On the $\mathcal{O}(\frac{\sqrt{d}}{K^{1/4}})$ Convergence Rate of AdamW Measured by ℓ_1 Norm

Huan Li [1] ✉ lihuanss@nankai.edu.cn
Yiming Dong [2] ymdong@stu.pku.edu.cn
Zhouchen Lin [2,3,4] ✉ zlin@pku.edu.cn

1. College of Artificial Intelligence, Nankai University, Tianjin, China
2. State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
3. Institute for Artificial Intelligence, Peking University, Beijing, China
4. Pazhou Laboratory (Huangpu), Guangzhou, China

Abstract

As the default optimizer for training large language models, AdamW has achieved remarkable success in deep learning. However, its convergence behavior is not theoretically well-understood. This paper establishes the convergence rate $\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq \mathcal{O}(\frac{\sqrt{d}C}{K^{1/4}})$ for AdamW measured by ℓ_1 norm, where K represents the iteration number, d denotes the model dimension, and C matches the constant in the optimal convergence rate of SGD. Theoretically, we have $\|\nabla f(\mathbf{x})\|_2 \ll \|\nabla f(\mathbf{x})\|_1 \leq \sqrt{d}\|\nabla f(\mathbf{x})\|_2$ for any high-dimensional vector $\mathbf{x}$ and $\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_1\right] \geq \sqrt{\frac{2d}{\pi}}\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_2\right]$ when each element of $\nabla f(\mathbf{x})$ is generated from Gaussian distribution $\mathcal{N}(0,1)$. Empirically, our experimental results on real-world deep learning tasks reveal $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$. Both support that our convergence rate can be considered to be analogous to the optimal $\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_2\right] \leq \mathcal{O}(\frac{C}{K^{1/4}})$ convergence rate of SGD.

1 Introduction

AdamW, which modifies Adam by decoupling weight decay from gradient-based updates, has emerged as the dominant optimizer for training deep neural networks, particularly for large language models. AdamW represents the pinnacle of adaptive gradient algorithms, having developed through the progression of AdaGrad [1, 2], RMSProp [3], Adam [4], and finally AdamW [5] itself. Although the literature on the convergence analysis of adaptive gradient algorithms is quite extensive, there has been little research on the convergence properties of AdamW.

Recently, Xie and Li [6] proved that if the iterates of AdamW converge to some $\mathbf{x}_\infty$, then $\mathbf{x}_\infty$ is a KKT point of the constrained problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \quad s.t. \quad \|\mathbf{x}\|_\infty \leq \frac{1}{\lambda}, \tag{1}$$

where $f(\mathbf{x})$ is the nonconvex objective function, $\|\cdot\|_\infty$ is the infinity norm, and λ is the weight decay parameter. Moreover, $\mathbf{x}$ is a KKT point of problem (1) iff [6]

$$\|\mathbf{x}\|_\infty \leq \frac{1}{\lambda} \quad \text{and} \quad \langle \lambda\mathbf{x}, \nabla f(\mathbf{x})\rangle + \|\nabla f(\mathbf{x})\|_1 = 0. \tag{2}$$

Xie and Li characterized which solution does AdamW converge to, if it indeed converges. The next fundamental question to address is whether and how fast AdamW converges. Zhou et al. [7]

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

conducted preliminary exploration on this problem. However, their analysis requires the weight decay parameter to decrease exponentially, making AdamW reduce to Adam finally. To the best of our knowledge, aside from [7], we have not found any other literature addressing the convergence issue of AdamW.

In practical deep learning training, we often initialize the network weights small and employ modest weight decay, for example, $\lambda = 0.01$, which empirically confines the optimization trajectory within the ℓ_∞ norm constraint, as empirically demonstrated in Figure 3. That is, $\|\mathbf{x}\|_\infty \leq \frac{c}{\lambda}$ for some $c < 1$, making $\langle \lambda \mathbf{x}, \nabla f(\mathbf{x}) \rangle + \|\nabla f(\mathbf{x})\|_1$ lower bounded by $(1-c)\|\nabla f(\mathbf{x})\|_1$. This key property enables the use of $\|\nabla f(\mathbf{x})\|_1$ as an effective yet significantly simpler convergence metric for AdamW in practical settings.

Building on the above observation, this paper focuses on the convergence rate of AdamW within the constraint in problem (1). Specifically, we prove the following convergence rate for AdamW

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq \mathcal{O}\left(\frac{\sqrt{d}}{K^{1/4}}\sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1)-f^*)} + \sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}\right) \tag{3}$$

by proper parameter settings such that $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$ for all iterates, where K is the total iteration number, d is the model dimension, σ_s is the gradient noise variance, L is the Lipschtiz smooth constant, and f^* is a lower bound of $f(\mathbf{x})$. Recall the classical convergence rate of SGD [8]

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

which matches the lower bound of nonconvex stochastic optimization [9]. Comparing (3) with (4), we see that our convergence rate (3) also achieves the same lower bound with respect to K, σ_s, L, and $f(\mathbf{x}^1) - f^*$). The only coefficient left unclear whether it is tight is the dimension d. Theoretically, we have $\|\nabla f(\mathbf{x})\|_2 \ll \|\nabla f(\mathbf{x})\|_1 \leq \sqrt{d}\|\nabla f(\mathbf{x})\|_2$ for any high-dimensional vector $\mathbf{x}$ and $\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_1\right] \geq \sqrt{\frac{2d}{\pi}}\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_2\right]$ when each element of $\nabla f(\mathbf{x})$ is generated from Gaussian distribution $\mathcal{N}(0,1)$. Empirically, we have observed $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$ on real-world deep learning tasks, as shown in Figure 2. Thus, we could say that our convergence rate (3) can be considered to be analogous to (4) of SGD in the ideal case.

As a special case, we also establish the same convergence rate (3) for Adam under slightly relaxed parameter settings than AdamW. To the best of our knowledge, this convergence rate only appears for RMSProp firstly proved in [10], and similar results for AdaGrad subsequently appeared in [11, 12] and RMSProp in [13] under different assumptions. Notably, comparable convergence guarantees remain unproven for AdamW and Adam.

2 Convergence Rate of AdamW

This section presents our convergence rate analysis for AdamW. We first describe the assumptions used throughout this paper as follows, where we denote $\mathcal{F}_k = \sigma(\mathbf{g}^1, \mathbf{g}^2, \cdots, \mathbf{g}^k)$ to be the sigma field of the stochastic gradients up to k, denote $\mathbb{E}_{\mathcal{F}_k}[\cdot]$ as 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 $\mathcal{F}_{k-1}$.

Assumptions:

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

Algorithm 1 AdamW

Hyper parameters: $\eta, \theta, \beta, \lambda, \varepsilon$
Initialize $\mathbf{x}^1$, $\mathbf{m}^0 = 0$, $\mathbf{v}^0 = 0$
for $k = 1, 2, \cdots, K$ **do**
  $\mathbf{g}^k = \text{GradOracle}(\mathbf{x}^k)$
  $\mathbf{m}^k = \theta\mathbf{m}^{k-1} + (1-\theta)\mathbf{g}^k$
  $\mathbf{v}^k = \beta\mathbf{v}^{k-1} + (1-\beta)(\mathbf{g}^k)^{\odot 2}$
  $\mathbf{x}^{k+1} = (1-\lambda\eta)\mathbf{x}^k - \frac{\eta}{\sqrt{\mathbf{v}^k+\varepsilon}} \odot \mathbf{m}^k$
end for

Denoting $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_d]$ as the noise variance vector and $\sigma_s = \|\boldsymbol{\sigma}\|_2 = \sqrt{\sum_{i=1}^d \sigma_i^2}$, we have the

following standard bounded noise variance assumption

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

Algorithm 1 provides the complete AdamW implementation, where we denote $\odot$ for the Hadamard product. Setting the weight decay parameter $\lambda = 0$ recovers the standard Adam. For analytical simplicity, we omit the bias correction term in our analysis.

Based on Assumptions 1-3, we provide the convergence rate of AdamW in the following theorem. Note that we do not assume the boundedness of the gradient $\nabla f(\mathbf{x}^k)$ or stochastic gradient $\mathbf{g}^k$.

Theorem 1 *Suppose that Assumptions 1-3 hold. Define* $\hat{\sigma}_s^2 = \max\left\{\sigma_s^2, \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}\right\}$ *with any constant* $\gamma \in (0, 1]$. *Let* $1-\theta = \sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{K\hat{\sigma}_s^2}}$, $\theta \leq \beta \leq \sqrt{\theta}$[1], $\eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}}$, $\varepsilon = \frac{\hat{\sigma}_s^2}{d}$, $\lambda \leq \frac{\sqrt{d}}{\sqrt{72}K^{3/4}}\sqrt[4]{\frac{L^3}{\hat{\sigma}_s^2(f(\mathbf{x}^1)-f^*)}}$, *and* $\|\mathbf{x}^1\|_\infty \leq \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}$. *Then for AdamW, we have* $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$ *for all* $k = 1, 2, \cdots, K$ *and*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq \frac{8\sqrt{d}}{K^{1/4}}\sqrt[4]{\hat{\sigma}_s^2 L(f(\mathbf{x}^1)-f^*)} + 30\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}.$$

Specially, when $\sigma_s^2 \leq \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}$, *we have* $1-\theta = \gamma$, $\theta \leq \beta \leq \sqrt{\theta}$, $\eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}}$, $\varepsilon = \frac{L(f(\mathbf{x}^1)-f^*)}{dK\gamma^2}$, $\lambda \leq \sqrt{\frac{dL\gamma}{72K(f(\mathbf{x}^1)-f^*)}}$, $\|\mathbf{x}^1\|_\infty \leq \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}$, $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$, *and accordingly*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq 38\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K\gamma}}.$$

Theorem 1 demonstrates that AdamW minimizes the gradient norm directly while restricting $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$. As a comparison, ℓ_2 regularized Adam only minimizes $\|\nabla f(\mathbf{x}) + \lambda\mathbf{x}\|$, rather than $\|\nabla f(\mathbf{x})\|$.

As a special case, we also establish the same convergence rate for Adam in the following corollary under slightly relaxed parameter settings. The complete description of Corollary 1 is given in Appendix B.

