \left(\frac{\|\boldsymbol{\sigma}\|_1}{\sqrt{T}} + \frac{r(\delta)}{T} \right), \\ \text{where } g(\delta) &= \mathcal{O} \left(\frac{f(\mathbf{x}^1) - f^*}{\eta} + \left(d \|\boldsymbol{\sigma}\|_\infty + \sum_{i=1}^d c_i(\delta) \right) \sqrt{\log \frac{dT}{\delta}} + dL\eta \log \left(\|\boldsymbol{\sigma}\|_1 \sqrt{T} + r(\delta) \right) \right), \\ c_i(\delta) &= \mathcal{O} \left(\sigma_i^3 \log \frac{d}{\delta} + \sigma_i \log \left(1 + \sigma_i^2 T + \sigma_i^2 \log \frac{d}{\delta} \right) + \|\boldsymbol{\sigma}\|_1 \log(\|\boldsymbol{\sigma}\|_1 \sqrt{T} + r(\delta)) \right), \\ r(\delta) &= \mathcal{O} \left(f(\mathbf{x}^1) - f^* + \|\boldsymbol{\sigma}^2\|_1 \log \frac{d}{\delta} + \|\boldsymbol{\sigma}\|_1 \sqrt{\log \frac{d}{\delta}} + Ld \log L \right). \end{aligned}$$

It is not easy to simplify the above convergence rate and the dependence on $\boldsymbol{\sigma}$ is not optimal due to the second term in $g(\delta)$. Ignoring $\sum_{i=1}^d c_i(\delta)$ and the logarithmic term in $g(\delta)$, their $\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_1$ is upper bounded by a constant not less than $\frac{\sqrt{d} \|\boldsymbol{\sigma}\|_\infty \|\boldsymbol{\sigma}\|_1 + \sqrt[4]{\|\boldsymbol{\sigma}\|_1^2 d L (f(\mathbf{x}^1) - f^*)}}{T^{1/4}}$, which is inferior to our convergence rate (5).

3.3 Convergence Rate of RMSProp in (Shi et al., 2020)

Shi et al. (2020, Theorem 4.3) studied problem $\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{j=0}^{n-1} f_j(\mathbf{x})$ and they assumed $\sum_{j=0}^{n-1} \|\nabla f_j(\mathbf{x})\|_2^2 \leq D_1 \|\nabla f(\mathbf{x})\|_2^2 + D_0$. They did not give the explicit dependence on the dimension in their theorem 4.3 and we try to recover it from their proof. On page 37 in (Shi et al., 2020), the authors gave

$$\min_{k \in [t_{init}, T]} \min \left\{ \|\nabla f(\mathbf{x}^{k,0})\|_1, \|\nabla f(\mathbf{x}^{k,0})\|_2^2 \sqrt{\frac{D_1 d}{D_0}} \right\} \leq \frac{Q_{1,3} + Q_{2,3} \log T}{\sqrt{T} - \sqrt{t_{init} - 1}} + \sqrt{D_0} Q_{3,3},$$

$$\text{where } Q_{1,3} = \frac{f(\mathbf{x}^{t_{init},0}) - f^* - C_6 \log t_{init}}{2\eta_1} \sqrt{10dnD_1d}, \quad Q_{2,3} = \frac{C_6 \sqrt{10dnD_1d}}{2\eta_1},$$

$$C_6 = L\eta_1^2 \left(\frac{nd}{2(1-\beta_2)} + \frac{C_4 \sqrt{d}}{n\sqrt{1-\beta_2}} \right), \quad C_4 \geq \frac{dn^2}{(1-\beta_2)^{1.5}}, \quad 1 - \beta_2 \leq \mathcal{O} \left(\frac{1}{n^{3.5}} \right),$$

and the other notations can be found in their proof. Since

$$\begin{aligned} Q_{1,3} + Q_{2,3} \log T &\geq \Omega \left(\left(\frac{f(\mathbf{x}^{t_{init},0}) - f^*}{\eta_1} + \frac{L\eta_1 n d^{1.5}}{(1 - \beta_2)^2} (\log T - \log t_{init}) \right) \sqrt{dnD_1d} \right) \\ &\geq \Omega \left(\sqrt{D_1 L (f(\mathbf{x}^{t_{init},0}) - f^*) d^{9/4}} \right). \end{aligned}$$

We see that their $\min_k \|\nabla f(\mathbf{x}^{k,0})\|_2^2$ is upper bounded by a constant not less than $\tilde{\Theta} \left(\frac{D_0 Q_{3,3}}{\sqrt{dD_1}} + \frac{d^{7/4} \sqrt{D_0 L (f(\mathbf{x}^{t_{init},0}) - f^*)}}{\sqrt{T}} \right)$. That is, $\min_k \|\nabla f(\mathbf{x}^{k,0})\|_2$ is upper bounded by a constant not less than $\tilde{\mathcal{O}} \left(\frac{\sqrt{D_0 Q_{3,3}}}{\sqrt{dD_1}} + \frac{d^{7/8} \sqrt{D_0 L (f(\mathbf{x}^{t_{init},0}) - f^*)}}{T^{1/4}} \right)$, which is at least $d^{3/8}$ times slower than our convergence rate (5).

3.4 Convergence Rate of RMSProp in (Défossez et al., 2022)

Défossez et al. (2022, Theorem 2) studied the convergence rate of RMSProp under the bounded stochastic gradient assumption, that is, there is $R \geq \sqrt{\varepsilon}$ so that $\|\mathbf{g}^k\|_\infty \leq R - \sqrt{\varepsilon}$ almost surely. They proved the following bound for RMSProp

$$\mathbb{E} [\|\nabla f(\mathbf{x}^\tau)\|_2^2] \leq \mathcal{O} \left(R \frac{f(\mathbf{x}^1) - f^*}{\eta T} + \left(\frac{dR^2}{\sqrt{1 - \beta}} + \frac{\eta dRL}{1 - \beta} \right) \left(\frac{1}{T} \ln \left(1 + \frac{R^2}{(1 - \beta)\varepsilon} \right) - \ln \beta \right) \right).$$

Letting $\beta = 1 - \frac{1}{T}$ and $\eta = \frac{1}{\sqrt{dT}} \sqrt{\frac{f(\mathbf{x}^1) - f^*}{L}}$, the above convergence rate can be simplified to

$$\mathbb{E} [\|\nabla f(\mathbf{x}^\tau)\|_2^2] \leq \mathcal{O} \left(\left(\frac{\sqrt{d}}{\sqrt{T}} R \sqrt{L(f(\mathbf{x}^1) - f^*)} + \frac{dR^2}{\sqrt{T}} \right) \ln(RT) \right),$$

where $dR^2 \geq \|\mathbf{g}^k\|_2^2, \forall k$. Due to the bounded stochastic gradient assumption, comparing their convergence rate with ours is unfair. Our proof follows the analytical framework in (Défossez et al., 2022). However, unlike their work, we cannot lower bound $\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}}$ by $\frac{|\nabla_i f(\mathbf{x}^k)|^2}{R}$ without the bounded stochastic gradient assumption. Instead, we use

$$\left(\sum_{k=1}^K \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \right)^2 \leq \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \right) \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E} [\sqrt{\tilde{\mathbf{v}}_i^k}] \right)$$

and rigorously derive a tight upper bound for $\sum_{k=1}^K \sum_{i=1}^d \mathbb{E} [\sqrt{\tilde{\mathbf{v}}_i^k}]$. Furthermore, in the absence of the bounded stochastic gradient assumption, we cannot use $\mathbf{v}_i^k \leq R^2$ to upper bound $\sum_{k=1}^K \mathbb{E} \left[\frac{\|\mathbf{g}_i^k\|^2}{\mathbf{v}_i^k} \right]$. To address this, we employ mathematical induction to establish the bound in (14). Additionally, we provide a sharper upper bound for the error term in Lemma 4 to ensure the tight dependence on σ_s .

