ADOPT: MODIFIED ADAM CAN CONVERGE WITH THE OPTIMAL RATE WITH ANY HYPERPARAMETERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive gradient methods based on exponential moving averages, such as `Adam` and `RMSprop`, are widely used for deep learning. However, it is known that they do not converge unless choosing hyperparameters in a problem-dependent manner. There have been many attempts to fix their convergence (e.g., `AMSGrad`), but they require an impractical assumption that the stochastic gradient is uniformly bounded. In this paper, we propose a new adaptive gradient method named `ADOPT`, which achieves the optimal convergence rate of $\mathcal{O}(1/\sqrt{T})$ with any hyperparameter choice without the bounded stochastic gradient assumption. `ADOPT` addresses the non-convergence issue of `Adam` by removing the current gradient from the second moment estimate and changing the order of the momentum calculation and the scaling operation by the second moment estimate. We also conduct intensive numerical experiments, and verify that our `ADOPT` achieves competitive or even better results compared to `Adam` and its variants across a wide range of tasks, including image classification, generative modeling, natural language processing, and deep reinforcement learning.

1 INTRODUCTION

Stochastic optimization algorithms, such as stochastic gradient descent (SGD), play a central role in deep learning. In particular, adaptive gradient methods based on exponential moving averages are widely used in practice. Despite the empirical success, it is known that some of the most popular algorithms, including Adam (Kingma & Ba, 2014) and RMSprop (Hinton et al., 2012), do not converge in theory. For example, Reddi et al. (2018) show that Adam and RMSprop fail to converge to a correct solution in a simple example where the objective function at time t is given as:

$$f_t(\theta) = \begin{cases} C\theta, & \text{for } t \bmod 3 = 1 \\ -\theta, & \text{otherwise,} \end{cases} \tag{1}$$

where $C > 2$ and $\theta \in [-1, 1]$. In this online optimization setting, Adam and RMSprop with specific hyperparameters converge to a wrong solution (i.e., $\theta = 1$) instead of the true solution (i.e., $\theta = -1$). There have been several attempts to fix the non-convergent behavior of Adam (Reddi et al., 2018; Zou et al., 2019). For example, AMSGrad (Reddi et al., 2018) ensures the convergence for online convex optimization by making slight modifications to the Adam algorithm. Subsequent studies (Chen et al., 2019; Zhou et al., 2018) show that AMSGrad also converges to a stationary point for smooth nonconvex stochastic optimization problems. However, the convergence proofs rely on the assumption that the stochastic gradient is uniformly bounded. This assumption is stronger than the one used for the analysis of vanilla SGD (Ghadimi & Lan, 2013; Bertsekas & Tsitsiklis, 2000; Khaled & Richtárik, 2023), and is often violated in practice. For example, when Gaussian noise is used in the gradient estimation (e.g., the reparameterization trick in variational autoencoders (Kingma & Welling, 2014)), the stochastic gradient is no longer bounded.

Concurrently, Zhou et al. (2019) analyze the cause of non-convergence in Adam and RMSprop in the problem described in Eq. (1) from the perspective of the correlation between the current gradient and the second moment estimate based on the exponential moving average. Specifically, they show that the issue can be resolved by excluding the gradient of the most recent n steps from the calculation of the second moment estimate, where n is a hyperparameter that is equal to or larger than 1. They extend the analysis to the case where momentum is incorporated, as in Adam, and propose AdaShift, which calculates momentum using only the gradient of the most recent n steps to ensure that momentum is uncorrelated with the second moment estimate. However, their theoretical analysis is limited to a single online convex problem described in Eq. (1), and the convergence of AdaShift for general nonconvex problems is unclear. Moreover, this approach involves a trade-off in the choice of n: when n is small, momentum has limited information about past gradients, and when n is large, the second moment estimate has limited information about recent gradients.

More recently, some works have demonstrated that Adam can converge by choosing the hyperparameters in a problem-dependent manner (Shi et al., 2020; Zhang et al., 2022; Wang et al., 2022; Li et al., 2023). However, tuning the hyperparameters for each specific problem is troublesome; hence developing algorithms with the problem-independent convergence guarantee is still important to safely apply adaptive gradient methods to general machine learning problems.

In this paper, we propose an alternative approach to addressing the non-convergence issue of Adam without encountering trade-offs in hyperparameters or relying on strong assumptions such as the bounded stochastic gradient assumption. To derive our algorithm, we first examine the case without momentum, analyzing the convergence bound of RMSprop for general smooth nonconvex optimization problems. Through this analysis, we uncover the fundamental cause of divergence, which stems from the correlation between the second moment estimate and the current gradient. This finding aligns with the results demonstrated by Zhou et al. (2019) for online convex optimization. To resolve the divergence problem, we introduce slight modifications to the RMSprop algorithm that eliminate the correlation. Subsequently, we extend our findings to the case where momentum is incorporated, as in Adam, and discover that the Adam-style momentum also contributes to non-convergence. Although AdaShift addresses this issue by removing past gradients from momentum, it introduces a trade-off as previously described. However, we propose a modification that overcomes this trade-off by changing the order of the momentum calculation and the scaling operation using the second moment estimate. With this small adjustment, we successfully eliminate the non-convergence problem of Adam without relying on a specific hyperparameter choice and the bounded stochastic gradient assumption. We provide theoretical evidence demonstrating that our derived algorithm, named ADaptive gradient method with the OPTimal convergence rate (ADOPT), can achieve convergence with the optimal rate of $\mathcal{O}(1/\sqrt{T})$ for smooth nonconvex optimization.

In our experiments, we begin by assessing the performance of ADOPT in a toy example where Adam typically fails to converge. This toy example is an extension of the one presented in Eq. (1) by Reddi et al. (2018), but we consider a scenario where the bounded stochastic gradient assumption does not hold. Our results demonstrate that ADOPT rapidly converges to the solution, while Adam fails to converge, and AMSGrad exhibits slow convergence due to the violation of the assumption. Next, we conduct an experiment using a simple multi-layer perceptron on the MNIST classification task to evaluate the performance of ADOPT in nonconvex optimization. Our findings indicate that ADOPT outperforms existing adaptive gradient methods, including Adam, AMSGrad, and AdaShift. Finally, we evaluate the performance of ADOPT in various practical applications, such as ImageNet classification using modern neural networks (SwinTransformer), training of deep generative models (NVAE), fine-tuning of language models (LLaMA), and deep reinforcement learning for continuous control. Our empirical results demonstrate that ADOPT achieves competitive or even superior results over existing algorithms (e.g., Adam) in these practical applications.

2 Preliminary

2.1 Problem Definition

We consider the minimization of the objective function $f : \mathbb{R}^D \to \mathbb{R}$ with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^D$. In this context, we focus on first-order stochastic optimization methods, where only the stochastic gradient $\boldsymbol{g}$ is accessible. As the objective f can be nonconvex, the goal is to find a stationary point where $\nabla f(\boldsymbol{\theta}) = 0$ (Blair, 1985; Vavasis, 1995). In order to analyze the convergence behavior of stochastic optimization algorithms, we adopt the following assumptions commonly employed in the literature[1] (Ghadimi & Lan, 2013; Zou et al., 2019; Défossez et al., 2022):

Assumption 1. *The objective function $f(\boldsymbol{\theta})$ is lower-bounded, i.e., $f(\boldsymbol{\theta}) \geq f_{\inf} > -\infty$ for all $\boldsymbol{\theta}$.*

[1] Note that Assumption 4 is often relaxed to an assumption that the variance (instead of the second moment) of the stochastic gradient is uniformly bounded, but we adopt Assumption 4 for the simplicity of our proofs.

Assumption 2. *The stochastic gradient $\boldsymbol{g}_t$ is an unbiased estimator of the objective $f(\boldsymbol{\theta}_{t-1})$, i.e., $\mathbb{E}[\boldsymbol{g}_t] = \nabla f(\boldsymbol{\theta}_{t-1})$ for all $t \geq 1$.*

Assumption 3. *The objective function is L-smooth on Θ, i.e., there exists a constant $L > 0$ such that $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Theta$.*

Assumption 4. *The stochastic gradient has a finite second moment, i.e., there exists a constant $G > 0$ such that $\mathbb{E}[\|\boldsymbol{g}_t\|^2] \leq G^2$.*

In the literature on convergence analysis, it is common to analyze the convergence rate of $\min_t\{\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t))\|^2]\}$, where $\boldsymbol{\theta}_t$ represents the parameter value after t parameter updates.

For the analysis of adaptive gradient methods (e.g., Adam and AMSGrad), many of previous works (Chen et al., 2019; Zhou et al., 2018; Défossez et al., 2022) make an additional assumption that the stochastic gradient $\boldsymbol{g}_t$ is uniformly bounded:

Assumption 5. *The stochastic gradient is uniformly upper-bounded, i.e., there exists a constant $G > 0$ such that $\|\boldsymbol{g}_t\| \leq G$.*

Note that when Assumption 5 holds, Assumption 4 is automatically satisfied. Therefore, Assumption 5 is a stronger assumption compared to Assumption 4. When we omit Assumption 5, it becomes challenging to analyze $\min_t\{\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t))\|^2]\}$ for adaptive gradient methods. As a result, the analysis often considers $\min_t\{\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t))\|^{4/3}]^{3/2}\}$ instead. In this paper, we focus on analyzing $\min_t\{\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t))\|^{4/3}]^{3/2}\}$, because one of our motivations is to address the omission of Assumption 5.

2.2 Review of Stochastic Optimization Algorithms for Nonconvex Objectives

The convergence of the vanilla SGD have been studied extensively in previous works. For smooth nonconvex functions, Ghadimi & Lan (2013) showed that SGD with a constant learning rate converges with an $\mathcal{O}(1/\sqrt{T})$ rate by setting $\alpha_t = \alpha = \Theta(1/\sqrt{T})$, where α_t is a learning rate at the t-th step, and T is a total number of parameter updates. This convergence rate is known to be minimax optimal up to a constant (Drori & Shamir, 2020). For the diminishing learning rate scheme, the convergence bound of $\mathcal{O}(\log T/\sqrt{T})$ is well-known for $\alpha_t = \alpha/\sqrt{t}$ (Ghadimi & Lan, 2013). Recently, Wang et al. (2021) have proved that SGD with $\alpha_t = \alpha/\sqrt{t}$ can also achieve the optimal rate $\mathcal{O}(1/\sqrt{T})$ by additionally assuming that the objective f is upper-bounded.