Corollary 1 *With the same assumptions and parameter settings of* $1-\theta$, η, *and* ε *as Theorem 1, but only requiring* $0 \leq \beta \leq 1$ *rather than both* $\theta \leq \beta \leq \sqrt{\theta}$ *and* $\|\mathbf{x}^1\|_\infty \leq \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}$, *we have the same convergence rate for Adam as established in Theorem 1.*

2.1 Optimality of Our Convergence Rate

When comparing our convergence rate (3) with the optimal rate (4) of SGD, which aligns with the lower bound in nonconvex stochastic optimization, we observe that our rate is also optimal with respect to K, σ_s, L, and $f(\mathbf{x}^1) - f^*$. The only remaining uncertainty concerns the tightness of the dimension d. Theoretically, $\|\nabla f(\mathbf{x})\|_2 \ll \|\nabla f(\mathbf{x})\|_1 \leq \sqrt{d}\|\nabla f(\mathbf{x})\|_2$ holds for any high-dimensional vector $\mathbf{x}$, and when each element of $\nabla f(\mathbf{x})$ is drawn from Gaussian distribution $\mathcal{N}(0, 1)$, we have $\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_1\right] \geq \sqrt{\frac{2d}{\pi}}\mathbb{E}\left[\|\nabla f(\mathbf{x})\|_2\right]$ from Lemma 1. Empirically, experiments on real deep neural networks training confirm $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$, as demonstrated in Figure 2. Thus, our convergence rate (3) can be regarded to be analogous to SGD's optimal rate (4).

Lemma 1 *When each entry of* $\mathbf{x} \in \mathbb{R}^d$ *is generated from Gaussian distribution with zero mean and unit variance, we have* $\mathbb{E}\left[\|\mathbf{x}\|_1\right] \geq \sqrt{\frac{2d}{\pi}}\mathbb{E}\left[\|\mathbf{x}\|_2\right]$.

[1]We gratefully thank the anonymous NeurIPS reviewer to derive this looser bound. Our original bound is $\theta \leq \beta \leq \frac{(1+\theta)^2}{4}$.

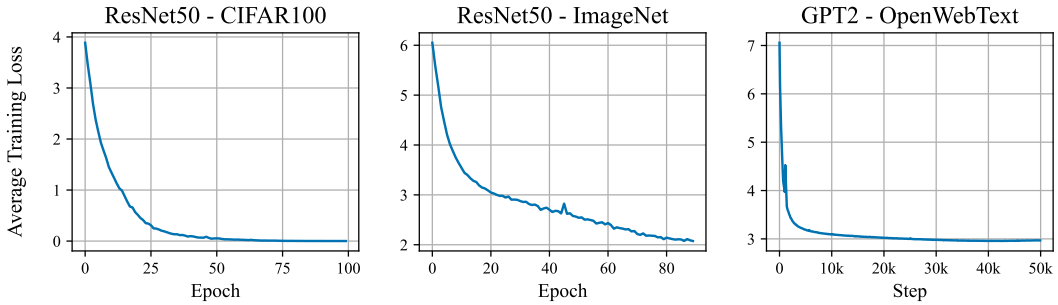

Figure 1: Illustration of average training loss $f(\mathbf{x}^k)$ over epochs/steps, and at the initialization, $f(\mathbf{x}^1) \leq 8$.

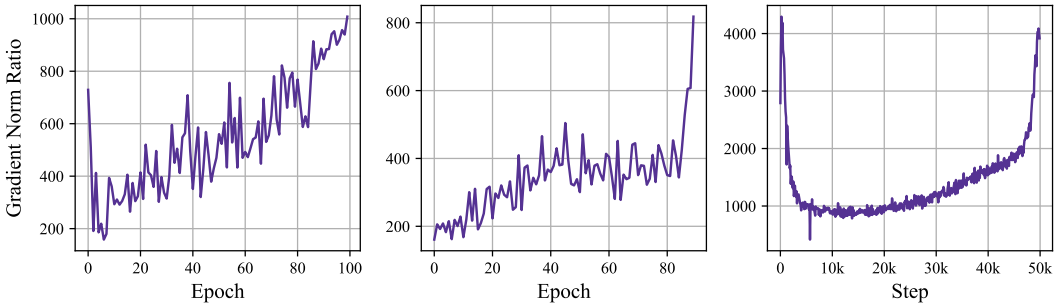

Figure 2: Illustration of $\|\nabla f(\mathbf{x}^k)\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x}^k)\|_2$ over epochs/steps. The gradient norm ratio shows $\frac{\|\nabla f(\mathbf{x}^k)\|_1}{\|\nabla f(\mathbf{x}^k)\|_2}$, and $\sqrt{d} = 4868$, 5060, and 11136, respectively.

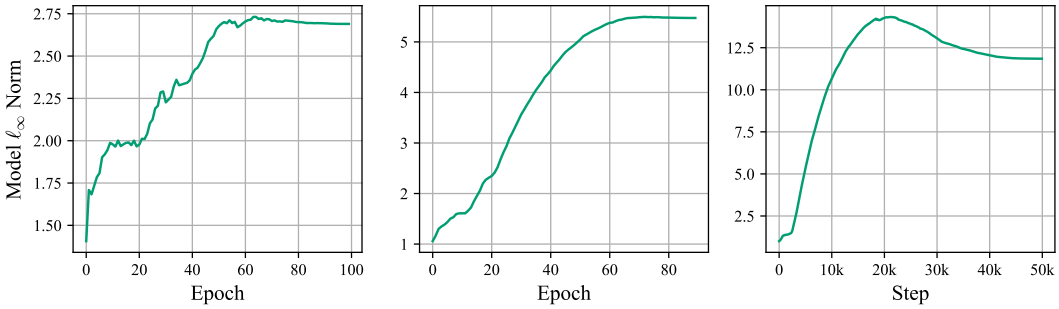

Figure 3: Illustration of $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$ over epochs/steps. The model ℓ_∞ norm shows $\|\mathbf{x}^k\|_\infty$, and $\lambda = 0.01$, 0.1, and 0.05, respectively.

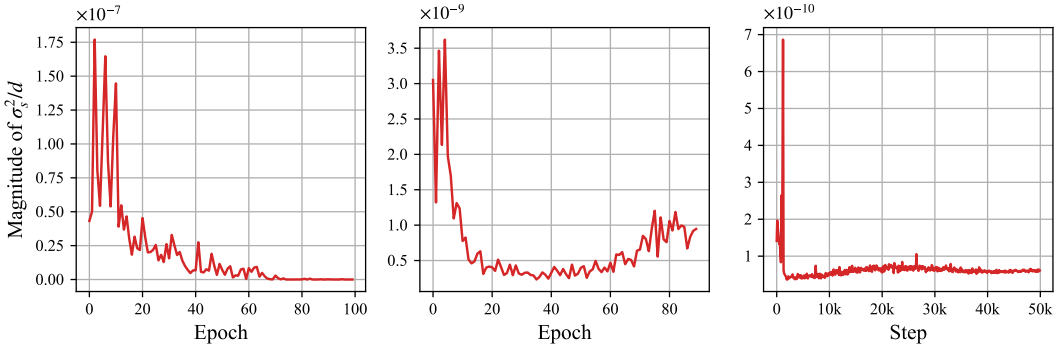

Figure 4: Illustration of small $\frac{\sigma_s^2}{d}$ over epochs/steps. The magnitude σ_s^2 is approximated by $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2$ without taking expectation, and $d = 2.37 \times 10^7$, 2.56×10^7, and 1.24×10^8, respectively.

Recently, Jiang et al. [11] established a fundamental lower bound for SGD when measuring gradients by ℓ_1 norm, which is of order $\Omega\left(\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}+\sqrt[4]{\frac{dL(f(\mathbf{x}^1)-f^*)\|\boldsymbol{\sigma}\|_1^2}{K}}\right)$ under Assumptions 1-3. When $\|\boldsymbol{\sigma}\|_1 \approx \sqrt{d}\|\boldsymbol{\sigma}\|_2 = \sqrt{d}\sigma_s$, this lower bound precisely aligns with our convergence rate in (3). We further conjecture that this lower bound applies more broadly to general first-order stochastic optimization algorithms under ℓ_1 norm gradient measurement. This would imply that our derived convergence rate is nearly tight.

2.2 Separating the Convergence Rate by the Noise Variance

In Theorem 1, we separate the convergence rate by the magnitude of σ_s. When $\sigma_s^2 \geq \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}$, both the convergence rates of AdamW and Adam are $\mathcal{O}(\frac{\sqrt{d}}{K^{1/4}})$. When σ_s^2 becomes smaller than $\frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}$, the convergence rates improve to $\mathcal{O}(\sqrt{\frac{d}{K}})$, matching that of gradient descent measured by ℓ_1 norm.

2.3 Reasonable Weight Decay Parameter and Initialization Interval

In Theorem 1, we set the weight decay parameter λ smaller than $\frac{\sqrt{d}}{\sqrt{72}K^{3/4}}\sqrt[4]{\frac{L^3}{\hat{\sigma}_s^2(f(\mathbf{x}^1)-f^*)}}$. In modern deep neural networks, the dimension d is typically extremely large, for example, $d = 1.75 \times 10^{11}$ in GPT-3, making $\frac{\sqrt{d}}{K^{3/4}}$ almost certainly exceed 0.01, which is the default setting of λ in PyTorch official implementation. For example, in the experiments of our paper, we train ResNet-50 on i) CIFAR-100 and ii) ImageNet dataset, and GPT-2 on iii) OpenWebText, and observe $(K, d) = (39100, 2.37\times10^7)$, $(28080, 2.56 \times 10^7)$, and $(50000, 1.24 \times 10^8)$, resulting in $\frac{\sqrt{d}}{K^{3/4}} \approx 1.75$, 2.33, and 3.33, respectively. We empirically show in Appendix D that large λ may cause AdamW not converge and thus a upper bound is necessary. We also initialize $\|\mathbf{x}^1\|_\infty \leq \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}}$. Although $\sqrt{\frac{K}{d}}$ is typically smaller than 1 in large language models training, it remains not too small. In practical configurations, we often initialize the network weights very small. On the other hand, although we always initialize the scale parameter in BatchNorm/LayerNorm to 1, we do not use weight decay for the scale parameter in practice.

2.4 Small ε Setting

In practice, ε is typically set to a very small value, for example, approximately 10^{-16} in PyTorch implementation[2], to prevent division by zero while maintaining the adaptive properties of AdamW and Adam. Larger ε values would make AdamW and Adam behave similarly to SGD, losing its adaptive learning rate adjustment. In Theorem 1, we set $\varepsilon = \frac{\hat{\sigma}_s^2}{d} = \max\left\{\frac{\sigma_s^2}{d}, \frac{L(f(\mathbf{x}^1)-f^*)}{dK\gamma^2}\right\}$, which remains small due to extremely large d and modest σ_s^2. We have empirically shown in Figure 4 that $\frac{\sigma_s^2}{d} \approx 10^{-7}$, 10^{-9}, and 10^{-10} in our experiments of ResNet-50 on Cifar-100 and ImageNet and GPT-2 on OpenWebText, respectively. Intuitively, ε should be smaller than the square of stochastic gradient at each coordinate, otherwise, ε would dominate the magnitude of $\mathbf{v}_i^k$ in $\frac{1}{\sqrt{\mathbf{v}_i^k+\varepsilon}}$. Our setting of ε, the coordinate-wise average of gradient noise variance, approximately resides at this critical threshold. Notably, our convergence rates for both AdamW and Adam do not depend on ε explicitly. In comparison, existing convergence rates for AdamW and Adam in the literature either explicitly depend on ε or exhibit a higher dependence on the dimension d.