3.5 Convergence Rate of Adam in (Li et al., 2023)

Li et al. (2023, Theorem 4.1) introduced a new proof of boundedness of gradients along the optimization trajectory. Although their Theorem 4.1 has no explicit dependence on d , it has a higher dependence on $L(f(\mathbf{x}^1) - f^*)$, σ_s and the constant λ as a compromise, where λ appears in the adaptive step-size $\frac{\eta}{\sqrt{\mathbf{v}^k} + \lambda}$, which is usually small in practice, for example, $\lambda = 10^{-8}$ in PyTorch implementation. Specifically, they assumed $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\| \leq \sigma_s$ with probability 1 and proved $\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_2^2 \leq \epsilon^2$ with high probability by letting

$$T = \max \left\{ \frac{1}{\beta^2}, \frac{G(f(\mathbf{x}^1) - f^*)}{\eta \epsilon^2} \right\}, \quad \eta \leq \min \left\{ \frac{\sigma_s \lambda \beta}{LG}, \frac{\lambda^{3/2} \beta}{L\sqrt{G}} \right\}, \quad \beta \leq \mathcal{O} \left(\frac{\lambda \epsilon^2}{\sigma_s^2 G} \right),$$

and G to be a large constant satisfying $G \geq \max\{\lambda, \sigma_s, \sqrt{L(f(\mathbf{x}^1) - f^*)}\}$. From their setting, we see that $T \geq \frac{G^{2.5} \sigma_s^2 L(f(\mathbf{x}^1) - f^*)}{\lambda^{2.5} \epsilon^4}$. Consequently, their $\frac{1}{T} \sum_{k=1}^T \|\nabla f(\mathbf{x}^k)\|_2$ is upper bounded by a constant not less than $(\frac{G}{\lambda})^{5/8} \frac{\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)}}{T^{1/4}}$, which is at least $(\frac{G}{\lambda})^{5/8}$ times slower than SGD. It is not easy to compare with our convergence rate (5) due to different measurement. When $\|\nabla f(\mathbf{x})\|_1 \geq \Omega((\frac{\lambda}{G})^{5/8} \sqrt{d}) \|\nabla f(\mathbf{x})\|_2$, our convergence rate is superior, and in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d}) \|\nabla f(\mathbf{x})\|_2$, our convergence rate is also $(\frac{G}{\lambda})^{5/8}$ times faster.

3.6 Other works

There are other literature that analyze adaptive gradient methods, including (Ward et al., 2020; Kavis et al., 2022; Faw et al., 2022; Wang et al., 2023b; Attia and Koren, 2023) for AdaGrad-norm, (Wang et al., 2023b) for AdaGrad, (Zou et al., 2019; Défossez et al., 2022) for RMSProp, (Reddi et al., 2018; Zou et al., 2019; Défossez et al., 2022; Guo et al., 2021; Chen et al., 2022; Zhang et al., 2022; Wang et al., 2023a; Hong and Lin, 2023, 2024b; Zhang et al., 2025) for Adam, and (Zaheer et al., 2018; Loshchilov and Hutter, 2018; Chen et al., 2019; Luo et al., 2019; You et al., 2019; Zhuang et al., 2020; Chen et al., 2021; Savarese et al., 2021; Crawshaw et al., 2022; Xie et al., 2024) for other variants. However, none have established a convergence rate comparable to that of SGD.

4 Proof of Theorem 1

Denote $\mathbf{x}^0 = \mathbf{x}^1$, which corresponds to $\mathbf{m}^0 = 0$. The second and third steps of Algorithm 2 can be rewritten in the heavy-ball style equivalently as follows,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\eta(1-\theta)}{\sqrt{\mathbf{v}^k}} \odot \mathbf{g}^k + \theta(\mathbf{x}^k - \mathbf{x}^{k-1}), \forall k \geq 1, \quad (6)$$

which leads to

$$\mathbf{x}^{k+1} - \theta \mathbf{x}^k = \mathbf{x}^k - \theta \mathbf{x}^{k-1} - \frac{\eta(1-\theta)}{\sqrt{\mathbf{v}^k}} \odot \mathbf{g}^k.$$

We follow (Liu et al., 2020) to define

$$\mathbf{z}^k = \frac{1}{1-\theta} \mathbf{x}^k - \frac{\theta}{1-\theta} \mathbf{x}^{k-1}, \forall k \geq 1. \quad (7)$$

Specially, we have $\mathbf{z}^1 = \mathbf{x}^1$ since $\mathbf{x}^1 = \mathbf{x}^0$. Thus, we have

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{\eta}{\sqrt{\mathbf{v}^k}} \odot \mathbf{g}^k, \quad (8)$$

and

$$\mathbf{z}^k - \mathbf{x}^k = \frac{\theta}{1-\theta}(\mathbf{x}^k - \mathbf{x}^{k-1}). \quad (9)$$

We follow (Défossez et al., 2022; Faw et al., 2022) to define

$$\tilde{\mathbf{v}}_i^k = \beta \mathbf{v}_i^{k-1} + (1-\beta) \left(|\nabla_i f(\mathbf{x}^k)|^2 + \sigma_i^2 \right). \quad (10)$$

Then with the supporting lemmas in Section 4.1 we can prove Theorem 1.

Proof As the gradient is L -Lipschitz, we have

$$\begin{aligned} \mathbb{E}_k \left[f(\mathbf{z}^{k+1}) | \mathcal{F}_{k-1} \right] - f(\mathbf{z}^k) &\leq \mathbb{E}_k \left[\left\langle \nabla f(\mathbf{z}^k), \mathbf{z}^{k+1} - \mathbf{z}^k \right\rangle + \frac{L}{2} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 | \mathcal{F}_{k-1} \right] \\ &= \mathbb{E}_k \left[-\eta \sum_{i=1}^d \left\langle \nabla_i f(\mathbf{z}^k), \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle + \frac{L\eta^2}{2} \sum_{i=1}^d \frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} | \mathcal{F}_{k-1} \right], \end{aligned}$$

where we use (8). Decomposing the first term into

$$-\left\langle \nabla_i f(\mathbf{x}^k), \frac{\mathbf{g}_i^k}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right\rangle + \left\langle \nabla_i f(\mathbf{x}^k), \frac{\mathbf{g}_i^k}{\sqrt{\tilde{\mathbf{v}}_i^k}} - \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle + \left\langle \nabla_i f(\mathbf{x}^k) - \nabla_i f(\mathbf{z}^k), \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle$$