While the vanilla SGD is still one of the most popular choices for stochastic optimization, adaptive gradient methods are dominantly used especially for deep learning. In adaptive gradient methods, the parameter $\boldsymbol{\theta}$ is updated additionally using the second moment estimate $\boldsymbol{v}_t$ in the following form:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_t \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}}, \tag{2}$$

where ϵ is a small constant, the division between vectors is applied in an element-wise manner, and the addition between a vector $\boldsymbol{a}$ and a scalar b is defined as $(\boldsymbol{a} + b)_i := a_i + b$. In AdaGrad (Duchi et al., 2011), $\boldsymbol{v}_t$ is defined as $\boldsymbol{v}_0 = \boldsymbol{0}$ and $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \boldsymbol{g}_t \odot \boldsymbol{g}_t$. In RMSprop (Hinton et al., 2012), an exponential moving average is substituted for the simple summation, i.e., $\boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t \odot \boldsymbol{g}_t$, where $0 \leq \beta_2 < 1$. Adam (Kingma & Ba, 2014) uses momentum in addition to the second moment estimate to accelerate the convergence as follows:

$$\begin{cases} \boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\,\boldsymbol{g}_t, \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_t \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}}, \end{cases} \tag{3}$$

where $\boldsymbol{m}_0 = \boldsymbol{0}$. Here, we omit the bias correction technique used in the original paper for clarity. Unfortunately, RMSprop and Adam are not guaranteed to converge even in a simple convex optimization problem as demonstrated by Reddi et al. (2018), whereas AdaGrad with a constant learning rate is known to converge with an $\mathcal{O}(\log T/\sqrt{T})$ rate under Assupmtions 1-4 for smooth nonconvex cases (Li & Orabona, 2019; Ward et al., 2020; Zou et al., 2018; Chen et al., 2019; Défossez et al., 2022). Although the convergence of Adam can be assured by choosing the hyperparameters (i.e., β_1 and β_2) in a problem-dependent manner (Shi et al., 2020; Zhang et al., 2022; Wang et al., 2022; Li et al., 2023), it is difficult to know the proper hyperparameters for each problem before training. To fix the non-convergence of Adam without depending on a hyperparameter choice, some researchers have proposed variants of Adam. Reddi et al. (2018) proposed AMSGrad, which substitute $\tilde{\boldsymbol{v}}_t$ for $\boldsymbol{v}$ in Eq. (3), where $\tilde{\boldsymbol{v}}_0 = \mathbf{0}$ and $\tilde{\boldsymbol{v}}_t = \max\left\{\tilde{\boldsymbol{v}}_{t-1}, \boldsymbol{v}_t\right\}$. The idea behind AMSGrad is that the scaling factor $\sqrt{\tilde{\boldsymbol{v}}_t + \epsilon^2}$ should be non-decreasing to ensure the convergence. After Reddi et al. (2018) originally proved the convergence of AMSGrad for online convex optimization, Chen et al. (2019) showed that AMSGrad with $\alpha_t = \alpha/\sqrt{t}$ converges with $\mathcal{O}(\log T/\sqrt{T})$ for nonconvex settings. Zhou et al. (2018) also analyzed the convergence of AMSGrad for nonconvex optimization, and derived the convergence rate of $\mathcal{O}(1/\sqrt{T})$ for a constant learning rate of $\alpha_t = \alpha = \Theta(1/\sqrt{T})$. However, their results depend on Assumption 5, which is often violated in practice. For example, variational autoencoders (Kingma & Welling, 2014) and diffusion models (Ho et al., 2020) are typical examples in which Assumption 5 does not hold because they utilize unbounded Gaussian noise in the gradient estimation. The cause of requirement for Assumption 5 is the max operation in the definition of $\tilde{\boldsymbol{v}}_t$. Since the max operation is convex, $\mathbb{E}[\tilde{\boldsymbol{v}}_t] \leq \max_t\{\mathbb{E}[\boldsymbol{v}_t]\}$ does not hold; hence Assumption 5 is required to upper-bound $\mathbb{E}[\tilde{\boldsymbol{v}}_t]$ in their proofs. Zhou et al. (2019) also tried to fix the non-convergent behavior of Adam. Their proposed AdaShift uses $\boldsymbol{v}_{t-n}$ instead of $\boldsymbol{v}_t$ for the second moment estimate, and calculate the momentum using the latest n gradients as follows:

$$\begin{cases} \boldsymbol{m}_t = \frac{\sum_{k=0}^{n-1} \beta_1^k \boldsymbol{g}_{t-k}}{\sum_{k=0}^{n-1} \beta_1^k}, \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_t \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t-n} + \epsilon^2}}. \end{cases} \tag{4}$$

In the original paper, some additional techniques (e.g., the block-wise adaptive learning rate) are used, but we omit them for clarity here. Though they give theoretical analysis for a single online convex example, any convergence bounds are not provided for nonconvex cases. More detailed discussion on related works is provided in Appendix A.

3 Cause of Non-convergence of Adam and How to Fix It

In this section, to derive an algorithm that can converge with any hyperparameter choice without the bounded stochastic gradient assumption, we analyze the cause of non-convergence of Adam, and discuss how it can be eliminated. To start from a simple case, we first analyze the case without momentum. Subsequently, we extend it to the case with momentum and provide a way to fix the convergence issue of Adam.

3.1 Case without Momentum

We first analyze the convergence of RMSprop, which corresponds to the no-momentum case of Adam when we omit the bias correction. For RMSprop, we derive the following convergence bound.

Theorem 1. *Under Assumptions 1, 2, 3, and 4, the following holds for the RMSprop with a constant learning rate $\alpha_t = \alpha$:*

$$\begin{aligned} &\min_{t=1,\ldots,T} \left\{ \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_{t-1}))\|^{4/3} \right]^{3/2} \right\} \\ &\leq 2\sqrt{\left(1-\beta_2^T\right) G^2 + \epsilon^2} \left(\frac{f\left(\boldsymbol{\theta}_0\right) - f_{\inf}}{\alpha T} + \frac{C}{T} \log\left(1 + \frac{\left(1-\beta_2^T\right) G^2}{\epsilon^2} \right) - C \log \beta_2 \right), \end{aligned} \tag{5}$$

where $C = \frac{\alpha D L}{2(1-\beta_2)} + \frac{2DG}{\sqrt{1-\beta_2}}$.

Sketch of proof. By Assumption 3, the following holds:

$$\mathbb{E}\left[f\left(\boldsymbol{\theta}_t\right)\right] \leq \mathbb{E}\left[f\left(\boldsymbol{\theta}_{t-1}\right)\right] - \alpha \mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top \left(\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}} \right) \right] + \frac{\alpha^2 L}{2} \mathbb{E}\left[\left\| \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}} \right\|^2 \right] \tag{6}$$

Applying Lemmas 7 and 9 in the appendix to this, the following inequality is derived:

$$\begin{aligned} \mathbb{E}\left[f\left(\boldsymbol{\theta}_t\right)\right] &\leq \mathbb{E}\left[f\left(\boldsymbol{\theta}_{t-1}\right)\right] - \frac{\alpha}{2} \mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top \left(\frac{\boldsymbol{g}_t}{\sqrt{\tilde{\boldsymbol{v}}_t + \epsilon^2}} \right) \right] + \left(\frac{\alpha^2 L}{2} + 2\alpha G\sqrt{1-\beta_2} \right) \mathbb{E}\left[\left\| \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}} \right\|^2 \right] \qquad (7) \\ &\leq \mathbb{E}\left[f\left(\boldsymbol{\theta}_{t-1}\right)\right] - \frac{\alpha}{2} \frac{\mathbb{E}\left[\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\|^{4/3} \right]^{3/2}}{\sqrt{\left(1-\beta_2^T\right) G^2 + \epsilon^2}} + \left(\frac{\alpha^2 L}{2} + 2\alpha G\sqrt{1-\beta_2} \right) \mathbb{E}\left[\left\| \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}} \right\|^2 \right], \qquad (8) \end{aligned}$$

where $\tilde{\boldsymbol{v}}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\mathbb{E}[\boldsymbol{g}_t \odot \boldsymbol{g}_t]$. Telescoping this for $t = 1, \ldots, T$ and rearranging the terms, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_{t-1})\|^{4/3}\right]^{3/2} \\ \leq 2\sqrt{\left(1-\beta_2^T\right)G^2 + \epsilon^2}\left(\frac{f(\boldsymbol{\theta}_0) - f_{\text{inf}}}{\alpha} + C\log\left(1 + \frac{\left(1-\beta_2^T\right)G^2}{\epsilon^2}\right) - CT\log\beta_2\right), \tag{9}$$

where the last inequality holds due to Assumption 1 and Lemma 8. Therefore, the bound in Eq. (5) is derived using the following fact:

$$\min_{t=1,\ldots,T}\left\{\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_{t-1}))\|^{4/3}\right]^{3/2}\right\} \leq \frac{\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_{t-1})\|^{4/3}\right]^{3/2}}{T}. \tag{10}$$

□

A detailed proof is provided in the appendix. When the learning rate α is chosen so that $\alpha = \Theta(1/\sqrt{T})$, the first and second terms on the right hand side of Eq. (5) converge with $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ rates, respectively. However, the last term includes a constant factor in terms of T, which represents the non-convergent behavior of RMSprop in the smooth nonconvex setting. More precisely, RMSprop is guaranteed to converge only to a bounded region around a stationary point, and the size of the bounded region depends on the hyperparameter β_2 and the problem-dependent factors D, G, and L. Therefore, we need to choose β_2 dependently on each problem to make the bounded region adequately small. Basically, the size of the bounded region can be made small by setting β_2 to a value close to 1, but how close to 1 it should be relies on the problem-dependent factors, which cannot be observed in advance. This result is consistent with recent results of convergence analyses of Adam and RMSprop (Shi et al., 2020; Zhang et al., 2022).