2.5 Unpractical Settings of η, θ, and β

In Theorem 1, we set the learning rate η very small and the parameters θ and β nearly equal to 1 to satisfy the proof requirements. This differs from standard implementations where $(\theta, \beta) = (0.9, 0.999)$ is typically used. Although investigating AdamW/Adam's property under realistic

[2]In PyTorch official implementation, ε appeared in a different place in $\frac{\eta}{\sqrt{\mathbf{v}^k}+\varepsilon}$ and $\varepsilon = 10^{-8}$, while we use $\frac{\eta}{\sqrt{\mathbf{v}^k+\varepsilon}}$.

configurations represents an important research direction, as it could yield valuable insights for deep learning hyperparameters tuning, the practical configurations may not guarantee the convergence. For instance, prior work [14, 15] demonstrates through constructed examples that Adam with common hyperparameters ($\theta = 0.9$, $\beta = \{0.999, 0.997, 0.995, 0.993\}$, and $\eta_k = \frac{0.1}{\sqrt{k}}$) fail to converge to stationary points (see [15, Figuire 2]). On the other hand, empirical evidence from recent studies [16, Figure 9] [17, Figure 6] demonstrate that during the training of language models with practical parameter configurations, the gradient norm hardly decreases during the training run, although the objective function decreases sufficiently.

2.6 No Conflict with [6]

For sufficiently large weight decay parameter λ where no critical points exist within the constrained domain of problem (1), the KKT conditions (2) serve as a natural convergence metric. As λ diminishes, problem (1) asymptotically approaches an unconstrained optimization problem, and AdamW reduces to Adam in the limit. There exists a critical threshold beyond which $\|\nabla f(\mathbf{x})\|_1$ also becomes a viable metric for convergence. Consequently, our results do not conflict with [6].

3 Proof Sketch

In this section, we outline the proof sketch of Theorem 1. The detailed proofs are provided in Appendix A. From the Lipschitz smoothness of $f(\mathbf{x})$ and the update of $\mathbf{x}^{k+1}$ in Algorithm 1, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{x}^{k+1})|\mathcal{F}_{k-1}\right] - f(\mathbf{x}^k)\\
\leq&\mathbb{E}_k\left[\left\langle\nabla f(\mathbf{x}^k), \mathbf{x}^{k+1}-\mathbf{x}^k\right\rangle + \frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\middle|\mathcal{F}_{k-1}\right]\\
=&\mathbb{E}_k\left[\underbrace{-\eta\sum_{i=1}^{d}\left\langle\nabla_i f(\mathbf{x}^k), \frac{\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\right\rangle}_{\text{term (a)}} + \underbrace{\frac{L\eta^2}{2}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\mathbf{v}_i^k+\varepsilon}}_{\text{term (b)}}\middle|\mathcal{F}_{k-1}\right].
\end{aligned}\tag{5}$$

Decompose term (a) into

$$-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\underbrace{-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}}_{\text{term (c)}} + \underbrace{\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)-\mathbf{m}_i^k-\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}}_{\text{term (d)}}\tag{6}$$

and relax term (b) as follows to absorb it within term (c)

$$\text{term (b)} \leq \frac{L\eta^2}{2\sqrt{\varepsilon}}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}} \overset{\eta\leq\frac{\sqrt{\varepsilon}}{2L}}{\leq} \frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}.\tag{7}$$

Next, we consider term (d) and relax it as follows

$$\text{term (d)} \leq \frac{\eta}{\sqrt{\varepsilon}}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2 + \eta\sum_{i=1}^{d}|\lambda\mathbf{x}_i^k|^2\sqrt{\mathbf{v}_i^k+\varepsilon}.\tag{8}$$

We see that the parameter ε plays a pivotal rule in steps (7) and (8). Combing (5)-(8), we have

$$\begin{aligned}
\mathbb{E}_k\left[f(\mathbf{x}^{k+1})|\mathcal{F}_{k-1}\right] - f(\mathbf{x}^k) \leq&\mathbb{E}_k\Bigg[-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}} - \frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\\
&+\underbrace{\frac{\eta}{\sqrt{\varepsilon}}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2}_{\text{term (e)}} + \eta\sum_{i=1}^{d}|\lambda\mathbf{x}_i^k|^2\sqrt{\mathbf{v}_i^k+\varepsilon}\Bigg|\mathcal{F}_{k-1}\Bigg].
\end{aligned}\tag{9}$$

Considering term (e), we can use standard techniques in the analysis of momentum SGD to build a recursion (Lemma 4) as follows

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2|\mathcal{F}_{k-1}\right]\\
\leq&\theta\left\|\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right\|^2+\frac{L^2\eta^2}{\sqrt{\varepsilon}(1-\theta)}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}}+(1-\theta)^2\sigma_s^2.
\end{aligned}\tag{10}$$

Multiplying both sides of (10) by $\frac{\eta}{\sqrt{\varepsilon}(1-\theta)}$, adding it to (9), and letting $\eta^2\leq\frac{\varepsilon(1-\theta)^2}{4L^2}$, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{x}^{k+1})-f^*+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2+\frac{\eta}{4}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\Big|\mathcal{F}_{k-1}\right]\\
\leq&f(\mathbf{x}^k)-f^*+\sum_{i=1}^d\mathbb{E}_k\left[\underbrace{-\frac{\eta}{2}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}}_{\text{term (f)}}+\underbrace{\eta|\lambda\mathbf{x}_i^k|^2\sqrt{\mathbf{v}_i^k+\varepsilon}}_{\text{term (g)}}\Big|\mathcal{F}_{k-1}\right]\\
&+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^{k-1})-\mathbf{m}^{k-1}\right\|^2+\frac{\eta}{4}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}}+\frac{\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}.
\end{aligned}\tag{11}$$

The above analysis comes from the standard framework and contains nothing new. We can recursively eliminate certain terms in (11) after telescoping, except for the troublesome term (g). The following outlines the key technical components of our proof to address term (g) and achieve the tight convergence rate.

3.1 Bounding $|\lambda\mathbf{x}_i^k|$ in Term (g) by $O(\frac{1}{K^{1/4}})$

The analysis for AdamW proves more challenging than for SignSGD-type methods with weight decay [18], because SignSGD maintains a fixed update size 1, whereas AdamW's updates can be arbitrarily large. Specifically, for AdamW and Adam, we have $\frac{|\mathbf{m}_i^k|^2}{\mathbf{v}_i^k}\leq\frac{(1-\theta)^2\beta}{(1-\beta)(\beta-\theta^2)}$ (Lemma 2), where the latter is minimized to be 1 by setting $\theta=\beta$. However, when setting $\theta=O(1)$ (for example, $\theta=0.9$) and $\beta=1-\frac{1}{K}$, we have $\frac{(1-\theta)^2\beta}{(1-\beta)(\beta-\theta^2)}=O(K)$, leading to unbounded updates in AdamW. This fundamental difficulty prevents direct extension of the proof framework in [10] to AdamW. We set $\theta\leq\beta\leq\sqrt{\theta}$ in Theorem 1 such that $\frac{(1-\theta)^2\beta}{(1-\beta)(\beta-\theta^2)}\leq 4$ (Lemma 2). Then for the update of $\mathbf{x}^{k+1}$ in AdamW, we have

$$\|\mathbf{x}^{k+1}\|_\infty-\frac{2}{\lambda}\leq(1-\eta\lambda)^k\left(\|\mathbf{x}^1\|_\infty-\frac{2}{\lambda}\right).$$

When $(1-\eta\lambda)^k$ decreases fast, we have $(1-\eta\lambda)^k\left(\|\mathbf{x}^1\|_\infty-\frac{2}{\lambda}\right)\to 0$ and $\|\mathbf{x}^{k+1}\|_\infty$ is loosely bounded by $\frac{2}{\lambda}$, which is far from our target $\lambda\|\mathbf{x}^{k+1}\|_\infty\leq O(\frac{1}{K^{1/4}})$. To address this issue, we control the decrease of $(1-\eta\lambda)^k$ by setting parameter λ properly such that $\eta\lambda\leq\frac{\sqrt{\nu}}{2K^{5/4}}$ and $(1-\eta\lambda)^k\geq e^{-\frac{\sqrt{\nu}}{K^{1/4}}}\geq 1-\frac{\sqrt{\nu}}{K^{1/4}}$ for some ν and any $k\leq K$. Equipped with proper initialization of $\|\mathbf{x}^1\|_\infty\leq\frac{\sqrt{\nu}}{K^{1/4}\lambda}$, we finally have (Lemma 3)

$$\|\mathbf{x}^{k+1}\|_\infty\leq\frac{2}{\lambda}-\left(1-\frac{\sqrt{\nu}}{K^{1/4}}\right)\left(\frac{2}{\lambda}-\frac{\sqrt{\nu}}{K^{1/4}\lambda}\right)\leq\frac{3}{\lambda}\frac{\sqrt{\nu}}{K^{1/4}},$$

and

$$\text{term (g)}\leq\frac{9\eta\nu}{K^{1/2}}\sqrt{\mathbf{v}_i^k+\varepsilon}.$$

Intuitively, when the initialization is far from the boundary of problem (1) and $(1-\eta\lambda)^k\approx 1$, the iterates $\mathbf{x}^{k+1}$ are guaranteed to be far from the boundary throughout the optimization process.