and using Assumption 2, we have

$$\begin{aligned} \mathbb{E}_k \left[f(\mathbf{z}^{k+1}) | \mathcal{F}_{k-1} \right] - f(\mathbf{z}^k) &\leq -\eta \sum_{i=1}^d \frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{L\eta^2}{2} \sum_{i=1}^d \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} | \mathcal{F}_{k-1} \right] \\ &\quad + \underbrace{\eta \sum_{i=1}^d \mathbb{E}_k \left[\left\langle \nabla_i f(\mathbf{x}^k), \frac{\mathbf{g}_i^k}{\sqrt{\tilde{\mathbf{v}}_i^k}} - \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle | \mathcal{F}_{k-1} \right]}_{\text{term (a)}} \\ &\quad + \underbrace{\eta \sum_{i=1}^d \mathbb{E}_k \left[\left\langle \nabla_i f(\mathbf{x}^k) - \nabla_i f(\mathbf{z}^k), \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle | \mathcal{F}_{k-1} \right]}_{\text{term (b)}}. \end{aligned} \quad (11)$$

We can use Lemma 4 to bound term (a). For term (b), we have

$$\begin{aligned}
 & \eta \sum_{i=1}^d \left\langle \nabla_i f(\mathbf{x}^k) - \nabla_i f(\mathbf{z}^k), \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle \\
 & \leq \frac{(1-\theta)^{0.5}}{2L\theta^{0.5}} \sum_{i=1}^d |\nabla_i f(\mathbf{x}^k) - \nabla_i f(\mathbf{z}^k)|^2 + \frac{L\theta^{0.5}\eta^2}{2(1-\theta)^{0.5}} \sum_{i=1}^d \frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \\
 & = \frac{(1-\theta)^{0.5}}{2L\theta^{0.5}} \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|^2 + \frac{L\theta^{0.5}\eta^2}{2(1-\theta)^{0.5}} \sum_{i=1}^d \frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \\
 & \stackrel{(1)}{\leq} \frac{L\theta^{1.5}\eta^2}{2(1-\theta)^{0.5}} \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} + \frac{L\theta^{0.5}\eta^2}{2(1-\theta)^{0.5}} \sum_{i=1}^d \frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \\
 & = \frac{L\sqrt{\theta}\eta^2}{2\sqrt{1-\theta}} \left(\sum_{t=1}^{k-1} \theta^{k-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} + \sum_{i=1}^d \frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right) = \frac{L\sqrt{\theta}\eta^2}{2\sqrt{1-\theta}} \sum_{t=1}^k \theta^{k-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t},
 \end{aligned}$$

where we use Lemma 5 in $\stackrel{(1)}{\leq}$. Plugging the above inequality and Lemma 4 into (11), taking expectation on \mathcal{F}_{k-1} , and rearranging the terms, we have

$$\begin{aligned}
 & \mathbb{E}_{\mathcal{F}_k} [f(\mathbf{z}^{k+1})] - \mathbb{E}_{\mathcal{F}_{k-1}} [f(\mathbf{z}^k)] + \frac{\eta}{2} \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \\
 & \leq \frac{2\eta e(1-\beta)}{\sqrt{\lambda}} \sum_{i=1}^d \sigma_i \mathbb{E}_{\mathcal{F}_k} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right] + \frac{L\eta^2}{\sqrt{1-\theta}} \sum_{t=1}^k \theta^{k-t} \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right].
 \end{aligned}$$

Summing over $k = 1, \dots, K$, we have

$$\begin{aligned}
 & \mathbb{E}_{\mathcal{F}_K} [f(\mathbf{z}^{K+1})] - f^* + \frac{\eta}{2} \sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \\
 & \leq f(\mathbf{z}^1) - f^* + \frac{2\eta e(1-\beta)}{\sqrt{\lambda}} \sum_{i=1}^d \sigma_i \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_k} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right] + \frac{L\eta^2}{\sqrt{1-\theta}} \sum_{i=1}^d \sum_{k=1}^K \sum_{t=1}^k \theta^{k-t} \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] \quad (12) \\
 & \leq f(\mathbf{z}^1) - f^* + \underbrace{\frac{2\eta e(1-\beta)}{\sqrt{\lambda}} \sum_{i=1}^d \sigma_i \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_k} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right]}_{\text{term (c)}} + \underbrace{\frac{L\eta^2}{(1-\theta)^{1.5}} \sum_{i=1}^d \sum_{t=1}^K \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right]}_{\text{term (d)}},
 \end{aligned}$$

where we use

$$\sum_{i=1}^d \sum_{k=1}^K \sum_{t=1}^k \theta^{k-t} \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] = \sum_{i=1}^d \sum_{t=1}^K \sum_{k=t}^K \theta^{k-t} \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] \leq \frac{1}{1-\theta} \sum_{i=1}^d \sum_{t=1}^K \mathbb{E}_{\mathcal{F}_t} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right].$$

Using Lemmas 6 and 7 to bound terms (c) and (d), respectively, letting $\eta = \frac{\gamma}{\sqrt{dT}}$ and $\beta = 1 - \frac{1}{T}$, we have

$$\begin{aligned}
 & \mathbb{E}_{\mathcal{F}_K} [f(\mathbf{z}^{K+1})] - f^* + \frac{\eta}{2} \sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\widehat{\mathbf{v}}_i^k}} \right] \\
 & \leq f(\mathbf{z}^1) - f^* + \left(\frac{2\eta e \sqrt{d} \sigma_s}{\sqrt{\lambda}} + \frac{L\eta^2 d}{(1-\theta)^{1.5}(1-\beta)} \right) \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) \right. \right. \\
 & \quad \left. \left. + \frac{L\theta^2 \eta^2}{2(1-\theta)^2} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + e + 1 \right) \quad (13) \\
 & = f(\mathbf{z}^1) - f^* + \left(\frac{2e\gamma \sigma_s}{\sqrt{\lambda T}} + \frac{L\gamma^2}{(1-\theta)^{1.5}} \right) \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) \right. \right. \\
 & \quad \left. \left. + \frac{L\theta^2 \gamma^2}{2(1-\theta)^2} \frac{1-\beta}{d} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + e + 1 \right).
 \end{aligned}$$