As can be seen from Eqs. (6) and (7), the constant term in Eq. (5) is derived from the second term of Eq. (6). Because $\boldsymbol{g}_t$ and $\boldsymbol{v}_t$ are not statistically independent, this term is first decomposed into the second and third terms of Eq. (7) by using Lemma 7. After the decomposition, $\boldsymbol{g}_t$ and $\tilde{\boldsymbol{v}}_t$ is now conditionally independent given $\boldsymbol{g}_0, \ldots, \boldsymbol{g}_{t-1}$, so Eq. (8) is derived using the following fact in Lemma 9:

$$\mathbb{E}\left[\frac{\boldsymbol{g}_t}{\sqrt{\tilde{\boldsymbol{v}}_t + \epsilon^2}}\right] = \mathbb{E}\left[\frac{\nabla f(\boldsymbol{\theta}_{t-1})}{\sqrt{\tilde{\boldsymbol{v}}_t + \epsilon^2}}\right]. \tag{11}$$

In other words, if the second moment estimate is designed to be conditionally independent to $\boldsymbol{g}_t$, the constant term in the convergence bound will be removed, because the second term of Eq. (6) can be directly lower-bounded by a quantity propotional to $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t))\|^{4/3}]^{3/2}$ as in Lemma 9. A simple way to achieve the conditional independence is to substitute $\boldsymbol{v}_{t-1}$ for $\boldsymbol{v}_t$ as a second moment estimate, because $\boldsymbol{v}_{t-1}$ does not have information about $\boldsymbol{g}_t$. This solution is similar to AdaShift, in which $\boldsymbol{v}_{t-n}$ is substituted for $\boldsymbol{v}_t$ as described in Eq. (4). In fact, the modified version of RMSprop is identical to AdaShift with $n = 1$ and $\beta_1 = 0$ except for the additional techniques (e.g., the block-wise adaptive learning rate).

3.2 Case with Momentum

As we have described, RMSprop can be modified to be convergent by removing the current gradient $\boldsymbol{g}_t$ from the second moment estimate $\boldsymbol{v}_t$. However, when we combine adaptive gradient methods with momentum like Adam, the convergence analysis becomes more complicated. Unfortunately, when Adam-style momentum in Eq. (3) is applied, the algorithm does not converge in general even when using $\boldsymbol{v}_{t-1}$ as a second moment estimate instead of $\boldsymbol{v}_t$. This is because the momentum $\boldsymbol{m}_t$ contains all history of the past gradients $\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t$; hence the second moment estimate always correlates with $\boldsymbol{m}_t$. AdaShift prevents this problem by calculating the momentum $\boldsymbol{m}_t$ only using the latest n gradients as described in Eq. (4). In that case, the momentum $\boldsymbol{m}_t$ and the second moment estimate $\boldsymbol{v}_{t-n}$ are conditionally independent, so the convergence can be retained. However, this approach has a trade-off in the choice of n. When n is small, $\boldsymbol{m}_t$ has little information about the past gradients; when n is large, $\boldsymbol{v}_{t-n}$ only has access to the gradient information in the distant past.

Algorithm 1 ADOPT algorithm

```
Require: Learning rate $\{\alpha_t\}$, initial parameter $\boldsymbol{\theta}_0$
Require: Exponential decay rate $0 \leq \beta_1 < 1, 0 \leq \beta_2 \leq 1$, small constant $\epsilon \geq 0$
  $\boldsymbol{m}_0 \leftarrow \mathbf{0}, \boldsymbol{v}_0 \leftarrow \mathbf{1}$
  for $t = 1$ to $T$ do
    $\boldsymbol{m}_t \leftarrow \beta_1 \cdot \boldsymbol{m}_{t-1} + (1 - \beta_1) \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}}$
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha_t \boldsymbol{m}_t$
    $\boldsymbol{v}_t \leftarrow \beta_2 \cdot \boldsymbol{v}_{t-1} + (1 - \beta_2) \, \boldsymbol{g}_t \odot \boldsymbol{g}_t$
  end for
  return $\{\boldsymbol{\theta}_t\}_{t=1}^T$
```

To remove this trade-off, instead of truncating the momentum to the latest n steps, we propose to use momentum of the following form:

$$\begin{cases} \boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1) \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}}, \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_t \boldsymbol{m}_t, \end{cases} \tag{12}$$

The main difference to the Adam-style momentum in Eq. (3) is the order of momentum calculation and the scaling operation by $\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}$. In Eq. (3), the scaling operation is performed after the momentum calculation, whereas in Eq. (12), the scaling operation is first applied to the current gradient $\boldsymbol{g}_t$ in advance to the momentum calculation. In this case, the second moment estimate $\boldsymbol{v}_{t-1}$ is only used to scale the current gradient $\boldsymbol{g}_t$, so the convergence can be guaranteed. A more detailed convergence analysis is provided in Section 4.

4 ADOPT: Adaptive Gradient Method with the Optimal Convergence Rate

Based on the analysis in the previous section, we propose a new adaptive gradient method named ADOPT (ADaptive gradient method with the OPTimal convergence rate). The entire procedure is summarized in Algorithm 4. In ADOPT, to ensure the convergence, we use $\boldsymbol{v}_{t-1}$ as a second moment estimate instead of $\boldsymbol{v}_t$, and the scaling operation by $\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}$ is applied not to the momentum $\boldsymbol{m}_t$ but to the current gradient $\boldsymbol{g}_t$. To prevent the initial scaling factor $\sqrt{\boldsymbol{v}_0 + \epsilon^2}$ from being too small, we initialize $\boldsymbol{v}_0$ with 1 instead of 0. By this modification, ADOPT can converge with the optimal rate $\mathcal{O}(1/\sqrt{T})$ for the smooth nonconvex optimization as follows:

Theorem 2. *Under Assumptions 1, 2, 3, and 4, the following holds for the ADOPT algorithm with a constant learning rate $\alpha_t = \alpha$:*

$$\min_{t=1,\ldots,T} \left\{ \mathbb{E} \left[\|\nabla f(\boldsymbol{\theta}_{t-1}))\|^{4/3} \right]^{3/2} \right\} \leq C_1(T) \left(\frac{f(\boldsymbol{\theta}_0) - f_{\inf}}{\alpha T} + \alpha C_2 \left(1 - \frac{1}{T \log \beta_2} \log \left(\frac{\beta_2^T + \epsilon^2}{1 + \epsilon^2} \right) \right) \right), \tag{13}$$

where $C_1(T) = \sqrt{\max\left\{G^2 + (1 - G^2)\beta_2^T, 1\right\} + \epsilon^2}$, $C_2 = \frac{(1+\beta_1)G^2 L}{2(1-\beta_1)\epsilon^2}$.

The detailed proof and related lemmas are provided in the appendix. When we choose the learning rate so that $\alpha = \Theta(1/\sqrt{T})$, the right hind side of Eq. (13) converges with an $\mathcal{O}(1/\sqrt{T})$ rate. We also provide the convergence bound for the case of diminishing learning rate (i.e., $\alpha_t = \alpha/\sqrt{t}$) in the appendix, which is closer to practical situations. In that case, ADOPT also converges with the optimal rate of $\mathcal{O}(1/\sqrt{T})$.

5 Experiments

In the experiments, we first validate our ADOPT algorithm using a simple toy example in which Adam is known to fail to converge, and confirm our theoretical findings through numerical simulation. Secondly, we run an experiment of training a simple multi-layer perceptron (MLP) for the MNIST dataset to verify the effectiveness of our ADOPT for nonconvex optimization problems.

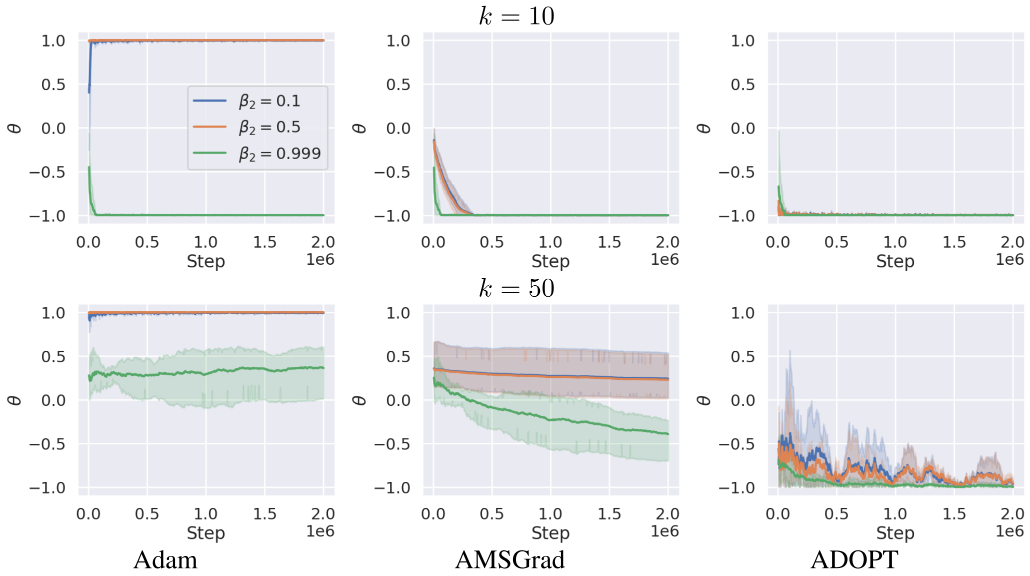

Figure 1: Performance comparison between Adam, AMSGrad and ADOPT in a simple univariate convex optimization problem. The plots show transitions of the parameter value, which should converge to the solution $\theta = -1$.

Finally, we evaluate our ADOPT in a wide range of practical applications, including image classification, natural language processing (NLP) tasks, generative modeling, and deep reinforcement learning. Detailed experimental settings are described in the appendix.