3.2 Absorbing Term (g) within Term (f)

To absorb term (g) within term (f), we first relax $\sqrt{\mathbf{v}_i^k+\varepsilon}$ to $\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}$ as follows by the concavity of $\sqrt{x}$ and $-\frac{1}{\sqrt{x}}$

$$\mathbb{E}_k\left[\underbrace{-\frac{\eta}{2}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}}_{\text{term (f)}}+\underbrace{\eta|\lambda\mathbf{x}_i^k|^2\sqrt{\mathbf{v}_i^k+\varepsilon}}_{\text{term (g)}}\Big|\mathcal{F}_{k-1}\right]\leq-\frac{\eta}{2}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}+\frac{9\eta\nu}{K^{1/2}}\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon},$$

where we define $\widetilde{\mathbf{v}}_i^k=\beta\mathbf{v}_i^{k-1}+(1-\beta)\left(\left|\nabla_i f(\mathbf{x}^k)\right|^2+\sigma_i^2\right)$. Then, we can bound $\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}$ as follows (Lemma 5) and absorb term (h) within $-\frac{\eta}{2}\sum_{k=1}^K\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}$ derived from term (f),

$$\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{k-1}}\left[\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\right]\leq K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}+\underbrace{2\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{t-1}}\left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]}_{\text{term (h)}}. \tag{12}$$

Summing (11) over k and combing the above analysis, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{x}^{K+1})-f^*+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^K)-\mathbf{m}^K\right\|^2+\frac{\eta}{4}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^K+\lambda\mathbf{x}_i^K\sqrt{\mathbf{v}_i^K+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^K+\varepsilon}}\right]\\
\leq&-\frac{\eta}{2}\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]+\frac{9\eta\nu}{K^{1/2}}\left(\underbrace{K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}}_{\text{term (i)}}+2\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{|\nabla_i f(\mathbf{x}^k)|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\right)\\
&+\underbrace{f(\mathbf{x}^1)-f^*+\frac{\eta}{\sqrt{\varepsilon}(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1)-\mathbf{m}^1\|^2\right]+\frac{K\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}}_{\text{term (j)}}\\
\leq&-\frac{\eta}{4}\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]+18\eta\nu d\sqrt{K\varepsilon}+\text{term (j)}
\end{aligned}$$

by letting $\frac{9\nu}{K^{1/2}}\leq\frac{1}{8}$ and $\varepsilon=\frac{\sigma_s^2}{d}$ such that $K\|\boldsymbol{\sigma}\|_1\leq Kd\sqrt{\varepsilon}$. Letting $\nu=\frac{1}{72d}\sqrt{\frac{\sigma_s^2L(f(\mathbf{x}^1)-f^*)}{\varepsilon^2}}$, $1-\theta=\sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{K\sigma_s^2}}$ and $\eta=\sqrt{\frac{\varepsilon(f(\mathbf{x}^1)-f^*)}{4K\sigma_s^2L}}$, both term (j) and $\eta\nu d\sqrt{K\varepsilon}$ are of the order $\eta\sqrt{\frac{K\sigma_s^2L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}$. This accounts for why bounding $|\lambda\mathbf{x}_i^k|$ by $\mathcal{O}(\frac{\sqrt{\nu}}{K^{1/4}})$, otherwise, term (i) would slow the convergence rate established in Theorem 1. Intuitively, when $\nabla_i f(\mathbf{x}^k)\approx 0$ such that $\widetilde{\mathbf{v}}_i^k=\beta^k\mathbf{v}_i^0+(1-\beta)\sum_{r=1}^k\beta^{k-r}\left(|\nabla_i f(\mathbf{x}^r)|^2+\sigma_i^2\right)\approx\sigma_i^2$, we have $\sum_{k=1}^K\sum_{i=1}^d\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\approx K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}$, making term (i) non-negligible in (12).

3.3 Eliminating ε in the Final Convergence Rate

Based on the above analysis, we get the following bound

$$\sum_{k=1}^K\sum_{i=1}^d\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\leq\mathcal{O}\left(\sqrt{\frac{K\sigma_s^2L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}\right).$$

Using Holder's inequality and (12) again, we finally have

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

$$\leq \left(\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\right)\left(K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}+2\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\right)$$
$$\leq \mathcal{O}\left(\sqrt{\frac{K\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}\left(\sqrt{\frac{K\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}+K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}\right)\right)$$

and

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\|\nabla f(\mathbf{x}^k)\|_1\right]$$
$$\leq \mathcal{O}\left(\frac{1}{K}\left(\sqrt{\frac{K\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}+\sqrt[4]{\underbrace{\frac{K\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon}\left(K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}\right)^2}_{\text{term (k)}}}\right)\right).$$

The above convergence rate is not optimal due to its explicit dependence on ε, which is absent from the optimal rate (4) of SGD. By setting $\varepsilon = \frac{\sigma_s^2}{d}$, we obtain $K\|\boldsymbol{\sigma}\|_1 \leq Kd\sqrt{\varepsilon}$ and $(K\|\boldsymbol{\sigma}\|_1 + Kd\sqrt{\varepsilon})^2 \leq 4K^2d^2\varepsilon$, which allows us to eliminate ε in the denominator of term (k). This yields the following final convergence rate

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

Although smaller value of ε does not affect the convergence of AdamW and Adam, this term cannot be eliminated any more and consequently slows the convergence rate by introducing explicit ε-dependence. On the other hand, while larger ε does not impact the convergence rate, it makes AdamW closer to SGD.

At last, in order to incorporate the scenario when $\sigma_s^2 \leq \frac{L(f(\mathbf{x}^1)-f^*)}{K}$, we define $\hat{\sigma}_s^2 = \max\left\{\sigma_s^2, \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}\right\}$ with any constant $\gamma \in (0,1]$ and replace σ_s^2 by $\hat{\sigma}_s^2$ in the definitions of ε, ν, $1-\theta$, and η.

4 Literature Comparisons

In this section, we compare our theoretical results with representative ones in the literature. A substantial amount of literature exists regarding the convergence analysis of adaptive gradient algorithms, such as [19, 20, 21, 22, 23] for AdaGrad-norm, [22, 11, 12, 24, 25] for AdaGrad, [26, 27, 28, 10, 13] for RMSProp, [29] for Adam-norm, [30, 26, 27, 31, 32, 14, 33, 15, 34, 35, 36, 37, 38, 39] for Adam, and [40, 41, 42, 43, 44, 45, 46, 47, 48, 49] for other variants. We primarily compare with the literature on AdamW and Adam. For Adam, we restrict our comparison to studies with the state-of-the-art convergence rates that do not require the bounded gradient assumption.

4.1 AdamW: Comparison with [7]

To the best of our knowledge based on a comprehensive literature review, [7] appears to be the only existing paper addressing AdamW's convergence and convergence rate. We compare with [7] in the following aspects. Firstly, the assumptions in [7] are stronger than ours. Denoting $f(\mathbf{x}) = \mathbb{E}_{\zeta\in D}[f(\mathbf{x};\zeta)]$, they assumed $\|\nabla f(\mathbf{y};\zeta) - \nabla f(\mathbf{x};\zeta)\| \leq L\|\mathbf{y}-\mathbf{x}\|$ (under which the lower bound is $\mathcal{O}(\frac{1}{\epsilon^3})$, rather than $\mathcal{O}(\frac{1}{\epsilon^4})$ [9]) and $\|\mathbf{g}^k\|_\infty \leq c_\infty$, while we only assume $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y}-\mathbf{x}\|$ without the bounded gradient assumption. Secondly, they set the weight decay parameter $\lambda_k = \lambda(1-\frac{\beta c_\infty^2}{\varepsilon})^k$, which decreases exponentially, making AdamW reduce to standard Adam in the limit. Thirdly, they establish the complexity of $\mathcal{O}(\max\{\frac{c_\infty^{2.5}L\sigma_s^2(f(\mathbf{x}^1)-f^*)}{\varepsilon^{1.25}\epsilon^4}, \frac{c_\infty^2\sigma_s^4}{\varepsilon\epsilon^4}\})$ to achieve $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla F_k(\mathbf{x}^k)\|^2] \leq \epsilon^2$, where F_k is a dynamic ℓ_2 regularized objective. Their complexity depends on ε explicitly, which is usually small in practice, for example, $\varepsilon \approx 10^{-16}$ in PyTorch implementation. As a comparison, our convergence rate does not depend on ε explicitly.

4.2 Adam: Comparison with [10]

Li et al. [10] studied RMSProp and its momentum extension, where RMSProp is a special case of Adam by letting $\theta = 0$ and $\lambda = 0$ in Algorithm 1. The convergence analysis of Adam presents substantially greater challenges than RMSProp and we cannot extend the proofs in [10] to Adam. Alternatively, this paper uses a different proof framework to establish for Adam the same convergence rate achieved by [10] under identical assumptions. As a trade-off, one limitation of our proof is that it relies on a larger value of parameter ε, although $\varepsilon = \frac{\hat{\sigma}_s^2}{d}$ is very small in practice. Specifically, under the parameter settings of $\beta = 1 - \frac{1}{K}$, $\mathbf{v}_i^0 = \lambda \max\{\sigma_i^2, \frac{1}{dK}\}$, and $\lambda \geq \frac{\sigma_s^2}{KL(f(\mathbf{x}^1)-f^*)}$ in [10], we have $\frac{1}{e^2} \leq \beta^t \leq 1$ for any $t \leq K$ and

$$\mathbf{v}_i^k = \beta^k \mathbf{v}_i^0 + (1-\beta)\sum_{t=1}^{k} \beta^{k-t} |\mathbf{g}_i^t|^2 \approx \frac{\sigma_i^2}{K} \frac{\sigma_s^2}{L(f(\mathbf{x}^1) - f^*)} + \frac{1}{K}\sum_{t=1}^{k} |\mathbf{g}_i^t|^2,$$

where $\beta^k \mathbf{v}_i^0$ plays the role of ε in Algorithm 1, which is of the order $\frac{\sigma_i^2}{K}$, or approximately $\frac{\sigma_s^2}{dK}$. As a comparison, in this paper, we have

$$\mathbf{v}_i^k + \varepsilon = \varepsilon + (1-\beta)\sum_{t=1}^{k} \beta^{k-t} |\mathbf{g}_i^t|^2 = \frac{\hat{\sigma}_s^2}{d} + (1-\beta)\sum_{t=1}^{k} \beta^{k-t} |\mathbf{g}_i^t|^2 \approx \sigma_i^2 + (1-\beta)\sum_{t=1}^{k} \beta^{k-t} |\mathbf{g}_i^t|^2.$$

When $\nabla f(\mathbf{x}^t) \approx 0$ such that $|\mathbf{g}_i^t| \approx \sigma_i$, we have $(1-\beta)\sum_{t=1}^{k} \beta^{k-t} |\mathbf{g}_i^t|^2 \approx \sigma_i^2$. Thus, ε accounts for nearly half of $(\mathbf{v}_i^k + \varepsilon)$'s size, while in [10], $\beta^k \mathbf{v}_i^0$ only makes up close to $\frac{1}{k}$ of $\mathbf{v}_i^k$'s total size. Other representative studies [34, 35, 33] have derived convergence guarantees for Adam built upon weak ε-dependent analysis. However, these results all yield slower convergence rates than ours with a higher dependence on the dimension d.