Next, we bound the right hand side of (13) by the constant F defined in (4). Specifically, we will prove

$$\mathbb{E}_{\mathcal{F}_{k-1}} [f(\mathbf{z}^k)] - f^* \leq F \text{ and } \frac{1-\beta}{d} \sum_{i=1}^d \sum_{t=1}^{k-1} \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] \leq \frac{F}{L\gamma^2} \quad (14)$$

by induction. (14) holds for $k = 1$ from the definition of F in (4), $\mathbf{x}^1 = \mathbf{z}^1$, and $\sum_{i=1}^d \sum_{t=1}^{k-1} \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} = 0$ when $k = 1$ given in Lemma 5. Suppose that the two inequalities hold for all $k = 1, 2, \dots, K$. Now, we consider $k = K + 1$. From Lemma 7, we have

$$\begin{aligned}
 & \frac{1-\beta}{d} \sum_{i=1}^d \sum_{t=1}^K \mathbb{E}_{\mathcal{F}_K} \left[\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] \leq \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) \right. \right. \\
 & \quad \left. \left. + \frac{L\theta^2 \gamma^2}{2(1-\theta)^2} \frac{1-\beta}{d} \sum_{i=1}^d \sum_{t=1}^{K-1} \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + e + 1 \right) \\
 & \leq \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(KF + \frac{\theta^2}{2(1-\theta)^2} F \right) + \frac{e^2}{\lambda} + e + 1 \right) \quad (15) \\
 & \leq \ln \left(\frac{4L\gamma e^2}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(1 + \frac{\theta^2}{2T(1-\theta)^2} \right) \frac{F}{\gamma} + \frac{12}{\lambda} \frac{F}{\gamma} \right) \\
 & \stackrel{(2)}{\leq} \frac{F}{L\gamma^2},
 \end{aligned}$$

where we use $(1 - \beta)K = \frac{K}{T} \leq 1$ and let $\frac{F}{\gamma} \geq 1$. Inequality $\stackrel{(2)}{\leq}$ will be verified later. Using the similar proof to (15), we derive from (13) that

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_K} [f(\mathbf{z}^{K+1})] - f^* + \frac{\eta}{2} \sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \\ & \leq f(\mathbf{z}^1) - f^* + \left(\frac{2e\gamma\sigma_s}{\sqrt{\lambda T}} + \frac{L\gamma^2}{(1-\theta)^{1.5}} \right) \ln \left(\frac{4L\gamma e^2}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(1 + \frac{\theta^2}{2T(1-\theta)^2} \right) \frac{F}{\gamma} + \frac{12}{\lambda} \frac{F}{\gamma} \right) \quad (16) \\ & \stackrel{(3)}{\leq} F. \end{aligned}$$

We construct F for $\stackrel{(2)}{\leq}$ and $\stackrel{(3)}{\leq}$ to hold by letting

$$\begin{aligned} 1 & \leq \frac{F}{\gamma}, & \ln \left(\frac{4L\gamma e^2}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(1 + \frac{\theta^2}{2T(1-\theta)^2} \right) + \frac{12}{\lambda} \right) & \leq \frac{F}{2L\gamma^2}, \\ \ln \frac{F}{\gamma} & \leq \frac{F}{2L\gamma^2}, & f(\mathbf{z}^1) - f^* & \leq \frac{F}{3}, & \left(\frac{2e\gamma\sigma_s}{\sqrt{\lambda T}} + \frac{L\gamma^2}{(1-\theta)^{1.5}} \right) \ln \frac{F}{\gamma} & \leq \frac{F}{3}, \\ \left(\frac{2e\gamma\sigma_s}{\sqrt{\lambda T}} + \frac{L\gamma^2}{(1-\theta)^{1.5}} \right) \ln & \left(\frac{4L\gamma e^2}{\lambda \max\{d \min_i \sigma_i^2, \frac{1}{T}\}} \left(1 + \frac{\theta^2}{2T(1-\theta)^2} \right) + \frac{12}{\lambda} \right) & \leq \frac{F}{3}, \end{aligned}$$

which are satisfied by setting of F in (4), where we use $\mathbf{x}^1 = \mathbf{z}^1$ and $c \ln x \leq x$ for all $x \geq 3c \ln c$ and $c \geq 3$ proved in Appendix B. So (14) also holds for $k = K + 1$. Thus, (14) holds for all $k = 1, 2, \dots, T$ by induction.

Using Holder's inequality, Lemma 8, and (16), we have

$$\begin{aligned} & \left(\sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} [\|\nabla f(\mathbf{x}^k)\|_1] \right)^2 = \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} [|\nabla_i f(\mathbf{x}^k)|] \right)^2 \\ & \leq \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \right) \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} [\sqrt{\tilde{\mathbf{v}}_i^k}] \right) \\ & \leq \left(\sum_{k=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \right) \left(K\|\boldsymbol{\sigma}\|_1 + \sqrt{dT} + 2 \sum_{k=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right] \right) \\ & \leq \frac{2F}{\eta} \left(K\|\boldsymbol{\sigma}\|_1 + \frac{F}{\eta} + \frac{4F}{\eta} \right) = \frac{2F}{\eta} \left(K\|\boldsymbol{\sigma}\|_1 + \frac{5F}{\eta} \right) \end{aligned}$$

for all $K \leq T$, where we use $\eta = \frac{\gamma}{\sqrt{dT}}$ and $\frac{F}{\gamma} \geq 1$. So we have

$$\frac{1}{T} \sum_{k=1}^T \mathbb{E}_{\mathcal{F}_{k-1}} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \frac{1}{T} \left(\sqrt{\frac{2FT\|\boldsymbol{\sigma}\|_1}{\eta}} + \frac{4F}{\eta} \right) = \frac{d^{1/4}}{T^{1/4}} \sqrt{\frac{2F\|\boldsymbol{\sigma}\|_1}{\gamma}} + \frac{\sqrt{d}}{\sqrt{T}} \frac{4F}{\gamma}.$$

■

4.1 Supporting Lemmas

In this section, we give some technical lemmas that will be used in our analysis.

Lemma 3 (Défossez et al., 2022) *Let $v_t = \beta v_{t-1} + (1 - \beta)g_t^2$. Then we have*

$$(1 - \beta) \sum_{t=1}^k \frac{g_t^2}{v_t} \leq \ln \frac{v_k}{\beta^k v_0}.$$

The next lemma is motivated by (Défossez et al., 2022). However, the key distinction is that following the proof in (Défossez et al., 2022), we can only get $\sigma_i \sqrt{1 - \beta} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right]$ in the last component of (17), where $1 - \beta = \frac{1}{T}$. We strengthen the constant from $\sqrt{1 - \beta}$ to $1 - \beta$, which is crucial to achieve the tight dependence on σ_s in our theory.

Lemma 4 *Suppose that Assumption 3 holds. Define $\tilde{\mathbf{v}}_i^k$ as in (10). Let $\mathbf{v}_i^0 = \lambda \max\{\sigma_i^2, \frac{1}{dT}\}$, $\beta = 1 - \frac{1}{T}$, and $T \geq \frac{e^2}{\lambda} \geq 2$. Then we have*

$$\mathbb{E}_k \left[\left\langle \nabla_i f(\mathbf{x}^k), \frac{\mathbf{g}_i^k}{\sqrt{\tilde{\mathbf{v}}_i^k}} - \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle \middle| \mathcal{F}_{k-1} \right] \leq \frac{|\nabla_i f(\mathbf{x}^k)|^2}{2\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{2\sigma_i e(1 - \beta)}{\sqrt{\lambda}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right]. \quad (17)$$