Toy problem: We consider a convex optimization problem with an objective $f(\theta) = \theta$ for $\theta \in [-1, 1]$. It is obvious that a solution for the problem is $\theta = -1$. Through the optimization, we only have access to the stochastic objective f_t as follows:

$$f_t(\theta) = \begin{cases} k^2\theta, & \text{with probability } 1/k \\ -k\theta, & \text{with probability } 1 - 1/k \end{cases}, \tag{14}$$

where $k \geq 1$. Because $\mathbb{E}[f_t(\theta)] = f(\theta)$ holds, the stochastic gradient $g_t = \nabla f_t(\theta)$ is an unbiased estimator of the true gradient ∇f regardless of the choice of k, satisfying Assumption 2. This problem is equivalent, except for scaling, to the stochastic optimization version of Eq. (1) provided by Reddi et al. (2018) as a case where Adam (and RMSprop) with specific hyperparameters fail to converge. The constant k controls the magnitude of gradient noise. When $k = 1$, it corresponds to the noiseless case where $f_t = f$ with probability 1. As k gets large, stochastic gradient becomes noisy, making G in Assumptions 4 or 5 large. Therefore, the optimization will be more difficult when k becomes larger. In the experiment, we set $k = 10$ or 50, and compare the robustness of Adam, AMSGrad, and ADOPT for various hyperparameter settings by changing β_2 from $0.1 \sim 0.999$. We set $\beta_1 = 0.9$ for all the algorithms, which is a common choice in practice. We set the learning rate to $\alpha_t = 0.01/\sqrt{1 + 0.01t}$. The parameter θ is initialized to 0 for all cases.

The result is shown in Figure 5. It can be seen that, when $k = 10$, Adam fails to converge except for $\beta_2 = 0.999$ while AMSGrad and ADOPT rapidly converge to the correct solution, i.e., $\theta = -1$. In a more extreme case where $k = 50$, Adam fails to converge even with β_2 very close to 1. This aligns with Theorem 1, since, when the gradient noise is large (i.e., G is large), the bounded region of the convergence bound also gets large, leading to divergence of Adam. Moreover, when $k = 50$, it is observed that the convergence of AMSGrad also becomes much slower than ADOPT. In fact, this phenomenon is also consistent with theory. In this problem setting, the second moment $\mathbb{E}[g_t^2]$ is $\mathcal{O}(k^3)$, while the squared norm of the stochastic gradient g_t^2 is $\mathcal{O}(k^4)$. Since the convergence bound of AMSGrad depends on the uniform bound of the stochastic gradient in Assumption 5, instead of the second moment in Assumption 4, its convergence also deteriorates with the order of g_t^2. Compared to AMSGrad, ADOPT only depends on the second moment bound for its convergence, so it converges much faster than AMSGrad even in such an extreme setting, although the convergence speed depends on the choice of hyperparameters.

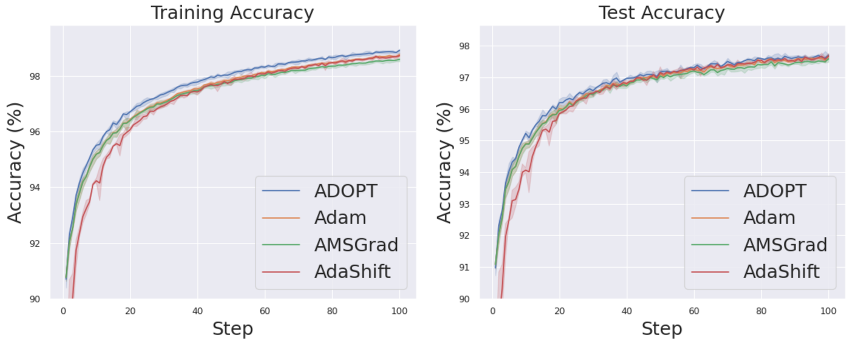

Figure 3: Accuracy for training data (left) and test data(right) in MNIST classification. The error bars show the 95% confidence intervals of three trials.

We also perform ablation study on how the two algorithmic changes from Adam to ADOPT affect the convergence. The differences between Adam and ADOPT are (1) decorrelation between the second moment estimate and the current gradient, and (2) change of order of momentum calculation and scaling operation by the second moment estimate. In this experiment, we remove each algorithmic change from ADOPT, and compare the result in the toy example. We set $k = 50$, and $(\beta_1, \beta_2) = (0.9, 0.999)$, since it is a common hyperparameter choice. The result is shown in Figure 2. It can be observed that ADOPT fails to converge with the exception of either algorithmic change. Therefore, applying both changes is essential to overcome the non-convergent issue of Adam, which also aligns with theory.

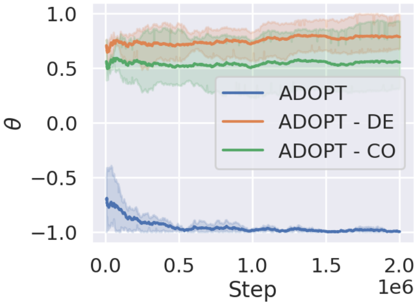

Figure 2: Ablation study of algorithmic changes between Adam and ADOPT. ”DE” and CO denote ”decorrelation” and ”change of order”, respectively.

These results correspond to the theoretical findings, showing the superiority of ADOPT to Adam and AMSGrad in terms of the convergence speed and its robustness to hyperparameter choices.

MNIST classification: To investigate the performance on nonconvex optimization, we compare ADOPT with Adam, AMSGrad and AdaShift, on the MNIST classification using an MLP with a single hidden layer. The number of hidden units is set to 784. We set the learning rate to $\alpha_t = \alpha / \sqrt{t}$, and α is tuned in the range of $\{1, 10^{-1}, 10^{-2}, 10^{-3}\}$. We apply weight decay of 1×10^{-4} to prevent over-fitting, and run 10K iterations of parameter updates. Figure 3 shows the learning curves of training and test accuracy. We observe our ADOPT performs slightly better than the others in terms of the convergence speed and the final performance. Thanks to the way of the momentum calculation in Eq. (12), ADOPT works better than AdaShift especially in the early phase of training.

ImageNet classification: We perform ImageNet classification using SwinTransformer (Liu et al., 2021) to confirm that our ADOPT works well for modern vision Transformers. We follow the official training recipe of Swin Transformer-tiny provided by Torchvision[2], and fix the training settings except for the optimizer choice. We use AdamW (Loshchilov & Hutter, 2019) as a baseline because it is set as the default official optimizer. We also compare with AMSGrad as another way to fix the divergence issue of Adam. Since AdamW uses decoupled weight decay, we also apply it to the other optimizers for fair comparison. We report the top-1 accuracy at $\frac{T}{3}$, $\frac{2}{3}T$ and T epochs in Tables 1, where T is the total number of training epochs. We observe that ADOPT outperforms AdamW and AMSGrad throughout the training in terms of the test accuracy, demonstrating the effectiveness of ADOPT for this setting.

Generative modeling: We train NVAE (Vahdat & Kautz, 2020) for MNIST using our ADOPT. In the official implementation of NVAE, Adamax (Kingma & Ba, 2014), an infinite-norm variant of Adam, is used as an optimizer, so we use Adamax as a baseline method. We use the exactly the same

[2] https://github.com/pytorch/vision/tree/main/references/classification

Table 1: Top-1 accuracy (%) of SwinTransformer on ImageNet.

Epoch	200	300
AdamW	79.29 ± 0.05	81.26 ± 0.04
AMSGrad	78.91 ± 0.03	81.17 ± 0.03
ADOPT	$\mathbf{79.62} \pm 0.03$	$\mathbf{81.50} \pm 0.04$

Table 2: Negative log-likelihood of NVAEs for MNIST.

Epoch	200	300
Adamax	80.19 ± 0.08	79.41 ± 0.07
ADOPT	$\mathbf{79.02} \pm 0.10$	$\mathbf{78.88} \pm 0.09$

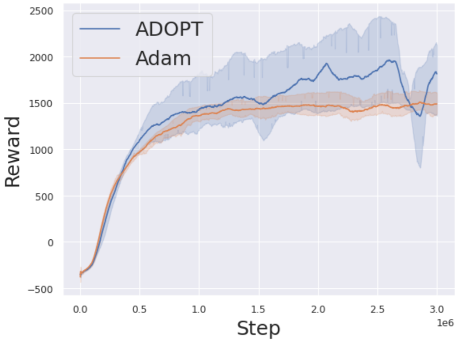

Figure 4: Performance comparison between Adam and ADOPT in reinforcement learning.

setting of the official implementation except that the learning rate for ADOPT is set to 2×10^{-4} since the default value 0.01 is too large for ADOPT. We report the negative log-likelihood for test data on Table 2. It is observed that the model trained with ADOPT shows the better likelihood.

Finetuning of large language models: We finetune the pretrained LLaMA-7B on 52K instruction-following data provided by Stanford Alpaca and compare the performance between the default optimizer (Adam,) and our ADOPT under the exactly same experimental setting. For evaluation, we use Multi-task Language Understanding (MMLU) Benchmark (Hendrycks et al., 2021), which is widely used to assess the performance of large language models. The MMLU score for LLaMA-7B without finetuning is 35.1. After fine-tuned via instruction-following using the baseline implementation with Adam, the score improves to 41.2. When we substitute ADOPT for Adam, the score even improves to 42.13. The detailed score comparison for each task is summarized in Figure 5 in the appendix.

Deep reinforcement learning: Lastly, we train reinforcement learning (RL) agents using the proximal policy optimization (PPO) with ADOPT for the optimizer. As a benchmark, we use a continuous control tasks of HalfCheetah on MuJoCo simulator. For comparison to ADOPT, Adam is used as a baseline optimizer. We follow the hyperparameter settings recommended by Stable-Baselines3 (Raffin et al., 2021), and just change the choice of an optimizer. The result is shown in Figure 4. It can be observed that ADOPT shows competitive or even better performance than Adam.

6 Limitations

One of the limitations of our analysis is that it still relies on the assumption that the second moment of stochastic gradient is uniformly bounded (i.e., Assumption 4). Although this assumption is weaker than the bounded stochastic gradient assumption (i.e., Assumption 5), it would be more desirable to relax it to an assumption that the variance of the stochastic gradient is uniformly bounded, which is often adopted in the analysis of the vanilla SGD (Ghadimi & Lan, 2013). Extending our result to weaker assumptions is an important direction of future work.

7 Conclusion

In this paper, we demystified the fundamental cause of divergence of adaptive gradient methods based on the exponential moving average, such as Adam and RMSprop, in general smooth nonconvex optimization problems, and demonstrate a way to fix the issue, proposing a new optimizer named ADOPT. Not only does ADOPT converge with the optimal rate without depending on a hyperparameter choice in theory, but ADOPT demonstrates competitive or even better results in the pracital applications, including ImageNet classification, generative modeling, finetuning of language models, and deep reinforcement learning. We expect that this work will serve as a bridge between theory and practice in the research of adaptive gradient methods. Since our ADOPT can be safely applied to many machine learning problems without careful tuning of hyperparameters, it can be expected to improve the training stability and the model performance in practice by substituting it for the existing adaptive gradien methods (e.g., Adam and RMSprop).