4.3 Adam: Comparison with [37]

Li et al. [37] studied Adam under assumption $\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\| \leq \sigma_s$ with probability 1 and proved $\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(\mathbf{x}^k)\|_2^2 \leq \epsilon^2$ with high probability within $\mathcal{O}(\frac{G^{2.5}\sigma_s^2 L(f(\mathbf{x}^1)-f^*)}{\widetilde{\varepsilon}^{2.5}\epsilon^4})$ iterations. That is, $\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(\mathbf{x}^k)\|_2 \leq (\frac{G}{\widetilde{\varepsilon}})^{5/8} \frac{1}{K^{1/4}} \sqrt[4]{\sigma_s^2 L(f(\mathbf{x}^1) - f^*)}$, where $G \geq \max\{\widetilde{\varepsilon}, \sigma_s, \sqrt{L(f(\mathbf{x}^1) - f^*)}\}$ and $\widetilde{\varepsilon}$ appeared in a different place in $\frac{\mathbf{m}^k}{\sqrt{\mathbf{v}^k + \widetilde{\varepsilon}}}$ (hence we may consider $\widetilde{\varepsilon}$ to be equal to $\sqrt{\varepsilon}$). When $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$, as empirically observed in real-world deep learning training, our convergence rate is $(\frac{G}{\widetilde{\varepsilon}})^{5/8}$ times faster than [37]. In PyTorch implementation, the default value of $\widetilde{\varepsilon}$ is typically set to 10^{-8}. To eliminate the dependence on ε, Li et al. [37] requires $\widetilde{\varepsilon}^2 (\approx \varepsilon) = G^2 \geq \max\{\sigma_s^2, L(f(\mathbf{x}^1) - f^*)\} \geq \sigma_s^2$, while we only need $\varepsilon = \frac{\sigma_s^2}{d}$, which is d times smaller.

Conclusion

This paper studies the popular AdamW optimizer in deep learning. We establish the convergence rate $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq \mathcal{O}(\frac{\sqrt{d}C}{K^{1/4}})$ for AdamW measured by ℓ_1 norm. It can be considered to be analogous to the optimal rate of SGD in the ideal case of $\|\nabla f(\mathbf{x})\|_1 = \Theta(\sqrt{d})\|\nabla f(\mathbf{x})\|_2$, which is verified on real-world deep learning tasks. An important direction for future research would be to investigate the optimal convergence rate using weak ε-dependent analysis (for example, $\log \frac{1}{\varepsilon}$) for AdamW and Adam. On the other hand, it is currently unclear whether our upper bound on λ is tight. Investigating how to prove the optimal convergence rate under a looser upper bound would be meaningful. This study is primarily concerned with theoretical analysis and it does not yield direct negative societal impacts.

Acknowledgements

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

References

[1] 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.

[2] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, 2010.

[3] 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.

[4] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[6] Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞ norm constrained optimization. In *International Conference on Machine Learning (ICML)*, 2024.

[7] Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6486–6493, 2024.

[8] Leon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[9] 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.

[10] Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\frac{\sqrt{d}}{T^{1/4}})$ convergence rate of RMSProp and its momentum extension measured by ℓ_1 norm. *Journal of Machine Learning Research*, 26(131):1–25, arXiv: 2402.00389, 2025.

[11] Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. In *Conference on Learning Theory (COLT), arXiv: 2406.04592*, 2025.

[12] Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under anisotropic smoothness. In *International Conference on Learning Representations (ICLR), arXiv: 2406.15244*, 2025.

[13] Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *International Conference on Learning Representations (ICLR), arXiv: 2410.08198*, 2025.

[14] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhiquan Luo. Adam can converge without any modification on update rules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[15] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of Adam under non-uniform smoothness. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[16] Hoang Tran, Qinzi Zhang, and Ashok Cutkosky. Empirical tests of optimization assumptions in deep learning. *arXiv: 2407.01825*, 2024.

[17] Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv: 2509.02046*, 2025.

[18] Yiming Dong, Huan Li, and Zhouchen Lin. Convergence rate analysis of LION. *arXiv: 2411.07724*, 2024.

[19] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[20] 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.

[21] 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.

[22] 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)*, 2023.

[23] 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.

[24] 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.

[25] Yusu Hong and Junhong Lin. Revisiting convergence of AdaGrad with relaxed assumptions. In *Uncertainty in Artificial Intelligence (UAI)*, 2024.

[26] 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.

[27] Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research*, 2022.

[28] Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyper-parameter. In *International Conference on Learning Representations (ICLR)*, 2020.

[29] Bohan Wang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, and Wei Chen. On the convergence of Adam under non-uniform smoothness: Separability from SGDM and beyond. *arXiv: 2403.15146*, 2024.

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

[31] 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.

[32] 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.

[33] 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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[34] Yusu Hong and Junhong Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *arXiv: 2311.02000*, 2023.

[35] Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[36] 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.

[37] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[38] Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[39] Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of Adam: Bridging the gap with SGD. In *International Conference on Machine Learning (ICML)*, 2025.

[40] Manzil Zaheer, Sashank J.Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Conference on Neural Information Processing Systems (NIPS)*, 2018.

[41] 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.

[42] 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.

[43] 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.

[44] 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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[45] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan 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.

[46] 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.

[47] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[48] 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.

[49] Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. Adopt: Modified Adam can converge with any β_2 with the optimal rate. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[54] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. OpenWebText corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

[55] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[56] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv:1909.08053*, 2019.

A Proof of Theorem 1

Proof 1 *As the gradient is L-Lipschitz, we have*

$$
\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{x}^{k+1})|\mathcal{F}_{k-1}\right]-f(\mathbf{x}^k)\\
\leq&\mathbb{E}_k\left[\left\langle\nabla f(\mathbf{x}^k),\mathbf{x}^{k+1}-\mathbf{x}^k\right\rangle+\frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\Big|\mathcal{F}_{k-1}\right]\\
=&\mathbb{E}_k\left[-\eta\sum_{i=1}^d\left\langle\nabla_i f(\mathbf{x}^k),\frac{\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\right\rangle+\frac{L\eta^2}{2}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\mathbf{v}_i^k+\varepsilon}\Big|\mathcal{F}_{k-1}\right]\\
=&\mathbb{E}_k\Bigg[-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\\
&\quad+\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)-\mathbf{m}_i^k-\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}+\frac{L\eta^2}{2}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\mathbf{v}_i^k+\varepsilon}\Big|\mathcal{F}_{k-1}\Bigg]\\
\leq&\mathbb{E}_k\Bigg[-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\\
&\quad+\eta\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)-\mathbf{m}_i^k\right|^2+\left|\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}+\frac{L\eta^2}{2\sqrt{\varepsilon}}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\Big|\mathcal{F}_{k-1}\Bigg]\\
\overset{(1)}{\leq}&\mathbb{E}_k\Bigg[-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}-\frac{\eta}{4}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}+\frac{\eta}{\sqrt{\varepsilon}}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2\\
&\quad+\eta\sum_{i=1}^d|\lambda\mathbf{x}_i^k|^2\sqrt{\mathbf{v}_i^k+\varepsilon}\Big|\mathcal{F}_{k-1}\Bigg]\\
\overset{(2)}{\leq}&\mathbb{E}_k\Bigg[-\frac{\eta}{2}\sum_{i=1}^d\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}-\frac{\eta}{4}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}+\frac{\eta}{\sqrt{\varepsilon}}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2\\
&\quad+\frac{9\eta\nu}{K^{1/2}}\sum_{i=1}^d\sqrt{\mathbf{v}_i^k+\varepsilon}\Big|\mathcal{F}_{k-1}\Bigg],
\end{aligned}
\tag{13}
$$

where we let $\eta\leq\frac{\sqrt{\varepsilon}}{2L}$ *in* $\overset{(1)}{\leq}$ *and use Lemma 3 in* $\overset{(2)}{\leq}$*. Denote*

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

From the concavity of $\sqrt{x}$ *and* $-\frac{1}{\sqrt{x}}$ *and Assumptions 2 and 3, we have*

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

Plugging into (13) and rearranging the terms, we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{x}^{k+1})-f^*+\frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}-\frac{\eta}{\sqrt{\varepsilon}}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2\Big|\mathcal{F}_{k-1}\right]\\
\leq& f(\mathbf{x}^k)-f^*-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}+\frac{9\eta\nu}{K^{1/2}}\sum_{i=1}^{d}\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}.
\end{aligned}\tag{14}$$

Multiplying both sides of (18) in Lemma 4 by $\frac{\eta}{\sqrt{\varepsilon}(1-\theta)}$ and adding it to (14), we have

$$\begin{aligned}
&\mathbb{E}_k\left[f(\mathbf{x}^{k+1})-f^*+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^k)-\mathbf{m}^k\right\|^2+\frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^k+\lambda\mathbf{x}_i^k\sqrt{\mathbf{v}_i^k+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^k+\varepsilon}}\Big|\mathcal{F}_{k-1}\right]\\
\leq& f(\mathbf{x}^k)-f^*-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}+\frac{9\eta\nu}{K^{1/2}}\sum_{i=1}^{d}\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^{k-1})-\mathbf{m}^{k-1}\right\|^2\\
&+\frac{L^2\eta^3}{\varepsilon(1-\theta)^2}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}}+\frac{\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}\\
\leq& f(\mathbf{x}^k)-f^*-\frac{\eta}{2}\sum_{i=1}^{d}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}+\frac{9\eta\nu}{K^{1/2}}\sum_{i=1}^{d}\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\\
&+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^{k-1})-\mathbf{m}^{k-1}\right\|^2+\frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}}+\frac{\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}},
\end{aligned}\tag{15}$$