Proof From the definition of $\tilde{\mathbf{v}}_i^k$, the recursion of \mathbf{v}_i^k , and the setting of \mathbf{v}_i^0 , we have

$$\tilde{\mathbf{v}}_i^k \geq \beta \mathbf{v}_i^{k-1} = \beta \left(\beta^{k-1} \mathbf{v}_i^0 + (1 - \beta) \sum_{t=1}^{k-1} \beta^{k-1-t} |\mathbf{g}_i^t|^2 \right) \geq \beta^k \mathbf{v}_i^0 \geq \frac{\lambda \sigma_i^2}{e^2}, \quad (18)$$

where we use $\beta^k \geq \frac{1}{e^2}$ for any $k \leq T$ from

$$k \ln \beta = -k \ln \frac{1}{\beta} \geq -T \frac{1 - \beta}{\beta} = -T \frac{\frac{1}{T}}{1 - \frac{1}{T}} \geq -2, \quad (19)$$

since $\ln x \leq x - 1$ for any $x > 0$. So we have

$$\begin{aligned} \left| \frac{1}{\sqrt{\tilde{\mathbf{v}}_i^k}} - \frac{1}{\sqrt{\mathbf{v}_i^k}} \right| &= \frac{|\mathbf{v}_i^k - \tilde{\mathbf{v}}_i^k|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k} (\sqrt{\mathbf{v}_i^k} + \sqrt{\tilde{\mathbf{v}}_i^k})} = (1 - \beta) \frac{||\mathbf{g}_i^k|^2 - |\nabla_i f(\mathbf{x}^k)|^2 - \sigma_i^2|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k} (\sqrt{\mathbf{v}_i^k} + \sqrt{\tilde{\mathbf{v}}_i^k})} \\ &\leq (1 - \beta) \frac{|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)| |\mathbf{g}_i^k + \nabla_i f(\mathbf{x}^k)| + \sigma_i^2}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k} (\sqrt{\mathbf{v}_i^k} + \sqrt{\tilde{\mathbf{v}}_i^k})} \\ &\stackrel{(1)}{\leq} \sqrt{1 - \beta} \frac{|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}} + \frac{e(1 - \beta)}{\sqrt{\lambda}} \frac{\sigma_i}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}}, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_k \left[\left\langle \nabla_i f(\mathbf{x}^k), \frac{\mathbf{g}_i^k}{\sqrt{\tilde{\mathbf{v}}_i^k}} - \frac{\mathbf{g}_i^k}{\sqrt{\mathbf{v}_i^k}} \right\rangle \middle| \mathcal{F}_{k-1} \right] \\ &\leq \sqrt{1 - \beta} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)| |\nabla_i f(\mathbf{x}^k)| |\mathbf{g}_i^k|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}} \middle| \mathcal{F}_{k-1} \right] + \frac{\sigma_i e(1 - \beta)}{\sqrt{\lambda}} \mathbb{E}_k \left[\frac{|\nabla_i f(\mathbf{x}^k)| |\mathbf{g}_i^k|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}} \middle| \mathcal{F}_{k-1} \right], \quad (20) \end{aligned}$$

where we use the definitions of $\tilde{\mathbf{v}}_i^k$ and \mathbf{v}_i^k and (18) in $\stackrel{(1)}{\leq}$. For the first term, we have

$$\begin{aligned} & \sqrt{1-\beta} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)| |\nabla_i f(\mathbf{x}^k)| |\mathbf{g}_i^k|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}} \middle| \mathcal{F}_{k-1} \right] \\ & \leq \frac{|\nabla_i f(\mathbf{x}^k)|^2}{4\sigma_i^2 \sqrt{\tilde{\mathbf{v}}_i^k}} \mathbb{E}_k \left[|\mathbf{g}_i^k - \nabla_i f(\mathbf{x}^k)|^2 \middle| \mathcal{F}_{k-1} \right] + \frac{\sigma_i^2(1-\beta)}{\sqrt{\tilde{\mathbf{v}}_i^k}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right] \\ & \stackrel{(2)}{\leq} \frac{|\nabla_i f(\mathbf{x}^k)|^2}{4\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{\sigma_i e(1-\beta)}{\sqrt{\lambda}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right], \end{aligned}$$

where we use Assumption 3 and (18) in $\stackrel{(2)}{\leq}$. For the second term, we have

$$\begin{aligned} \frac{\sigma_i e(1-\beta)}{\sqrt{\lambda}} \mathbb{E}_k \left[\frac{|\nabla_i f(\mathbf{x}^k)| |\mathbf{g}_i^k|}{\sqrt{\tilde{\mathbf{v}}_i^k} \sqrt{\mathbf{v}_i^k}} \middle| \mathcal{F}_{k-1} \right] & \leq \frac{|\nabla_i f(\mathbf{x}^k)|^2}{4\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{\sigma_i^2 e^2(1-\beta)^2}{\lambda \sqrt{\tilde{\mathbf{v}}_i^k}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right] \\ & \stackrel{(3)}{\leq} \frac{|\nabla_i f(\mathbf{x}^k)|^2}{4\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{\sigma_i e^3(1-\beta)^2}{\lambda^{1.5}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right] \\ & \leq \frac{|\nabla_i f(\mathbf{x}^k)|^2}{4\sqrt{\tilde{\mathbf{v}}_i^k}} + \frac{\sigma_i e(1-\beta)}{\sqrt{\lambda}} \mathbb{E}_k \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \middle| \mathcal{F}_{k-1} \right], \end{aligned}$$

where we use (18) again in $\stackrel{(3)}{\leq}$. Plugging the above two inequalities into (20), we have the conclusion. \blacksquare

The next lemma is used to bound the norm of the gradient by the function value gap and the second order term, where the latter corresponds to $\frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t}$ from the second order term in Taylor expansion.

Lemma 5 *Suppose that Assumption 1 holds. Letting $\mathbf{m}^0 = 0$, we have*

$$\begin{aligned} \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|^2 & \leq \frac{L^2 \theta^2 \eta^2}{1-\theta} \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t}, \\ \|\nabla f(\mathbf{x}^k)\|^2 & \leq 4L(f(\mathbf{z}^k) - f^*) + \frac{2L^2 \theta^2 \eta^2}{1-\theta} \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t}, \\ \sum_{k=1}^K \|\nabla f(\mathbf{x}^k)\|^2 & \leq 4L \left(\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) + \frac{L \theta^2 \eta^2}{2(1-\theta)^2} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right). \end{aligned}$$

Specially, denote $\sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} = 0$ when $K = 1$.