References

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Charles Blair. Problem complexity and method efficiency in optimization (as nemirovsky and db yudin). *Siam Review*, 27(2):264, 1985.

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*, 2019. URL https://openreview.net/forum?id=H1x-x309tm.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPQhzTSWA7.

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2658–2667. PMLR, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pp. 2624–2633. PMLR, 2019.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Lecture 6e rmsprop: Divide the gradient by a running average of its recent magnitude, 2012. URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AU4qHN2VkS. Survey Certification.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Haochuan Li, Ali Jadbabaie, and Alexander Rakhlin. Convergence of adam under relaxed assumptions. *arXiv preprint arXiv:2304.13972*, 2023.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pp. 983–992. PMLR, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *The Journal of Machine Learning Research*, 22(1):12348–12355, 2021.

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

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. Rmsprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Stephen A Vavasis. Complexity issues in global optimization: a survey. *Handbook of global optimization*, pp. 27–41, 1995.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

Xiaoyu Wang, Sindri Magnússon, and Mikael Johansson. On the convergence of step decay step-size for stochastic optimization. *Advances in Neural Information Processing Systems*, 34:14226–14238, 2021.

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

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=l5UNyaHqFdO.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgTkhRcKQ.

Fangyu Zou, Li Shen, Zequn Jie, Ju Sun, and Wei Liu. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2018.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11127–11135, 2019.

	Algorithm	Problem	Smoothness	Gradient Growth
Zhang et al. (2022)	Adam	Finite sum	L-Lipschitz $\boldsymbol{g}$	$\mathbb{E}[\|g\|^2] \leq G_0^2 + G_1^2\|\nabla f\|^2$
Wang et al. (2022)	Adam	Finite sum	(L_0, L_1)-Lipschitz $\boldsymbol{g}$	$\mathbb{E}[\|g\|^2] \leq G_0^2 + G_1^2\|\nabla f\|^2$
Li et al. (2023)	Adam	General	(L_0, L_ρ)-Lipschitz ∇f	$\|\boldsymbol{g} - \nabla f\| \leq \sigma$
Défossez et al. (2022)	Adam	General	L-Lipschitz ∇f	$\mathbb{E}[\|\boldsymbol{g}\|^2] \leq G^2$
Chen et al. (2019)	AMSGrad	General	L-Lipschitz ∇f	$\|\boldsymbol{g}\| \leq G$
Zhou et al. (2018)	AMSGrad	General	L-Lipschitz ∇f	$\|\boldsymbol{g}\| \leq G$
Ours	ADOPT	General	L-Lipschitz ∇f	$\mathbb{E}[\|\boldsymbol{g}\|^2] \leq G^2$

Table 3: Comparison of the problem settings between our analysis and other existing works.

	Constraints	Convergence
Zhang et al. (2022)	$\beta_1 < \sqrt{\beta_2}, \beta_2 \geq \gamma(n)$	$\mathcal{O}(\log T/\sqrt{T}) + \mathcal{O}(G_0)$
Wang et al. (2022)	$\beta_1 < \sqrt{\beta_2}, \delta(\beta_2) = \mathcal{O}(1/G_1)$	$\mathcal{O}(\log T/\sqrt{T}) + \mathcal{O}(G_0)$
Li et al. (2023)	$\beta_1 < \sqrt{\beta_2}, \beta_1 \leq c(L_0, L_\rho, G)$	$\mathcal{O}(1/\sqrt{T})$
Défossez et al. (2022)	$\beta_1 < \sqrt{\beta_2}, 1 - \beta_2 = \Theta(1/T)$	$\mathcal{O}(\log T/\sqrt{T})$
Chen et al. (2019)	$\beta_1 < \sqrt{\beta_2}$	$\mathcal{O}(\log T/\sqrt{T})$
Zhou et al. (2018)	$\beta_1 < \sqrt{\beta_2}$	$\mathcal{O}(1/\sqrt{T})$
Ours	-	$\mathcal{O}(1/\sqrt{T})$

Table 4: Comparison of the convergence rate and imposed constraints on the hyperparameters between our analysis and other existing works. Please refer to the original papers for the definitions of γ and c.

A Detailed Relationships to Existing Analyses

In this section, we discuss the relationships between our analysis and existing ones on the convergence of Adam-like optimizers in smooth nonconvex optimization problems. Tables 7 and 7 are a summary of comparisons between them in terms of their problem settings and derived convergence rates.

Zhang et al. (2022) focus on convergence of Adam in the finite sum problem, where the objective has a following form:

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_i(\boldsymbol{\theta}). \tag{15}$$

f_i is, for example, a loss function for i-th training sample. Although many deep learning problems can be formulated as a finite sum problem, training of the variational autoencoders (VAEs) or diffusion models is out of the finite-sum problem, since their objective is formulated as an infinite sum (i.e., an expectation over continuous variables). Moreover, they assume the stochastic gradient $\boldsymbol{g}$ is L-Lipschitz, whereas we only assume true gradient ∇f is L-Lipschitz. They also assume a growth condition as follows:

$$\mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right] \leq G_0^2 + G_1^2 \|\nabla f(\boldsymbol{\theta}_{t-1})\|^2. \tag{16}$$

This growth condition is weaker than our Assumption 4. Assumption 4 is a special case of the growth condition where $G_1 = 0$. Their derived convergence rate has a constant factor of $\mathcal{O}(G_0)$; hence the strong growth condition (i.e., $G_0 = 0$) is required to assure convergence. Moreover, to achieve the convergence rate, one needs to choose sufficiently large β_2, which has to be tuned in a problem-dependent manner.

Wang et al. (2022) also focus on convergence of Adam in the finite sum problem, but they relax the L-Lipschitz condition on $\boldsymbol{g}$ to the (L_0, L_1)-Lipschitz condition. They also assume the growth condition in Eq. (16), and their convergence rate has the same order with Zhang et al. (2022), so it still requires the strong growth condition (i.e., $G_0 = 0$) to assure convergence. The condition of β_2 is also similar to Zhang et al. (2022).

Li et al. (2023) consider Adam's convergence on general smooth nonconvex problems. Similar to Wang et al. (2022), they use (L_0, L_ρ)-Lipschitz condition on the true gradient ∇f. They also assume that the gradient noise is almost surely bounded:

$$\|\boldsymbol{g} - \nabla f\| \leq \sigma \tag{17}$$

The relationship between this assumption and our Assumption 4 is a little complicated. Assumption 4 is equivalent to a combination of the following two Assumptions:

Assumption 6. *The true gradient is uniformly bounded, i.e., there exist constants G and σ such that $\|\nabla f(\boldsymbol{\theta})\|^2 \leq G^2 - \sigma^2$ and $G > \sigma$.*

Assumption 7. *Variance of the stochastic gradient is uniformly bounded, i.e., there exist a constant σ such that $\mathbb{E}\left[\|\boldsymbol{g}_t - \nabla f(\boldsymbol{\theta}_{t-1})\|^2\right] \leq \sigma^2$.*

The bounded noise assumption of Eq. (17) is strictly stronger than Assumption 7, but they do not assume the bounded true gradient (i.e., Assumption 6). The bounded noise assumption is often violated in practice (e.g., training of VAEs), because the gradient is often estimated using unbounded noise (i.e., Gaussian noise). Their convergence rate $\mathcal{O}(1/\sqrt{T})$ is better than Zhang et al. (2022) and Wang et al. (2022), while it still requires constraints on the hyperparameters, which have to be chosen in a problem-dependent manner.

Défossez et al. (2022) analyzes the convergence of Adam under exactly the same assumptions with ours, and they derive the $\mathcal{O}(\log T/\sqrt{T})$ rate, which is worse than our ADOPT's convergence rate. Moereover, to assure the convergence, β_2 has to be chosen dependently on the total number of iterations T.

Chen et al. (2019) and **Zhou et al. (2018)** analyze the convergence of AMSGrad for general smooth nonconvex problems, and derive the convergence rate of $\mathcal{O}(\log T/\sqrt{T})$ and $\mathcal{O}(1/\sqrt{T})$, respectively. However, to guarantee the convergence, the stochastic gradient $\boldsymbol{g}$ has to be bounded almost surely (Assumption 5), which is often violated in practice. In addition, the hyperparameter β_1 and β_2 should be chosen satisfying $\beta_1 < \sqrt{\beta_2}$. This constraint is relatively minor compared to the constraint imposed in the analyses of Adam, since it can be satisfied in a problem-independent manner.

B With-Replacement vs. Without-Replacement

In the optimization of finite-sum problems, practitioners often use *without-replacement sampling*, which is also known as *random shuffling*, to obtain stochastic gradient. In this case, the stochastic gradient has a small bias due to the lack of replacement, so Assumption 2 is violated. However, the vanilla SGD is known to converge with the without-replacement strategy (Haochen & Sra, 2019), and some of the analyses of Adam also adopt without-replacement sampling (Zhang et al., 2022; Wang et al., 2022).

Unfortunately, we find that our ADOPT has a counter example, in which ADOPT fails to converge when using without-replacement sampling. For example, when we consider minimizing $f(\theta) = \sum_{i=1}^{3} f_i(\theta)$, where $\theta \in [-1, 1]$, $f_1(\theta) = 1.9\theta$ and $f_2(\theta) = f_2(\theta) = -\theta$, it can be easily observed that ADOPT with $\beta_1 = \beta_2 = 0$ fails to converge to the correct solution, i.e., $\theta = 1$.

This non-convergence can be easily avoided by using the with-replacement strategy. Moreover, the difference between with- and without-replacement sampling becomes negligible when n in the finite-sum $\sum_{i=1}^{n} f_i$ is large enough; hence it does not affect the practical performance very much. In fact, our experiments except for the toy example are performed using without-replacement sampling, but divergent behaviors are not observed. If one applies ADOPT to problems where the difference seems severe (e.g., when training with a small dataset), we recommend to use with-replacement sampling instead of random shuffling for stable training. When one uses PyTorch (Paszke et al., 2019) for the implementation, for example, with-replacement sampling can be easily applied by specifying `replacemnet=True` for `torch.utils.data.RandomSampler`, and feeding it to the `sampler` argument of `torch.utils.data.DataLoader`.