where we let $\eta^2\leq\frac{\varepsilon(1-\theta)^2}{4L^2}$ in the last inequality. For both (14) and (15), taking expectation with respect to $\mathcal{F}_{k-1}$, rearranging the terms, summing (14) with $k=1$ and (15) over $k=2,3,\cdots,K$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{x}^{K+1})-f^*+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^K)-\mathbf{m}^K\right\|^2+\frac{\eta}{4}\sum_{i=1}^{d}\frac{\left|\mathbf{m}_i^K+\lambda\mathbf{x}_i^K\sqrt{\mathbf{v}_i^K+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^K+\varepsilon}}\right]\\
\leq& f(\mathbf{x}^1)-f^*+\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[-\frac{\eta}{2}\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}+\frac{9\eta\nu}{K^{1/2}}\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\right]\\
&+\frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1)-\mathbf{m}^1\|^2\right]+\frac{\eta}{\sqrt{\varepsilon}}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1)-\mathbf{m}^1\|^2\right]+\frac{(K-1)\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}\\
\overset{(3)}{\leq}& f(\mathbf{x}^1)-f^*-\frac{\eta}{2}\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\\
&+\frac{9\eta\nu}{K^{1/2}}\left(K\|\boldsymbol{\sigma}\|_1+Kd\sqrt{\varepsilon}+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{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\right)\\
&+\frac{\eta}{\sqrt{\varepsilon}(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1)-\mathbf{m}^1\|^2\right]+\frac{(K-1)\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}\\
\overset{(4)}{\leq}& f(\mathbf{x}^1)-f^*-\frac{\eta}{4}\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\\
&+18\eta\nu d\sqrt{K\varepsilon}+\frac{\eta}{\sqrt{\varepsilon}(1-\theta)}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1)-\mathbf{m}^1\|^2\right]+\frac{(K-1)\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}\\
\overset{(5)}{\leq}& f(\mathbf{x}^1)-f^*-\frac{\eta}{4}\sum_{k=1}^{K}\sum_{i=1}^{d}\mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right]\\
&+18\eta\nu d\sqrt{K\varepsilon}+\frac{2\eta L(f(\mathbf{x}^1)-f^*)}{\sqrt{\varepsilon}(1-\theta)}+\frac{\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}++\frac{(K-1)\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}},
\end{aligned}\tag{16}$$

where we use Lemma 5 in $\overset{(3)}{\leq}$*, let* $\frac{9\nu}{K^{1/2}} \leq \frac{1}{8}$ *and* $\varepsilon \geq \frac{\sigma_s^2}{d}$ *such that* $\|\boldsymbol{\sigma}\|_1 \leq \sqrt{d}\|\boldsymbol{\sigma}\|_2 = \sqrt{d}\sigma_s \leq d\sqrt{\varepsilon}$ *in* $\overset{(4)}{\leq}$*, and use* $\mathbf{m}^0 = 0$*,*

$$f^* \leq f\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) \leq f(\mathbf{x}) - \frac{1}{L}\left\langle \nabla f(\mathbf{x}), \nabla f(\mathbf{x})\right\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\mathbf{x})\right\|^2 = f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2,$$

and

$$\begin{aligned}\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1) - \mathbf{m}^1\|^2\right] =& \mathbb{E}_{\mathcal{F}_1}\left[\|\theta\nabla f(\mathbf{x}^1) + (1-\theta)(\nabla f(\mathbf{x}^1) - \mathbf{g}^1)\|^2\right]\\ =& \theta^2\|\nabla f(\mathbf{x}^1)\|^2 + (1-\theta)^2\mathbb{E}_{\mathcal{F}_1}\left[\|\nabla f(\mathbf{x}^1) - \mathbf{g}^1\|^2\right]\\ \leq& 2L(f(\mathbf{x}^1) - f^*) + (1-\theta)^2\sigma_s^2\end{aligned}$$

in $\overset{(5)}{\leq}$*. So from (16), we have*

$$\begin{aligned}&\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}}\right]\\ \leq& \frac{4(f(\mathbf{x}^1) - f^*)}{\eta} + 72\nu d\sqrt{K\varepsilon} + \frac{8L(f(\mathbf{x}^1) - f^*)}{\sqrt{\varepsilon}(1-\theta)} + \frac{4K(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}}\\ \leq& \frac{4(f(\mathbf{x}^1) - f^*)}{\eta} + 72\nu d\sqrt{K\varepsilon} + \frac{8L(f(\mathbf{x}^1) - f^*)}{\sqrt{\varepsilon}(1-\theta)} + \frac{4K(1-\theta)\hat{\sigma}_s^2}{\sqrt{\varepsilon}},\end{aligned} \tag{17}$$

where we denote $\hat{\sigma}_s^2 = \max\left\{\sigma_s^2, \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}\right\}$ *with any constant* $\gamma \in (0, 1]$*.*

Recall that we require the parameters satisfying the following relations in the above proof

$$\eta \leq \frac{\sqrt{\varepsilon}}{2L}, \quad \eta^2 \leq \frac{\varepsilon(1-\theta)^2}{4L^2}, \quad \frac{9\nu}{K^{1/2}} \leq \frac{1}{8}, \quad \varepsilon \geq \frac{\sigma_s^2}{d}$$

and

$$\eta\lambda \leq \frac{\sqrt{\nu}}{2K^{5/4}}, \quad \frac{\sqrt{\nu}}{K^{1/4}} < 1, \quad \|\mathbf{x}^1\|_\infty \leq \frac{\sqrt{\nu}}{K^{1/4}\lambda}, \quad \theta \leq \beta \leq \sqrt{\theta} < 1$$

in Lemma 3.

Recalling the definition of $\hat{\sigma}_s$ *and letting* $\varepsilon = \frac{\hat{\sigma}_s^2}{d}$*,* $1 - \theta = \sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{K\hat{\sigma}_s^2}}$*,* $\eta = \sqrt{\frac{\varepsilon(f(\mathbf{x}^1)-f^*)}{4K\hat{\sigma}_s^2 L}} = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}}$*,* $\nu = \frac{1}{72d}\sqrt{\frac{\hat{\sigma}_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon^2}} = \frac{1}{72}\sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{\hat{\sigma}_s^2}}$*,* $\lambda \leq \frac{\sqrt{\nu}}{2K^{5/4}\eta} = \frac{\sqrt{d}}{\sqrt{72}K^{3/4}}\sqrt[4]{\frac{L^3}{\hat{\sigma}_s^2(f(\mathbf{x}^1)-f^*)}}$*, and* $\|\mathbf{x}^1\|_\infty \leq \sqrt{\frac{K(f(\mathbf{x}^1)-f^*)}{dL}} = 2K\eta \leq \frac{\sqrt{\nu}}{K^{1/4}\lambda}$*, the above requirements are satisfied. So we have from (17) that*

$$\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}}\right] \leq 21\sqrt{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}{\varepsilon}}.$$

Using Holder's inequality and Lemma 5, we have

$$\begin{aligned}&\left(\sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}}\left[\|\nabla f(\mathbf{x}^k)\|_1\right]\right)^2\\ \leq& \left(\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}}\right]\right)\left(\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}\right]\right)\\ \leq& \left(\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}}\right]\right)\left(K\|\boldsymbol{\sigma}\|_1 + Kd\sqrt{\varepsilon} + 2\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k + \varepsilon}}\right]\right)\\ \leq& \left(21\sqrt{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}{\varepsilon}}\right)\left(42\sqrt{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}{\varepsilon}} + K\|\boldsymbol{\sigma}\|_1 + Kd\sqrt{\varepsilon}\right)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{\mathcal{F}_{k-1}}\left[\|\nabla f(\mathbf{x}^k)\|_1\right]\\ \leq& \frac{1}{K}\left(30\sqrt{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}{\varepsilon}} + 5\sqrt[4]{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}{\varepsilon}\left(K\|\boldsymbol{\sigma}\|_1 + Kd\sqrt{\varepsilon}\right)^2}\right)\\ \leq& 30\sqrt{\frac{dL(f(\mathbf{x}^1) - f^*)}{K}} + \frac{8\sqrt{d}}{K^{1/4}}\sqrt[4]{\hat{\sigma}_s^2 L(f(\mathbf{x}^1) - f^*)}\end{aligned}$$

by letting $\varepsilon = \frac{\hat{\sigma}_s^2}{d}$ *and using* $\|\boldsymbol{\sigma}\|_1 \leq \sqrt{d}\|\boldsymbol{\sigma}\|_2 = \sqrt{d}\sigma_s \leq d\sqrt{\varepsilon}$*. At last, from Lemma 3 and the settings of* ν *and* $\hat{\sigma}_s$*, we have*

$$\lambda\|\mathbf{x}^k\|_\infty \leq \frac{3\sqrt{\nu}}{K^{1/4}} = \frac{3}{\sqrt{72}}\sqrt[4]{\frac{L(f(\mathbf{x}^1)-f^*)}{K\hat{\sigma}_s^2}} \leq \frac{3}{\sqrt{72}}$$

for all $k = 1, 2, \cdots, K$*, leading to* $\|\mathbf{x}^k\|_\infty < \frac{1}{\lambda}$*.*

B Proof of Corollary 1

We give the complete description of Corollary 1 in the following corollary.

Corollary 2 *Suppose that Assumptions 1-3 hold. Define* $\hat{\sigma}_s^2 = \max\left\{\sigma_s^2, \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}\right\}$ *with any constant* $\gamma \in (0, 1]$*. Let* $1-\theta = \sqrt{\frac{L(f(\mathbf{x}^1)-f^*)}{K\hat{\sigma}_s^2}}$*,* $0 \leq \beta \leq 1$*,* $\eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4dKL}}$*, and* $\varepsilon = \frac{\hat{\sigma}_s^2}{d}$*. Then for Adam, we have*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq \frac{6\sqrt{d}}{K^{1/4}}\sqrt[4]{\hat{\sigma}_s^2 L(f(\mathbf{x}^1)-f^*)} + 15\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K}}.$$

Specially, when $\sigma_s^2 \leq \frac{L(f(\mathbf{x}^1)-f^*)}{K\gamma^2}$*, we have* $1-\theta = \gamma$*,* $0 \leq \beta \leq 1$*,* $\eta = \sqrt{\frac{f(\mathbf{x}^1)-f^*}{4KdL}}$*,* $\varepsilon = \frac{L(f(\mathbf{x}^1)-f^*)}{dK\gamma^2}$*, and accordlingly*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_1\right] \leq 21\sqrt{\frac{dL(f(\mathbf{x}^1)-f^*)}{K\gamma}}.$$

Proof 2 *When* $\lambda = 0$*, the* $\frac{9\eta\nu}{K^{1/2}}\sum_{i=1}^d \sqrt{\mathbf{v}_i^k + \varepsilon}$ *term disappears in (13) in the proof of Theorem 1, and (16) becomes*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_K}\left[f(\mathbf{x}^{K+1}) - f^* + \frac{\eta\theta}{\sqrt{\varepsilon}(1-\theta)}\left\|\nabla f(\mathbf{x}^K) - \mathbf{m}^K\right\|^2 + \frac{\eta}{4}\sum_{i=1}^d \frac{\left|\mathbf{m}_i^K + \lambda\mathbf{x}_i^K\sqrt{\mathbf{v}_i^K+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^K+\varepsilon}}\right] \\
\leq& f(\mathbf{x}^1) - f^* - \frac{\eta}{2}\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right] + \frac{2\eta L(f(\mathbf{x}^1)-f^*)}{\sqrt{\varepsilon}(1-\theta)} + \frac{K\eta(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}},
\end{aligned}$$

where the term $18\eta\nu d\sqrt{K\varepsilon}$ *disappears because we do not need Lemma 5 to bound* $\frac{9\eta\nu}{K^{1/2}}\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}\right]$ *any more.*