Proof For the first part, as the gradient is L -Lipschitz, we have

$$\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|^2 \leq L^2 \|\mathbf{x}^k - \mathbf{z}^k\|^2 \stackrel{(1)}{=} \frac{L^2 \theta^2}{(1-\theta)^2} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \stackrel{(2)}{=} \frac{L^2 \theta^2 \eta^2}{(1-\theta)^2} \|\mathbf{m}^{k-1}\|^2,$$

where we use (9) in $\stackrel{(1)}{=}$ and the update of \mathbf{x} in $\stackrel{(2)}{=}$. From the update of \mathbf{m}^k in Algorithm 2, we have

$$\mathbf{m}_i^k = \theta^k \mathbf{m}_i^0 + (1-\theta) \sum_{t=1}^k \theta^{k-t} \frac{\mathbf{g}_i^t}{\sqrt{\mathbf{v}_i^t}} = (1-\theta) \sum_{t=1}^k \theta^{k-t} \frac{\mathbf{g}_i^t}{\sqrt{\mathbf{v}_i^t}}.$$

Using the convexity of $(\cdot)^2$, we have

$$|\mathbf{m}_i^k|^2 \leq (1-\theta)^2 \left(\sum_{t=1}^k \theta^{k-t} \right) \left(\sum_{t=1}^k \theta^{k-t} \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right) \leq (1-\theta) \sum_{t=1}^k \theta^{k-t} \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t}.$$

For the second part, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}^k)\|^2 &\leq 2\|\nabla f(\mathbf{z}^k)\|^2 + 2\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{z}^k)\|^2 \\ &\leq 4L(f(\mathbf{z}^k) - f^*) + \frac{2L^2 \theta^2 \eta^2}{1-\theta} \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t}, \end{aligned}$$

where we use

$$\begin{aligned} f^* &\leq f\left(\mathbf{z}^k - \frac{1}{L} \nabla f(\mathbf{z}^k)\right) \\ &\leq f(\mathbf{z}^k) - \frac{1}{L} \langle \nabla f(\mathbf{z}^k), \nabla f(\mathbf{z}^k) \rangle + \frac{L}{2} \left\| \frac{1}{L} \nabla f(\mathbf{z}^k) \right\|^2 \\ &= f(\mathbf{z}^k) - \frac{1}{2L} \|\nabla f(\mathbf{z}^k)\|^2. \end{aligned}$$

For the third part, we have

$$\sum_{k=1}^K \|\nabla f(\mathbf{x}^k)\|^2 \leq \sum_{k=1}^K \left(4L(f(\mathbf{z}^k) - f^*) + \frac{2L^2 \theta^2 \eta^2}{1-\theta} \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right).$$

Using

$$\sum_{k=1}^K \sum_{t=1}^{k-1} \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} = \sum_{t=1}^{K-1} \sum_{k=t+1}^K \theta^{k-1-t} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \leq \frac{1}{1-\theta} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t},$$

we have the conclusion. Specially, when $K = 1$, we have $\sum_{k=1}^K \|\nabla f(\mathbf{x}^k)\|^2 = \|\nabla f(\mathbf{x}^1)\|^2 = \|\nabla f(\mathbf{z}^1)\|^2 \leq 2L(f(\mathbf{z}^1) - f^*)$. So we can denote $\sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} = 0$ when $K = 1$ such that the third part holds for all $K \geq 1$. \blacksquare

The next two lemmas are used to bound the second order terms in (12).

Lemma 6 Suppose that Assumptions 1-3 hold. Let $\mathbf{v}_i^0 = \lambda \max \{\sigma_i^2, \frac{1}{dT}\}$ and $\beta = 1 - \frac{1}{T}$. Then for all $K \leq T$, we have

$$\begin{aligned} & \frac{1-\beta}{\sqrt{d}} \sum_{i=1}^d \sigma_i \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_K} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right] \\ & \leq \sigma_s \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max \{d \min_i \sigma_i^2, \frac{1}{T}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) + \frac{L\theta^2\eta^2}{2(1-\theta)^2} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + e + 1 \right). \end{aligned}$$

Proof From Lemma 3, the concavity of $\ln x$, Holder's inequality, and the definition of $\sigma_s = \sqrt{\sum_i \sigma_i^2}$, we have

$$\begin{aligned} & \frac{1-\beta}{\sqrt{d}} \sum_{i=1}^d \sigma_i \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_K} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right] \leq \frac{1}{\sqrt{d}} \sum_{i=1}^d \sigma_i \mathbb{E}_{\mathcal{F}_K} \left[\ln \frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \leq \frac{1}{\sqrt{d}} \sum_{i=1}^d \sigma_i \ln \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \\ & \leq \sqrt{\frac{1}{d} \sum_{i=1}^d \sigma_i^2 \sum_{i=1}^d \left(\ln \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \right)^2} = \sigma_s \sqrt{\frac{1}{d} \sum_{i=1}^d \left(\ln \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \right)^2}. \end{aligned} \quad (21)$$

From the recursion of \mathbf{v}^k , we have $\mathbf{v}_i^K \geq \beta^K \mathbf{v}_i^0$, which leads to $\ln \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \geq 0$. From the concavity of $(\ln x)^2$ for $x \geq e$, we have

$$\begin{aligned} & \frac{1}{d} \sum_{i=1}^d \left(\ln \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \right)^2 \leq \frac{1}{d} \sum_{i=1}^d \left(\ln \left(\mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] + e \right) \right)^2 \\ & \leq \left(\ln \left(\frac{1}{d} \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] + e \right) \right)^2. \end{aligned} \quad (22)$$

Using the recursive update of \mathbf{v}^k , Assumptions 2 and 3, (19), $\mathbf{v}_i^0 = \lambda \max \{\sigma_i^2, \frac{1}{dT}\}$, and Lemma 5, we have

$$\begin{aligned} & \frac{1}{d} \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_K} \left[\frac{\mathbf{v}_i^K}{\beta^K \mathbf{v}_i^0} \right] \\ & = \frac{1}{d} \sum_{i=1}^d \frac{(1-\beta) \sum_{k=1}^K \beta^{K-k} \mathbb{E}_{\mathcal{F}_K} [|\mathbf{g}_i^k|^2] + \beta^K \mathbf{v}_i^0}{\beta^K \mathbf{v}_i^0} \\ & \leq \frac{1}{d} \sum_{i=1}^d \frac{(1-\beta) \sum_{k=1}^K \beta^{K-k} \mathbb{E}_{\mathcal{F}_K} [|\nabla_i f(\mathbf{x}^k)|^2 + \sigma_i^2]}{\beta^K \mathbf{v}_i^0} + 1 \\ & \leq \frac{e^2}{d} \sum_{i=1}^d \frac{(1-\beta) \sum_{k=1}^K \beta^{K-k} \mathbb{E}_{\mathcal{F}_K} [|\nabla_i f(\mathbf{x}^k)|^2] + \sigma_i^2}{\lambda \max \{\sigma_i^2, \frac{1}{dT}\}} + 1 \\ & \leq \frac{e^2(1-\beta)}{\lambda \max \{d \min_i \sigma_i^2, \frac{1}{T}\}} \sum_{k=1}^K \beta^{K-k} \mathbb{E}_{\mathcal{F}_K} [\|\nabla f(\mathbf{x}^k)\|^2] + \frac{e^2}{\lambda} + 1 \\ & \leq \frac{4Le^2(1-\beta)}{\lambda \max \{d \min_i \sigma_i^2, \frac{1}{T}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) + \frac{L\theta^2\eta^2}{2(1-\theta)^2} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + 1. \end{aligned}$$

Plugging into (22) and (21), we have the conclusion. \blacksquare

Specially, replacing each σ_i by 1 and σ_s by \sqrt{d} in the derivation of (21), we have the following lemma.