C On the Bias Correction Technique

A bias correction technique is originally proposed in Adam (Kingma & Ba, 2014), in which the momentum and the second moment estimate are scaled as follows:

$$\tilde{\boldsymbol{m}}_t = \boldsymbol{m}_t / \left(1 - \beta_1^t\right) \tag{18}$$

$$\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_t / \left(1 - \beta_2^t\right) \tag{19}$$

This technique prevents $\boldsymbol{m}_t$ and $\boldsymbol{v}_t$ from being biased to the initial value (i.e., $\mathbf{0}$ for Adam), and contributes to stable training of Adam. Basically, this technique can be directly applicable to our ADOPT, but we empirically find that ADOPT works adequately well even when the bias correction is not applied. Our hypothesis about the reason of it is that $\boldsymbol{v}_t$ is initialized to $\mathbf{1}$ for ADOPT to prevent it from being too small at the first parameter update, so the effect of bias correction is limited compared to Adam. Moreover, as for the momentum $\boldsymbol{m}_t$, initialization with $\mathbf{0}$ works like a warm-up of the learning rate, so it may have a positive effect for stable training especially in the early phase of optimization. Although there might be a better way to introduce a technique like the bias correction for ADOPT, we leave it for future work.

D Recommendation of Hyperparameter Settings for ADOPT

We experimentally find that our ADOPT works similarly to Adam when the same hyperparameters are used, but ϵ should be set to a little larger value (e.g., 1×10^{-6}) for ADOPT compared to Adam, in which ϵ is set to 1×10^{-8} by default. Our recommendation of the hyperparameter settings for ADOPT is provided in Table 5.

β_1	0.9
β_2	0.999
ϵ	1×10^{-6}

Table 5: Recommended hyperparameters for the ADOPT algorithm

E Theorems

Theorem 3. *Under Assumptions 1, 2, 3, and 4, if the objective f is upper-bounded by $f_{\sup}$, the following holds for the ADOPT algorithm with a learning rate $\alpha_t = \alpha / \sqrt{t}$:*

$$\begin{aligned}
&\min_{t=1,\ldots,T} \left\{ \mathbb{E}\left[\left\| \nabla f\left(\boldsymbol{\theta}_t\right) \right\|^{4/3} \right]^{3/2} \right\} \\
&\leq \frac{3\sqrt{\max\left\{G^2, 1\right\} + \epsilon^2}}{2\left(\left(T+1\right)^{3/2} - 1\right)} \left(\frac{f_{\sup} - f_{\inf}}{\alpha} \left(T+1\right) + \left(\frac{\sqrt{2}\alpha\beta_1 G^2 L}{\epsilon^2 \left(1-\beta_1\right)} + \frac{\alpha G^2 L}{2\epsilon^2} \right) T \right) \\
&\quad + \frac{3\sqrt{\max\left\{G^2, 1\right\} + \epsilon^2}}{2\left(\left(T+1\right)^{3/2} - 1\right)} \left(\frac{2\sqrt{2}\beta_1 G^2}{\epsilon\left(1-\beta_1\right)} \left(\sqrt{T+1} - 1\right) + \frac{\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} \frac{T}{T+1} \right) \\
&\quad + \frac{3\sqrt{\max\left\{G^2, 1\right\} + \epsilon^2}}{2\left(\left(T+1\right)^{3/2} - 1\right)} \left(\frac{2\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha^2 \beta_1 G^2 L}{\sqrt{2}\epsilon^2 \left(1-\beta_1\right)} \right) \log\left(T+1\right).
\end{aligned} \tag{20}$$

F Proofs

Proof of Theorems 2 and 3. We define $\boldsymbol{\phi}_t$ for $t \geq 1$ as follows:

$$\boldsymbol{\phi}_t = \frac{1}{1-\beta_1}\boldsymbol{\theta}_t - \frac{\beta_1}{1-\beta_1}\boldsymbol{\theta}_{t-1}. \tag{21}$$

We also define $\boldsymbol{\phi}_0 = \boldsymbol{\theta}_0$. By Assumption 3, the following holds for $t \geq 1$:

$$\begin{aligned}
f\left(\boldsymbol{\phi}_t\right) &\leq f\left(\boldsymbol{\phi}_{t-1}\right) + \nabla f\left(\boldsymbol{\phi}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right) + \frac{L}{2}\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2 \qquad (22)\\
&= f\left(\boldsymbol{\phi}_{t-1}\right) + \nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\\
&\quad + \left(\nabla f\left(\boldsymbol{\phi}_{t-1}\right) - \nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right) + \frac{L}{2}\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2 \qquad (23)\\
&\leq f\left(\boldsymbol{\phi}_{t-1}\right) + \nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\\
&\quad + \left\|\nabla f\left(\boldsymbol{\phi}_{t-1}\right) - \nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\| + \frac{L}{2}\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2 \qquad (24)\\
&\leq f\left(\boldsymbol{\phi}_{t-1}\right) + \nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\\
&\quad + L\left\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}_{t-1}\right\|\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\| + \frac{L}{2}\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2, \qquad (25)
\end{aligned}$$

where the second inequality holds due to the Cauchy-Schwarz inequality, and the last inequality holds due to Assumption 3.

By taking the expectation, the following holds:

$$\begin{aligned}
\mathbb{E}\left[f\left(\boldsymbol{\phi}_t\right)\right] &\leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] + \mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\right]\\
&\quad + L\mathbb{E}\left[\left\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}_{t-1}\right\|\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|\right] + \frac{L}{2}\mathbb{E}\left[\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2\right] \qquad (26)\\
&\leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] + \frac{\left(\alpha_{t-1} - \alpha_t\right)\beta_1\left(1 - \beta_1^{t-1}\right)G^2}{\left(1 - \beta_1\right)\sqrt{\beta_2^{t-2} + \epsilon^2}}\\
&\quad - \alpha_t\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1 - G^2)\beta_2^T, 1\right\} + \epsilon^2}} \qquad (27)\\
&\quad + \frac{\alpha_{t-1}\left(\alpha_{t-1} - \alpha_t\right)\beta_1^2\left(1 - \beta_1^{t-1}\right)G^2L}{\left(\beta_2^{t-2} + \epsilon^2\right)\left(1 - \beta_1\right)^2} + \frac{\alpha_t\alpha_{t-1}\beta_1\sqrt{1 - \beta_1^{t-1}}G^2L}{\left(1 - \beta_1\right)\sqrt{\beta_2^{t-1} + \epsilon^2}\sqrt{\beta_2^{t-2} + \epsilon^2}}\\
&\quad + \frac{\left(\alpha_{t-1} - \alpha_t\right)^2\beta_1^2\left(1 - \beta_1^{t-1}\right)G^2L}{2\left(1 - \beta_1\right)^2\left(\beta_2^{t-2} + \epsilon^2\right)} + \frac{\alpha_t^2G^2L}{2\left(\beta_2^{t-1} + \epsilon^2\right)}\\
&\quad + \frac{\alpha_t\left(\alpha_{t-1} - \alpha_t\right)\beta_1\sqrt{1 - \beta_1^{t-1}}G^2L}{2\left(1 - \beta_1\right)\sqrt{\beta_2^{t-1} + \epsilon^2}\sqrt{\beta_2^{t-2} + \epsilon^2}}. \qquad (28)
\end{aligned}$$

When $\alpha_t = \alpha$, the following holds:

$$\begin{aligned}
\mathbb{E}\left[f\left(\boldsymbol{\phi}_t\right)\right] &\leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] - \alpha\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1 - G^2)\beta_2^T, 1\right\} + \epsilon^2}}\\
&\quad + \frac{\alpha^2\beta_1\sqrt{1 - \beta_1^{t-1}}G^2L}{\left(1 - \beta_1\right)\sqrt{\beta_2^{t-1} + \epsilon^2}\sqrt{\beta_2^{t-2} + \epsilon^2}} + \frac{\alpha^2G^2L}{2\left(\beta_2^{t-1} + \epsilon^2\right)} \qquad (29)\\
&\leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] - \alpha\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1 - G^2)\beta_2^T, 1\right\} + \epsilon^2}}\\
&\quad + \frac{\alpha^2\left(1 + \beta_1\right)G^2L}{2\left(1 - \beta_1\right)\left(\beta_2^{t-1} + \epsilon^2\right)}. \qquad (30)
\end{aligned}$$

Telescoping it for $t = 1, \ldots, T$, we have

$$\mathbb{E}\left[f\left(\boldsymbol{\phi}_T\right)\right] \leq f\left(\boldsymbol{\theta}_0\right) - \alpha \frac{\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} + \epsilon^2}} + \frac{\alpha^2\left(1+\beta_1\right)G^2 L}{2\left(1-\beta_1\right)} \sum_{t=0}^{T-1} \frac{1}{\beta_2^t + \epsilon^2} \tag{31}$$

$$\leq f\left(\boldsymbol{\theta}_0\right) - \alpha \frac{\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} + \epsilon^2}} + \frac{\alpha^2\left(1+\beta_1\right)G^2 L}{2\left(1-\beta_1\right)} \int_0^T \frac{1}{\beta_2^t + \epsilon^2} dt \tag{32}$$

$$\begin{aligned} &\leq f\left(\boldsymbol{\theta}_0\right) - \alpha \frac{\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} + \epsilon^2}} \\ &\quad + \frac{\alpha^2\left(1+\beta_1\right)G^2 L}{2\left(1-\beta_1\right)\epsilon^2} \left(T - \frac{1}{\log \beta_2} \log\left(\frac{\beta_2^T + \epsilon^2}{1+\epsilon^2}\right)\right) \end{aligned} \tag{33}$$

By rearranging the terms, we have

$$\min_{t=1,\ldots,T} \left\{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}\right\}$$

$$\leq \frac{\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{T} \tag{34}$$

$$\leq \sqrt{\max\left\{G^2, 1\right\} + \epsilon^2} \left(\frac{f\left(\boldsymbol{\theta}_0\right) - f_{\inf}}{\alpha T} + \alpha C_2 \left(1 - \frac{1}{T \log \beta_2} \log\left(\frac{\beta_2^T + \epsilon^2}{1+\epsilon^2}\right)\right)\right) \tag{35}$$

$$\leq C_1\left(T\right) \left(\frac{f\left(\boldsymbol{\theta}_0\right) - f_{\inf}}{\alpha T} + \alpha C_2 \left(1 - \frac{1}{T \log \beta_2} \log\left(\frac{\beta_2^T + \epsilon^2}{1+\epsilon^2}\right)\right)\right), \tag{36}$$

where $C_1\left(T\right) = \sqrt{\max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} + \epsilon^2}$, $C_2 = \frac{(1+\beta_1)G^2 L}{2(1-\beta_1)\epsilon^2}$.