Similar to the proof of Theorem 1, we have

$$\begin{aligned}
\sum_{k=1}^K\sum_{i=1}^d \mathbb{E}_{\mathcal{F}_{k-1}}\left[\frac{\left|\nabla_i f(\mathbf{x}^k)\right|^2}{\sqrt{\widetilde{\mathbf{v}}_i^k+\varepsilon}}\right] \leq& \frac{2(f(\mathbf{x}^1)-f^*)}{\eta} + \frac{4L(f(\mathbf{x}^1)-f^*)}{\sqrt{\varepsilon}(1-\theta)} + \frac{2K(1-\theta)\sigma_s^2}{\sqrt{\varepsilon}} \\
\leq& 10\sqrt{\frac{K\hat{\sigma}_s^2 L(f(\mathbf{x}^1)-f^*)}{\varepsilon}}.
\end{aligned}$$

Comparing with (17), we see that the term $72\nu d\sqrt{K\varepsilon}$ *disappears. Following the proof of Theorem 1, we have the conclusion. Note that we do not use Lemmas 2 and 3 in the proof of Corollary 1, so Corollary 1 does not require* $\theta \leq \beta \leq \sqrt{\theta}$ *and* $\|\mathbf{x}^1\|_\infty \leq \frac{\sqrt{\nu}}{K^{1/4}\lambda}$ *any more.*

C Supporting Lemmas

Lemma 2 *Suppose* $\mathbf{m}^0 = 0$*,* $\mathbf{v}^0 = 0$*, and* $\theta \leq \beta \leq \sqrt{\theta} < 1$*, then we have*

$$\frac{|\mathbf{m}_i^k|^2}{\mathbf{v}_i^k} \leq \frac{(1-\theta)^2\beta}{(1-\beta)(\beta-\theta^2)} \leq 4.$$

Proof 3 *From the recursions of* $\mathbf{m}_i^k$ *and* $\mathbf{v}_i^k$*, we have*

$$\begin{aligned}
\mathbf{m}_i^k &= \theta^k\mathbf{m}_i^0 + (1-\theta)\sum_{r=1}^k \theta^{k-r}\mathbf{g}_i^r = (1-\theta)\sum_{r=1}^k \theta^{k-r}\mathbf{g}_i^r, \\
\mathbf{v}_i^k &= \beta^k\mathbf{v}_i^0 + (1-\beta)\sum_{r=1}^k \beta^{k-r}|\mathbf{g}_i^r|^2 = (1-\beta)\sum_{r=1}^k \beta^{k-r}|\mathbf{g}_i^r|^2.
\end{aligned}$$

Using Holder's inequality, we have

$$\begin{aligned}|\mathbf{m}_i^k|^2 =&(1-\theta)^2\left(\sum_{r=1}^k \theta^{k-r}\mathbf{g}_i^r\right)^2 \le (1-\theta)^2\left(\sum_{r=1}^k \beta^{k-r}|\mathbf{g}_i^r|^2\right)\left(\sum_{r=1}^k\left(\frac{\theta^2}{\beta}\right)^{k-r}\right)\\ =&\mathbf{v}_i^k\frac{(1-\theta)^2}{1-\beta}\sum_{r=1}^k\left(\frac{\theta^2}{\beta}\right)^{k-r} \le \mathbf{v}_i^k\frac{(1-\theta)^2}{1-\beta}\frac{1}{1-\frac{\theta^2}{\beta}} \overset{(1)}{\le} \mathbf{v}_i^k\frac{(1-\theta)^2}{(1-\beta)^2}\\ \overset{(2)}{\le}&\mathbf{v}_i^k\frac{(1-\sqrt{\theta})^2(1+\sqrt{\theta})^2}{(1-\sqrt{\theta})^2} \le \mathbf{v}_i^k(1+\sqrt{\theta})^2 \le 4\mathbf{v}_i^k,\end{aligned}$$

where we use $\theta \le \beta$ in $\overset{(1)}{\le}$ and $\beta \le \sqrt{\theta}$ in $\overset{(2)}{\le}$.

Lemma 3 *Suppose $\eta\lambda \le \frac{\sqrt{\nu}}{2K^{5/4}}$, $\|\mathbf{x}^1\|_\infty \le \frac{\sqrt{\nu}}{K^{1/4}\lambda}$, $\frac{\sqrt{\nu}}{K^{1/4}} < 1$, and $\theta \le \beta \le \sqrt{\theta} < 1$, then we have*

$$\lambda\|\mathbf{x}^k\|_\infty \le \frac{3\sqrt{\nu}}{K^{1/4}}, \quad \forall k = 1, 2, \cdots, K.$$

Proof 4 *From the update of $\mathbf{x}^{k+1}$, we have*

$$\begin{aligned}\|\mathbf{x}^{k+1}\|_\infty - \frac{2}{\lambda} =&\left\|(1-\eta\lambda)\mathbf{x}^k - \frac{\eta}{\sqrt{\mathbf{v}^k+\varepsilon}}\odot\mathbf{m}^k\right\|_\infty - \frac{2}{\lambda}\\ \le&(1-\eta\lambda)\|\mathbf{x}^k\|_\infty + \left\|\frac{\eta}{\sqrt{\mathbf{v}^k+\varepsilon}}\odot\mathbf{m}^k\right\|_\infty - \frac{2}{\lambda}\\ \overset{(1)}{\le}&(1-\eta\lambda)\|\mathbf{x}^k\|_\infty + 2\eta - \frac{2}{\lambda}\\ =&(1-\eta\lambda)\left(\|\mathbf{x}^k\|_\infty - \frac{2}{\lambda}\right)\\ \le&(1-\eta\lambda)^k\left(\|\mathbf{x}^1\|_\infty - \frac{2}{\lambda}\right)\\ \le&-\frac{1}{\lambda}(1-\eta\lambda)^k\left(2-\frac{\sqrt{\nu}}{K^{1/4}}\right),\end{aligned}$$

where we use Lemma 2 in $\overset{(1)}{\le}$. Since $\ln x \le x-1$ and $e^x \ge x+1$ for any $x > 0$ and $\eta\lambda \le \frac{\sqrt{\nu}}{2K^{5/4}} \le \frac{1}{2}$, we have for any $k \le K$ that

$$k\ln(1-\eta\lambda) = -k\ln\frac{1}{1-\eta\lambda} \ge -K\left(\frac{1}{1-\eta\lambda}-1\right) = -\frac{K\eta\lambda}{1-\eta\lambda} \ge -\frac{\sqrt{\nu}}{K^{1/4}},$$

$$(1-\eta\lambda)^k \ge e^{-\frac{\sqrt{\nu}}{K^{1/4}}} \ge 1 - \frac{\sqrt{\nu}}{K^{1/4}},$$

and

$$\|\mathbf{x}^{k+1}\|_\infty - \frac{2}{\lambda} \le -\frac{1}{\lambda}\left(1-\frac{\sqrt{\nu}}{K^{1/4}}\right)\left(2-\frac{\sqrt{\nu}}{K^{1/4}}\right) \le -\frac{2}{\lambda} + \frac{3}{\lambda}\frac{\sqrt{\nu}}{K^{1/4}}.$$

Lemma 4 *Suppose that Assumptions 1-3 hold. Then we have*

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\mathbf{m}^k - \nabla f(\mathbf{x}^k)\right\|^2|\mathcal{F}_{k-1}\right]\\ \le&\theta\left\|\mathbf{m}^{k-1} - \nabla f(\mathbf{x}^{k-1})\right\|^2 + \frac{L^2\eta^2}{\sqrt{\varepsilon}(1-\theta)}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^{k-1} + \lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}} + (1-\theta)^2\sigma_s^2.\end{aligned} \tag{18}$$

Proof 5 *Denoting $\zeta^k = \mathbf{g}^k - \nabla f(\mathbf{x}^k)$, from the update of $\mathbf{m}^k$, we have*

$$\begin{aligned}\mathbf{m}^k - \nabla f(\mathbf{x}^k) =&\theta\mathbf{m}^{k-1} + (1-\theta)\mathbf{g}^k - \nabla f(\mathbf{x}^k)\\ =&\theta\left(\mathbf{m}^{k-1} - \nabla f(\mathbf{x}^{k-1})\right) + (1-\theta)\left(\nabla f(\mathbf{x}^k) + \zeta^k\right) - \nabla f(\mathbf{x}^k) + \theta\nabla f(\mathbf{x}^{k-1})\\ =&\theta\left(\mathbf{m}^{k-1} - \nabla f(\mathbf{x}^{k-1})\right) + (1-\theta)\zeta^k - \theta\left(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})\right)\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}_k\left[\|\mathbf{m}^k-\nabla f(\mathbf{x}^k)\|^2|\mathcal{F}_{k-1}\right]\\
\leq&\left\|\theta\left(\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right)-\theta\left(\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^{k-1})\right)\right\|^2+(1-\theta)^2\sigma_s^2\\
\leq&\theta^2\left(\left(1+\frac{1-\theta}{\theta}\right)\left\|\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right\|^2+\left(1+\frac{\theta}{1-\theta}\right)\left\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^{k-1})\right\|^2\right)+(1-\theta)^2\sigma_s^2\\
\leq&\theta\left\|\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right\|^2+\frac{L^2}{1-\theta}\left\|\mathbf{x}^k-\mathbf{x}^{k-1}\right\|^2+(1-\theta)^2\sigma_s^2\\
=&\theta\left\|\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right\|^2+\frac{L^2\eta^2}{1-\theta}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\mathbf{v}_i^{k-1}+\varepsilon}+(1-\theta)^2\sigma_s^2\\
\leq&\theta\left\|\mathbf{m}^{k-1}-\nabla f(\mathbf{x}^{k-1})\right\|^2+\frac{L^2\eta^2}{\sqrt{\varepsilon}(1-\theta)}\sum_{i=1}^d\frac{\left|\mathbf{m}_i^{k-1}+\lambda\mathbf{x}_i^{k-1}\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}\right|^2}{\sqrt{\mathbf{v}_i^{k-1}+\varepsilon}}+(1-\theta)^2\sigma_s^2.
\end{aligned}
$$

The following lemma is modified from [10]. We give the proof here only for the sake of completeness.