Lemma 7 *Suppose that Assumptions 1-3 hold. Let $\mathbf{v}_i^0 = \lambda \max \{\sigma_i^2, \frac{1}{dT}\}$ and $\beta = 1 - \frac{1}{T}$. Then for all $K \leq T$, we have*

$$\begin{aligned} & \frac{1-\beta}{d} \sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_K} \left[\frac{|\mathbf{g}_i^k|^2}{\mathbf{v}_i^k} \right] \\ & \leq \ln \left(\frac{4Le^2(1-\beta)}{\lambda \max \{d \min_i \sigma_i^2, \frac{1}{dT}\}} \mathbb{E}_{\mathcal{F}_K} \left[\sum_{k=1}^K (f(\mathbf{z}^k) - f^*) + \frac{L\theta^2\eta^2}{2(1-\theta)^2} \sum_{t=1}^{K-1} \sum_{i=1}^d \frac{|\mathbf{g}_i^t|^2}{\mathbf{v}_i^t} \right] + \frac{e^2}{\lambda} + e + 1 \right). \end{aligned}$$

In the next lemma, the key point is that the dominant part on the right hand side of (23) only depends on $\|\boldsymbol{\sigma}\|_1$ instead of $L(f(\mathbf{x}^1) - f^*)$, which is crucial to give tight dependence on $L(f(\mathbf{x}^1) - f^*)$ in our theoretical analysis. From (16), the third part on the right hand side of (23) is in fact of order $\mathcal{O}(\sqrt{dT})$, confirming that the first term indeed dominates the bound.

Lemma 8 *Suppose that Assumptions 1-3 hold. Let $0 \leq \beta \leq 1$ and $\mathbf{v}_i^0 = \lambda \max \{\sigma_i^2, \frac{1}{dT}\}$ with $\lambda \leq 1$. Then for all $K \leq T$, we have*

$$\sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right] \leq K\|\boldsymbol{\sigma}\|_1 + \sqrt{dT} + 2 \sum_{t=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right]. \quad (23)$$

Proof From the definition of $\tilde{\mathbf{v}}_i^k$, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right] \\ & = \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\beta \mathbf{v}_i^{k-1} + (1-\beta) (|\nabla_i f(\mathbf{x}^k)|^2 + \sigma_i^2)} \right] \\ & = \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{\beta \mathbf{v}_i^{k-1} + (1-\beta) \sigma_i^2}{\sqrt{\beta \mathbf{v}_i^{k-1} + (1-\beta) (|\nabla_i f(\mathbf{x}^k)|^2 + \sigma_i^2)}} + \frac{(1-\beta) |\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\beta \mathbf{v}_i^{k-1} + (1-\beta) (|\nabla_i f(\mathbf{x}^k)|^2 + \sigma_i^2)}} \right] \\ & \leq \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\beta \mathbf{v}_i^{k-1} + (1-\beta) \sigma_i^2} \right] + (1-\beta) \mathbb{E}_{\mathcal{F}_{k-1}} \left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\tilde{\mathbf{v}}_i^k}} \right]. \end{aligned}$$

Consider the first part in the general case. From the recursion of \mathbf{v}_i^k , we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{F}_{k-t}} \left[\sqrt{\beta^t \mathbf{v}_i^{k-t} + (1-\beta^t) \sigma_i^2} \right] \\ & = \mathbb{E}_{\mathcal{F}_{k-t}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\mathbf{g}_i^{k-t}|^2 + (1-\beta^t) \sigma_i^2} \right] \\ & = \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\mathbb{E}_{k-t} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\mathbf{g}_i^{k-t}|^2 + (1-\beta^t) \sigma_i^2} \middle| \mathcal{F}_{k-t-1} \right] \right] \end{aligned}$$

$$\begin{aligned}
 &\stackrel{(1)}{\leq} \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) \mathbb{E}_{k-t} \left[|\mathbf{g}_i^{k-t}|^2 | \mathcal{F}_{k-t-1} \right] + (1-\beta^t) \sigma_i^2} \right] \\
 &\stackrel{(2)}{\leq} \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) (|\nabla_i f(\mathbf{x}^{k-t})|^2 + \sigma_i^2) + (1-\beta^t) \sigma_i^2} \right] \\
 &= \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2 + (1-\beta^{t+1}) \sigma_i^2} \right] \\
 &= \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\frac{\beta^{t+1} \mathbf{v}_i^{k-t-1} + (1-\beta^{t+1}) \sigma_i^2}{\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2 + (1-\beta^{t+1}) \sigma_i^2}} \right] \\
 &\quad + \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\frac{\beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2}{\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2 + (1-\beta^{t+1}) \sigma_i^2}} \right] \\
 &\leq \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + (1-\beta^{t+1}) \sigma_i^2} \right] \\
 &\quad + \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\frac{\beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2}{\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + \beta^t (1-\beta) |\nabla_i f(\mathbf{x}^{k-t})|^2 + (1-\beta^{t+1}) \sigma_i^2}} \right] \\
 &= \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\sqrt{\beta^{t+1} \mathbf{v}_i^{k-t-1} + (1-\beta^{t+1}) \sigma_i^2} \right] + \sqrt{\beta^t} (1-\beta) \mathbb{E}_{\mathcal{F}_{k-t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^{k-t})|^2}{\sqrt{\tilde{\mathbf{v}}_i^{k-t}}} \right],
 \end{aligned}$$

where we use the concavity of \sqrt{x} in $\stackrel{(1)}{\leq}$ and Assumptions 2 and 3 in $\stackrel{(2)}{\leq}$. Applying the above inequality recursively for $t = 1, 2, \dots, k-1$, we have

$$\begin{aligned}
 &\mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\beta \mathbf{v}_i^{k-1} + (1-\beta) \sigma_i^2} \right] \\
 &\leq \sqrt{\beta^k \mathbf{v}_i^0 + (1-\beta^k) \sigma_i^2} + \sum_{t=1}^{k-1} \sqrt{\beta^{k-t}} (1-\beta) \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right]
 \end{aligned}$$

and

$$\begin{aligned}
 \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right] &\leq \sqrt{\beta^k \mathbf{v}_i^0 + (1-\beta^k) \sigma_i^2} + \sum_{t=1}^k \sqrt{\beta^{k-t}} (1-\beta) \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right] \\
 &\leq \sqrt{\sigma_i^2 + \frac{1}{dT}} + \sum_{t=1}^k \sqrt{\beta^{k-t}} (1-\beta) \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right],
 \end{aligned}$$

where we use $\mathbf{v}_i^0 = \lambda \max \{ \sigma_i^2, \frac{1}{dT} \} \leq \sigma_i^2 + \frac{1}{dT}$. Summing over $i = 1, 2, \dots, d$ and $k = 1, 2, \dots, K$, we have

$$\sum_{i=1}^d \sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}} \left[\sqrt{\tilde{\mathbf{v}}_i^k} \right] \leq K \sum_{i=1}^d \left(\sigma_i + \frac{1}{\sqrt{dT}} \right) + \sum_{k=1}^K \sum_{t=1}^k \sqrt{\beta^{k-t}} (1-\beta) \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right]$$

$$\begin{aligned}
&= K\|\boldsymbol{\sigma}\|_1 + \sqrt{dT} + \sum_{t=1}^K \sum_{k=t}^K \sqrt{\beta^{k-t}}(1-\beta) \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right] \\
&\leq K\|\boldsymbol{\sigma}\|_1 + \sqrt{dT} + \frac{1-\beta}{1-\sqrt{\beta}} \sum_{t=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right] \\
&= K\|\boldsymbol{\sigma}\|_1 + \sqrt{dT} + (1+\sqrt{\beta}) \sum_{t=1}^K \sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{t-1}} \left[\frac{|\nabla_i f(\mathbf{x}^t)|^2}{\sqrt{\tilde{\mathbf{v}}_i^t}} \right].
\end{aligned}$$

■

5 Experimental Details

We conduct experiments on both computer vision and natural language processing tasks to verify the relationship $\|\nabla f(\mathbf{x}^k)\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x}^k)\|_2$ during the training of RMSProp and its momentum variant. We call the `torch.optim.RMSprop` API in PyTorch for both optimizers. Code is released at <https://github.com/adonis-dym/Convergence-Rate-RMSProp>.