When $\alpha_t = \alpha/\sqrt{t}$, the following holds for $t \geq 2$:

$$\alpha_{t-1} - \alpha_t = \alpha\left(\frac{1}{\sqrt{t-1}} - \frac{1}{\sqrt{t}}\right) \tag{37}$$

$$= \frac{\alpha\left(\sqrt{t} - \sqrt{t-1}\right)}{\sqrt{t\left(t-1\right)}} \tag{38}$$

$$= \frac{\alpha}{\sqrt{t\left(t-1\right)}\left(\sqrt{t} + \sqrt{t-1}\right)} \tag{39}$$

$$\leq \frac{\alpha}{2\left(t-1\right)^{3/2}} \tag{40}$$

$$\leq \frac{\sqrt{2}\alpha}{t^{3/2}}. \tag{41}$$

This also holds for $t = 1$ by defining $\alpha_0 = \alpha$. Applying it to Eq. (28), we have

$$\begin{aligned}
\mathbb{E}\left[f\left(\boldsymbol{\phi}_t\right)\right] \leq\ & \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] + \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1\left(1-\beta_1^{t-1}\right)G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-2}+\epsilon^2}} \\
& - \alpha_t \frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T, 1\right\}+\epsilon^2}} \\
& + \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2\left(1-\beta_1^{t-1}\right)G^2L}{\left(\beta_2^{t-2}+\epsilon^2\right)\left(1-\beta_1\right)^2} + \frac{\alpha_t\alpha_{t-1}\beta_1\sqrt{1-\beta_1^{t-1}}G^2L}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-1}+\epsilon^2}\sqrt{\beta_2^{t-2}+\epsilon^2}} \\
& + \frac{\left(\alpha_{t-1}-\alpha_t\right)^2\beta_1^2\left(1-\beta_1^{t-1}\right)G^2L}{2\left(1-\beta_1\right)^2\left(\beta_2^{t-2}+\epsilon^2\right)} + \frac{\alpha_t^2G^2L}{2\left(\beta_2^{t-1}+\epsilon^2\right)} \\
& + \frac{\alpha_t\left(\alpha_{t-1}-\alpha_t\right)\beta_1\sqrt{1-\beta_1^{t-1}}G^2L}{2\left(1-\beta_1\right)\sqrt{\beta_2^{t-1}+\epsilon^2}\sqrt{\beta_2^{t-2}+\epsilon^2}} \qquad (42) \\
\leq\ & \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] + \frac{\sqrt{2}\alpha\beta_1\left(1-\beta_1^{t-1}\right)G^2}{t^{3/2}\left(1-\beta_1\right)\sqrt{\beta_2^{t-2}+\epsilon^2}} \\
& - \frac{\alpha}{\sqrt{t}} \frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T, 1\right\}+\epsilon^2}} \\
& + \frac{2\alpha^2\beta_1^2G^2L}{\left(\beta_2^{t-1}+\epsilon^2\right)\left(1-\beta_1\right)^2t^2} + \frac{\sqrt{2}\alpha^2\beta_1G^2L}{\left(1-\beta_1\right)\left(\beta_2^{t-1}+\epsilon^2\right)t} \\
& + \frac{\alpha^2\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^{t-1}+\epsilon^2\right)t^3} + \frac{\alpha^2G^2L}{2\left(\beta_2^{t-1}+\epsilon^2\right)t} \\
& + \frac{\alpha^2\beta_1G^2L}{\sqrt{2}\left(1-\beta_1\right)\left(\beta_2^{t-1}+\epsilon^2\right)t^2} \qquad (43) \\
=\ & \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] - \frac{\alpha}{\sqrt{t}} \frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T, 1\right\}+\epsilon^2}} \\
& + \frac{\alpha^2\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{2\left(1-\beta_1\right)\left(\beta_2^{t-1}+\epsilon^2\right)} \cdot t^{-1} + \frac{\sqrt{2}\alpha\beta_1G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}} \cdot t^{-\frac{3}{2}} \\
& + \frac{\alpha^2\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)} \cdot t^{-2} + \frac{\alpha^2\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)} \cdot t^{-3}. \qquad (44)
\end{aligned}$$

Multiplying t to the both sides and rearranging the terms, we have

$$\begin{aligned}
& \frac{\sqrt{t}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2, 1\right\}+\epsilon^2}} \\
& \leq \frac{\mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)-f\left(\boldsymbol{\phi}_t\right)\right]}{\alpha} \cdot t + \frac{\alpha\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{2\left(1-\beta_1\right)\left(\beta_2^{t-1}+\epsilon^2\right)} + \frac{\sqrt{2}\beta_1G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}} \cdot t^{-\frac{1}{2}} \qquad (45) \\
& \quad + \frac{\alpha\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)}t^{-1} + \frac{\alpha\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)}t^{-2} \qquad (46)
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\sqrt{t}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T,1\right\}+\epsilon^2}} \\
&\leq \frac{f\left(\boldsymbol{\phi}_0\right) - Tf\left(\boldsymbol{\phi}_T\right) + \sum_{t=1}^{T-1} f\left(\boldsymbol{\phi}_t\right)}{\alpha} + \frac{\alpha\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{2\left(1-\beta_1\right)}\sum_{t=1}^{T} \frac{1}{\beta_2^{t-1}+\epsilon^2} \\
&\quad + \frac{\sqrt{2}\beta_1 G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}}\sum_{t=1}^{T} t^{-\frac{1}{2}} + \frac{\alpha\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)}\sum_{t=1}^{T} t^{-1} \\
&\quad + \frac{\alpha\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)}\sum_{t=1}^{T} t^{-2} && (47) \\
&\leq \frac{f_{\sup}-f_{\inf}}{\alpha}T + \frac{\alpha\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{2\left(1-\beta_1\right)}\int_0^T \frac{1}{\beta_2^t+\epsilon^2}dt \\
&\quad + \frac{\sqrt{2}\beta_1 G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}}\left(1+\int_1^T t^{-\frac{1}{2}}dt\right) + \frac{\alpha\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(1+\int_1^T t^{-1}dt\right) \\
&\quad + \frac{\alpha\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(1+\int_1^T t^{-2}dt\right) && (48) \\
&\leq \frac{f_{\sup}-f_{\inf}}{\alpha}T + \frac{\alpha\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{2\left(1-\beta_1\right)\epsilon^2}\left(T-\frac{1}{\log\beta_2}\log\left(\frac{\beta_2^T+\epsilon^2}{1+\epsilon^2}\right)\right) \\
&\quad + \frac{\sqrt{2}\beta_1 G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}}\left(2\sqrt{T}-1\right) + \frac{\alpha\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)G^2L}{\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(1+\log T\right) \\
&\quad + \frac{\alpha\beta_1^2G^2L}{\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(2-\frac{1}{T}\right) && (49)
\end{aligned}$$

Therefore, the following bound is derived.

$$\begin{aligned}
&\min_{t=1,\ldots,T}\left\{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}\right\} \\
&\leq \frac{\sum_{t=1}^{T}\sqrt{t}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sum_{t=1}^{T}\sqrt{t}} && (50) \\
&\leq \frac{\sum_{t=1}^{T}\sqrt{t}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\int_0^T \sqrt{t}dt} && (51) \\
&\leq \frac{3C_T\left(f_{\sup}-f_{\inf}\right)}{2\alpha}\frac{1}{\sqrt{T}} + \frac{3\alpha\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)C_TG^2L}{4\left(1-\beta_1\right)\epsilon^2}\left(\frac{1}{\sqrt{T}}-\frac{1}{T^{3/2}\log\beta_2}\log\left(\frac{\beta_2^T+\epsilon^2}{1+\epsilon^2}\right)\right) \\
&\quad + \frac{3\beta_1C_TG^2}{\sqrt{2}\left(1-\beta_1\right)\sqrt{\beta_2^T+\epsilon^2}}\left(\frac{2}{T}-\frac{1}{T^{3/2}}\right) \\
&\quad + \frac{3\alpha\beta_1\left(1+\left(2\sqrt{2}-1\right)\beta_1\right)C_TG^2L}{2\sqrt{2}\left(1-\beta\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(\frac{1}{T^{3/2}}+\frac{\log T}{T^{3/2}}\right) \\
&\quad + \frac{3\alpha\beta_1^2C_TG^2L}{2\left(1-\beta_1\right)^2\left(\beta_2^T+\epsilon^2\right)}\left(\frac{2}{T^{3/2}}-\frac{1}{T^{5/2}}\right), && (52)
\end{aligned}$$

where $C_T = \sqrt{\max\left\{G^2+(1-G^2)\beta_2^T,1\right\}+\epsilon^2}$.

$$
\begin{aligned}
\mathbb{E}\left[f\left(\boldsymbol{\phi}_t\right)\right] \leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] &+ \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G^2}{\epsilon\left(1-\beta_1\right)} - \alpha_t \frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2,1\right\}+\epsilon^2}} \qquad (53)\\
&+ \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha_t\alpha_{t-1}\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)}\\
&+ \frac{\left(\alpha_{t-1}-\alpha_t\right)^2\beta_1^2 G^2 L}{2\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha_t^2 G^2 L}{2\epsilon^2} + \frac{\alpha_t\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G^2 L}{2\epsilon^2\left(1-\beta_1\right)} \qquad (54)\\
\leq \mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)\right] &+ \frac{\sqrt{2}\alpha\beta_1 G^2}{\epsilon\left(1-\beta_1\right)t^{3/2}} - \frac{\alpha}{\sqrt{t}}\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2,1\right\}+\epsilon^2}} \qquad (55)\\
&+ \frac{2\alpha^2\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2 t^2} + \frac{\sqrt{2}\alpha^2\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)t}\\
&+ \frac{\alpha^2\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2 t^3} + \frac{\alpha^2 G^2 L}{2\epsilon^2 t} + \frac{\alpha^2\beta_1 G^2 L}{\sqrt{2}\epsilon^2\left(1-\beta_1\right)t^2}. \qquad (56)
\end{aligned}
$$