Lemma 5 *Suppose that Assumptions 1-3 hold. Let $\beta\leq 1$ and $\mathbf{v}^0=0$. Then we have*

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

Proof 6 *From the definition of $\widetilde{\mathbf{v}}_i^k$, we have*

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

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

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

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

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

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

and

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

where we use $\mathbf{v}_i^0=0$. Summing over $i=1,2,\cdots,d$ and $k=1,2,\cdots,K$, we have

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

Lemma 6 *When each entry of $\mathbf{x}\in\mathbb{R}^d$ is generated from Gaussian distribution with zero mean and unit variance, we have $\mathbb{E}\left[\|\mathbf{x}\|_1\right]\geq\sqrt{\frac{2d}{\pi}}\mathbb{E}\left[\|\mathbf{x}\|_2\right]$.*

Proof 7 *When $\mathbf{x}_i\sim\mathcal{N}(0,1)$, we have*

$$
\begin{aligned}
& \mathbb{E}\left[|\mathbf{x}_i|\right]=\sqrt{\frac{2}{\pi}},\quad \mathbb{E}\left[\mathbf{x}_i^2\right]=1, \\
& \mathbb{E}\left[\|\mathbf{x}\|_1\right]=\sum_{i=1}^{d}\mathbb{E}\left[|\mathbf{x}_i|\right]=d\sqrt{\frac{2}{\pi}}, \\
& \mathbb{E}\left[\|\mathbf{x}\|_2^2\right]=\sum_{i=1}^{d}\mathbb{E}\left[\mathbf{x}_i^2\right]=d, \\
& \mathbb{E}\left[\|\mathbf{x}\|_2\right]=\mathbb{E}\left[\sqrt{\|\mathbf{x}\|_2^2}\right]\overset{(1)}{\leq}\sqrt{\mathbb{E}\left[\|\mathbf{x}\|_2^2\right]}=\sqrt{d}, \\
& \frac{\mathbb{E}\left[\|\mathbf{x}\|_1\right]}{\mathbb{E}\left[\|\mathbf{x}\|_2\right]}\geq\sqrt{\frac{2d}{\pi}}.
\end{aligned}
$$

where we use the concavity of $\sqrt{x}$ in $\overset{(1)}{\leq}$.

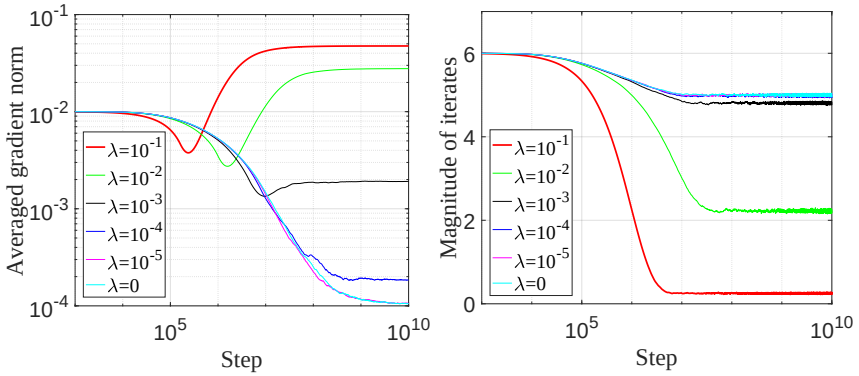

Figure 5: Illustrations of $\frac{1}{k}\sum_{t=1}^{k}|\nabla f(x^t)|$ (left) and x^k (right) over steps on the toy example.

D A Toy Example with Large λ

Consider the following function:

$$f(x)=\frac{(x-x^*)^2}{200},\ \text{with the stochastic gradient}\ g(x)=\begin{cases} x-x^*-1, & \text{with probability } p=0.1,\\ -\frac{1}{10}(x-x^*-\frac{10}{9}), & \text{with probability } 1-p.\end{cases}$$

We set $K=10^{10}$, $\theta=1-\frac{1}{\sqrt{K}}$, $\beta=\sqrt{\theta}$, $\eta=\frac{1}{\sqrt{K}}$, $\varepsilon=10^{-10}$, $m^0=0$, $v^0=0$, and $x^1=x^*+1$ for AdamW, where $x^*=5$ is the minimum solution of $f(x)$. We test $\lambda=\{10^{-1},10^{-2},10^{-3},10^{-4},10^{-5},0\}$ such that $x^*<\frac{1}{\lambda}$ and thus the KKT conditions (2) reduce to $|\nabla f(x^*)|=0$ at the minimum solution. So we can use the gradient norm $|\nabla f(x)|$ to measure the convergence. From Figure 5, we see that AdamW fails to converge to x^* when $\lambda=\{10^{-1},10^{-2},10^{-3}\}$, indicating that large values of λ exceeding a certain threshold may cause AdamW neither to converge to the minimum solution nor to a KKT point satisfying (2)[3]. In practical implementations, excessively large values of λ are typically avoided, as they may drive the parameters toward zero and away from the minimum solution.

E Experimental Details

In the main paper, we conduct several representative deep learning experiments to empirically support our claims, covering classic image classification and language processing tasks. For the vision tasks, we independently train ResNet50 [50] on CIFAR100 [51] and ImageNet [52] datasets; For the language task, we adopt the GPT-2 [53] architecture and pretrain it on the OpenWebText [54] dataset. Code is released at https://github.com/adonis-dym/Convergence-Rate-AdamW .

Our experiments involve the computation of the full training loss $f(\mathbf{x}^k)$ as well as the full gradient $\nabla f(\mathbf{x}^k)$. However, in the typical stochastic training paradigm, one often updates the parameter $\mathbf{x}^k$ on-the-fly immediately after obtaining the stochastic gradient $\mathbf{g}^k$ from the backward pass. To get an accurate measurement and avoid interfering with the normal training process, we propose to split each epoch into two separate phases: *training phase* and *logging phase*. In the training phase, we traverse the dataset once with stochastic updates, where the model parameters are updated upon processing each mini-batch. In the logging phase, we conduct a second traversal over the training dataset while keeping the model parameters frozen. Since the loss function is typically defined to be the average over all training samples and the gradient computation is inherently linear, we accumulate the losses and stochastic gradients across mini-batches during this phase. This yields the exact values of the full training loss $f(\mathbf{x}^k)$ and full gradient $\nabla f(\mathbf{x}^k)$ at the current iteration.

In the following, we detail each experimental setup individually:

i) ResNet50 - CIFAR100: CIFAR100 is a simple benchmark dataset that is widely used for quick and efficient evaluation of deep learning tasks. It contains a training split of 50000 examples and a test split of 10000 examples, although we do not perform evaluation on the test set in this work. Following the official implementation, we use the `torch.optim.AdamW` API to configure the optimizer. We initialize the learning rate to 3×10^{-3}, train the ResNet50 model for 100 epochs, and apply a cosine learning rate decay schedule during the whole training process. Setting the batch size to 128, each epoch consists of $\lfloor 50000/128\rfloor+1=391$ steps, where the additional step accounts for the final truncated batch which contains the remaining samples. The total number of

[3]This does not conflict with [6] because [6] only considered deterministic AdamW.

steps is $K = 391 \times 100 = 39100$. Without loss of generality, we compute the noise vector $\boldsymbol{\sigma}^k = \mathbf{g}^k - \nabla f(\mathbf{x}^k)$ using the stochastic gradient $\mathbf{g}^k$ obtained from the first batch at the logging phase. We leave the weight decay λ as its default value 0.01, and complete the training task with a single NVIDIA A100 GPU.

ii) ResNet50 - ImageNet: To evaluate the scalability of our conclusions on larger-scale dataset, we conduct experiments on the ImageNet dataset using the same ResNet50 architecture. ImageNet consists of approximately 1.28 million training images and 50,000 validation images across 1,000 classes, which also come with an official dataset split. We employ the training script from PyTorch Image Models (`timm`) [55], making only the necessary modifications to suit our experimental setup. We adopt the same optimizer configuration as previously, but compute the noise vector using the last batch at the logging phase, as the `timm` script discards incomplete batch and ensures uniform batch sizes. We follow the standard ImageNet training protocol for ResNet-50, which consists of 90 epochs as commonly adopted in the literature and official implementations [50, 55]. The first 10 epochs are used for learning rate linear warmup from 0 to 3×10^{-3}, followed by cosine decay over the remaining 80 epochs. We apply standard data augmentation techniques including RandAugment, Mixup (0.1), and CutMix (1.0). Setting the batchsize to 4096, each epoch consists of 312 minibatches and the total number of steps is $K = 28080$. We set $\lambda = 0.1$ and complete the training task using 8 NVIDIA A100 GPUs.

iii) GPT2 - OpenWebText: To assess the generality of our conclusions across different modalities, we further evaluate on a language modeling task using GPT-2. We pretrain this model on the OpenWebText dataset under the NVIDIA Megatron-LM codebase [56], which is a widely adopted framework for large-scale language model training. Unlike the previous settings, where computing the full training loss and gradient over the entire dataset is tractable, the OpenWebText dataset is substantially larger, containing approximately 9 billion tokens. Consequently, an entire pass through the dataset to get the full training loss $f(\mathbf{x}^k)$ and gradient $\nabla f(\mathbf{x}^k)$ is computationally infeasible. Instead, we approximate these quantities by accumulating their values over 100 consecutive mini-batches at the logging phase. We follow the Megatron-LM official GPT-2 training configuration with minimal modifications to suit our experimental needs. We train a GPT-2 Small model with approximately 125M parameters. The model is optimized using the fused implementation of AdamW from NVIDIA Apex package, which is the default setting in Megatron-LM. We set the learning rate to 3×10^{-3} and weight decay to 0.05. Following the de facto standard in large-scale language model training, we use $(\beta_1, \beta_2) = (0.9, 0.95)$ instead of the conventional $(0.9, 0.999)$ setting. The total training process runs for 50,000 iterations, where the learning rate is linearly warmed up for the first 2,000 iterations and then decayed following a cosine schedule. We set the global batch size to 640 and train the model for $K = 50000$ steps, and complete the training task using 8 NVIDIA A100 GPUs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contribution is to establish the convergence rate of AdamW. It is clearly written in the abstract and introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 4.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are provided in Section 2 and the proofs are provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental settings are provided in Appendix E. We also share the code in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is provided in the supplementary material. It will be available on Github after the decision.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in the Appendix E. This work is solely concerned with training aspects, with no consideration of testing performance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our figures show the convergence properties of AdamW for specific runs. People often do not plot the behaviors over multiple averaged runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resources information are provided in the Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper follows the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in the conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: They are properly cited based on their licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.