For computer vision experiments, we train ResNet-50 on both CIFAR-100 and ImageNet datasets. In the CIFAR-100 training task, we set the initial learning rate to 10^{-5} and employ a cosine decay scheduler over all 100 training epochs. We set the batch size to 64 and the weight decay to 0.1. For the ImageNet training task, we utilize the timm library protocol (Wightman, 2019). The training process spans 200 epochs with a 20-epoch linear warm-up period to increase the learning rate to 10^{-4} , a 170-epoch cosine decay period to decrease the learning rate to 10^{-5} , and 10 final epochs with a constant learning rate 10^{-5} . We set the batchsize to 2048. We maintain identical settings for both optimizers and configure the other parameters using the default settings in the PyTorch API, including assigning the momentum parameter to 0.9 for RMSProp with momentum. At the end of each epoch, we compute the full training loss and gradient by traversing the entire training dataset to accurately measure the gradient norm ratio $\frac{\|\nabla f(\mathbf{x}^k)\|_1}{\|\nabla f(\mathbf{x}^k)\|_2}$.

For natural language processing tasks, we train the classic GPT-2 model from scratch on the OpenWebText dataset using the Megatron-LM framework. Setting the batchsize to 640, we train the model for 50000 steps, equivalent to approximately 3.5 epochs. The training schedule includes a 2000-step linear warm-up period increasing the learning rate to 10^{-5} and a cosine decay period for the remaining steps. In our training setting, we employ a decoupled weight decay of 0.05 in the AdamW style, rather than the vanilla implementation of ℓ_2 regularization in the PyTorch API. Given the computational constraints inherent in large-scale language model training, we approximate full gradients by aggregating over a subset of 100 batches, providing an efficient yet representative estimate of the full gradients.

Our experimental results, as compiled in Figure 1 in Section 1, demonstrate that the gradient norm ratio $\frac{\|\nabla f(\mathbf{x}^k)\|_1}{\|\nabla f(\mathbf{x}^k)\|_2}$ consistently scales as $\Theta(\sqrt{d})$. This empirical observation confirms that the convergence rate derived in this study is in accordance with that of SGD with respect to the problem dimension d .

Conclusion and Future Work

This paper studies the classical RMSProp and its momentum extension. We establish the convergence rate of $\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_1] \leq \tilde{\mathcal{O}} \left(\frac{\sqrt{d}}{T^{1/4}} \left(\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} \right) \right)$ measured by ℓ_1 norm without the bounded gradient condition. Our convergence rate can be considered to be analogous to the $\frac{1}{T} \sum_{k=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}^k)\|_2] \leq \mathcal{O} \left(\frac{1}{T^{1/4}} \sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} \right)$ one of SGD in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$ for high-dimensional problems. One interesting future work is to establish the lower bound of adaptive gradient methods measured by ℓ_1 norm. We conjecture that the lower bound is $\mathcal{O} \left(\frac{\sqrt{d}}{T^{1/4}} \left(\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)} \right) \right)$.

Appendix A. Proof of Lemma 3

Proof From $\ln(1-x) \leq -x$ for any $x < 1$, we have

$$(1-\beta)\frac{g_t^2}{v_t} \leq -\ln \left(1 - (1-\beta)\frac{g_t^2}{v_t} \right) = -\ln \frac{v_t - (1-\beta)g_t^2}{v_t} = -\ln \frac{\beta v_{t-1}}{v_t} = \ln \frac{v_t}{\beta v_{t-1}}$$

and

$$(1-\beta) \sum_{t=1}^k \frac{g_t^2}{v_t} \leq \ln \frac{v_k}{\beta^k v_0}.$$

■

Appendix B. Proof of $c \ln x \leq x$ for all $x \geq 3c \ln c$ and $c \geq 3$

Proof Denote $f(x) = \frac{\ln x}{x}$. Since $f'(x) = \frac{1}{x^2} - \frac{\ln x}{x^2} \leq 0$ when $x \geq e$, $f(x)$ is decreasing when $x \geq e$ and $\frac{\ln x}{x} \leq \frac{\ln(3c \ln c)}{3c \ln c} = \frac{1}{c} \frac{\ln c + \ln(3 \ln c)}{3 \ln c} \leq \frac{1}{c} \frac{\ln c + \ln c^2}{3 \ln c} = \frac{1}{c}$, where we use $\ln c \leq c$ and $3 \ln c \leq c^2$ for $c \geq 3$. ■

Acknowledgments and Disclosure of Funding

H. Li was supported by NSF China (No. 62476142) and Z. Lin was supported by the Beijing Natural Science Foundation (No. L257007) and the NSF China (No. 62276004). Li and Lin are the corresponding authors.

References

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023.

- Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning (ICML)*, 2023.
- Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning (ICML)*, 2018.
- Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. SignSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, 2018.
- Leon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Congliang Chen, Li Shen, Fangyu Zou, and Wei Liu. Towards practical Adam: Non-convexity, convergence theory, and mini-batch acceleration. *Journal of Machine Learning Research*, 23(229):1–47, 2022.
- Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyang Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021.
- Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations (ICLR)*, 2019.
- Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research*, 2022.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.
- Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory (COLT)*, 2022.
- Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. *Arxiv: 2112.03459*, 2021.
- Yusu Hong and Junhong Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *Arxiv: 2311.02000*, 2023.
- Yusu Hong and Junhong Lin. Revisiting convergence of Adagrad with relaxed assumptions. In *Uncertainty in Artificial Intelligence (UAI)*, 2024a.

- Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.
- Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations (ICLR)*, 2022.
- Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning (ICML)*, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018.
- Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations (ICLR)*, 2019.
- H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, 2010.
- Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.
- Pedro Savarese, David McAllester, Sudarshan Babu, and Michael Maire. Domain-independent dominance of adaptive methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyper-parameter. In *International Conference on Learning Representations (ICLR)*, 2020.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural Networks for Machine Learning*, 2012.
- Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and the lower bound of Adam’s iteration complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

- Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: simple proofs and relaxed assumptions. In *Conference on Learning Theory (COLT)*, 2023b.
- Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):9047–9076, 2020.
- Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.
- Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.
- Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations (ICLR)*, 2019.
- Manzil Zaheer, Sashank J.Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.
- Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance. *Transactions on Machine Learning Research*, 2025.
- Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.