Multiplying t to the both sides and rearranging the terms, we have

$$
\begin{aligned}
&\frac{\sqrt{t}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2,1\right\}+\epsilon^2}}\\
&\leq \frac{\mathbb{E}\left[f\left(\boldsymbol{\phi}_{t-1}\right)-f\left(\boldsymbol{\phi}_t\right)\right]}{\alpha}t + \frac{\sqrt{2}\alpha\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)} + \frac{\alpha G^2 L}{2\epsilon^2} \qquad (57)\\
&+ \frac{\sqrt{2}\beta_1 G^2}{\epsilon\left(1-\beta_1\right)}t^{-\frac{1}{2}} + \left(\frac{2\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha^2\beta_1 G^2 L}{\sqrt{2}\epsilon^2\left(1-\beta_1\right)}\right)t^{-1} + \frac{\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2}t^{-2} \qquad (58)
\end{aligned}
$$

Telescoping it for $t = 2, \ldots, T+1$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\frac{\sqrt{t+1}\,\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2,1\right\}+\epsilon^2}}\\
&\leq \frac{2f\left(\boldsymbol{\phi}_1\right)-\left(T+1\right)f\left(\boldsymbol{\phi}_{T+1}\right)+\sum_{t=2}^{T}f\left(\boldsymbol{\phi}_t\right)}{\alpha} + \left(\frac{\sqrt{2}\alpha\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)} + \frac{\alpha G^2 L}{2\epsilon^2}\right)T \qquad (59)\\
&+ \frac{\sqrt{2}\beta_1 G^2}{\epsilon\left(1-\beta_1\right)}\sum_{t=2}^{T+1}t^{-\frac{1}{2}} + \left(\frac{2\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha^2\beta_1 G^2 L}{\sqrt{2}\epsilon^2\left(1-\beta_1\right)}\right)\sum_{t=2}^{T+1}t^{-1} + \frac{\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2}\sum_{t=2}^{T+1}t^{-2}\\
&\leq \frac{f_{\sup}-f_{\inf}}{\alpha}\left(T+1\right) + \left(\frac{\sqrt{2}\alpha\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)} + \frac{\alpha G^2 L}{2\epsilon^2}\right)T + \frac{\sqrt{2}\beta_1 G^2}{\epsilon\left(1-\beta_1\right)}\int_1^{T+1}t^{-\frac{1}{2}}dt \qquad (60)\\
&+ \left(\frac{2\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha^2\beta_1 G^2 L}{\sqrt{2}\epsilon^2\left(1-\beta_1\right)}\right)\int_1^{T+1}t^{-1}dt + \frac{\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2}\int_1^{T+1}t^{-2}dt\\
&\leq \frac{f_{\sup}-f_{\inf}}{\alpha}\left(T+1\right) + \left(\frac{\sqrt{2}\alpha\beta_1 G^2 L}{\epsilon^2\left(1-\beta_1\right)} + \frac{\alpha G^2 L}{2\epsilon^2}\right)T + \frac{2\sqrt{2}\beta_1 G^2}{\epsilon\left(1-\beta_1\right)}\left(\sqrt{T+1}-1\right)\\
&+ \left(\frac{2\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2} + \frac{\alpha^2\beta_1 G^2 L}{\sqrt{2}\epsilon^2\left(1-\beta_1\right)}\right)\log\left(T+1\right) + \frac{\alpha\beta_1^2 G^2 L}{\epsilon\left(1-\beta_1\right)^2}\frac{T}{T+1}. \qquad (61)
\end{aligned}
$$

Therefore, the following bound is derived.

$$\begin{aligned}
&\min_{t=1,\ldots,T}\left\{\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|^{4/3}\right]^{3/2}\right\} \\
&\leq \frac{\sum_{t=1}^{T}\sqrt{t+1}\,\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|^{4/3}\right]^{3/2}}{\sum_{t=1}^{T}\sqrt{t+1}} && (62)\\
&\leq \frac{\sum_{t=1}^{T}\sqrt{t+1}\,\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|^{4/3}\right]^{3/2}}{\int_0^T\sqrt{t+1}dt} && (63)\\
&\leq \frac{3\sqrt{\max\{G^2,1\}+\epsilon^2}}{2\left((T+1)^{3/2}-1\right)}\left(\frac{f_{\sup}-f_{\inf}}{\alpha}(T+1)+\left(\frac{\sqrt{2}\alpha\beta_1G^2L}{\epsilon^2(1-\beta_1)}+\frac{\alpha G^2L}{2\epsilon^2}\right)T\right)\\
&\quad+\frac{3\sqrt{\max\{G^2,1\}+\epsilon^2}}{2\left((T+1)^{3/2}-1\right)}\left(\frac{2\sqrt{2}\beta_1G^2}{\epsilon(1-\beta_1)}\left(\sqrt{T+1}-1\right)+\frac{\alpha\beta_1^2G^2L}{\epsilon(1-\beta_1)^2}\frac{T}{T+1}\right)\\
&\quad+\frac{3\sqrt{\max\{G^2,1\}+\epsilon^2}}{2\left((T+1)^{3/2}-1\right)}\left(\frac{2\alpha\beta_1^2G^2L}{\epsilon(1-\beta_1)^2}+\frac{\alpha^2\beta_1G^2L}{\sqrt{2}\epsilon^2(1-\beta_1)}\right)\log(T+1). && (64)
\end{aligned}$$

□

G Lemmas

Lemma 4. *For all $\boldsymbol{\theta}\in\mathbb{R}^D$ and $t\geq 1$, the following holds*

$$\|\nabla f(\boldsymbol{\theta}_{t-1})\|\leq G. \tag{65}$$

Proof.

$$\begin{aligned}
\|\nabla f(\boldsymbol{\theta}_{t-1})\| &= \sqrt{\|\mathbb{E}[\boldsymbol{g}_t]\|^2} && (66)\\
&\leq \sqrt{\mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]} && (67)\\
&\leq G. && (68)
\end{aligned}$$

The first inequality holds because $\mathbb{E}[(\boldsymbol{g}_t)_i]^2\leq\mathbb{E}[(\boldsymbol{g}_t)_i^2]$, and the second inequality holds due to Assumption 4. □

Lemma 5. *For all $\boldsymbol{\theta}\in\mathbb{R}^D$ and $t\geq 1$, the following holds*

$$\mathbb{E}[\|\boldsymbol{g}_t\|]\leq G \tag{69}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{g}_t\|] &\leq \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]^{1/2} && (70)\\
&\leq G, && (71)
\end{aligned}$$

where the first inequality holds due to the Hölder's inequality and the second one holds due to Assumption 4. □

Lemma 6. *For the RMSprop algorithm, the following holds for $t\geq 1$:*

$$\mathbb{E}\left[\sum_{i=1}^{D}(\boldsymbol{v}_t)_i\right]\leq\left(1-\beta_2^t\right)G^2 \tag{72}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right] &= \mathbb{E}\left[\left(1-\beta_2\right)\sum_{i=1}^{D}\sum_{k=1}^{t}\beta_2^{t-k}\left(\boldsymbol{g}_k\right)_i^2\right] && (73)\\
&\leq \left(1-\beta_2\right)G^2\sum_{k=1}^{t}\beta_2^{t-k} && (74)\\
&= \left(1-\beta_2^t\right)G^2. && (75)
\end{aligned}$$

□

Lemma 7. *For the RMSprop algorithm, the following holds:*

$$\begin{aligned}
&\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right)\right]\\
&\geq \frac{1}{2}\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\frac{\boldsymbol{g}_t}{\sqrt{\tilde{\boldsymbol{v}}_t+\epsilon^2}}\right)\right] - 2G\sqrt{1-\beta_2}\,\mathbb{E}\left[\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right\|^2\right] && (76)
\end{aligned}$$

Proof.

$$\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right)\right] = \sum_{i=1}^{D}\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}}\right] \tag{77}$$

We define $\tilde{\boldsymbol{v}}_t$ as follows:

$$\tilde{\boldsymbol{v}}_t = \beta_2\boldsymbol{v}_{t-1} + \left(1-\beta_2\right)\mathbb{E}\left[\boldsymbol{g}_t\odot\boldsymbol{g}_t\right] \tag{78}$$

Using this, the following holds:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}}\right]\\
&= \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] + \mathbb{E}\left[\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\left(\frac{1}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}} - \frac{1}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right)\right] && (79)\\
&= \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] + \mathbb{E}\left[\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\left(\frac{1}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}} - \frac{1}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right)\right] && (80)\\
&\geq \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] - \mathbb{E}\left[\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\left(\frac{1}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}} - \frac{1}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right)\right|\right], && (81)
\end{aligned}$$

where the last inequality holds due to $A \geq -|A|$. For the second term, the following holds:

$$\begin{aligned}
&\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\left(\frac{1}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}} - \frac{1}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right)\right|\\
&= \left(1-\beta_2\right)\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\frac{\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right] - \left(\boldsymbol{g}_t\right)_i^2}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}\left(\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}+\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}\right)}\right| && (82)\\
&\leq \left(1-\beta_2\right)\left(\frac{\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\right|\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\left(\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2\right)} + \frac{\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\right|\left(\boldsymbol{g}_t\right)_i^2}{\left(\left(\boldsymbol{v}_t\right)_i+\epsilon^2\right)\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right), && (83)
\end{aligned}$$

where the last inequality holds due to the triangle inequality. For the first term, the following holds:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\right| \mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}{\sqrt{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\left(\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2\right)}\right] \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]+\left(1-\beta_2\right) \mathbb{E}\left[\frac{\left(\boldsymbol{g}_t\right)_i^2 \mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]^2}{\left(\left(\boldsymbol{v}_t\right)_i+\epsilon^2\right)\left(\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2\right)^{3/2}}\right] && (84) \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]+\mathbb{E}\left[\frac{\left(\boldsymbol{g}_t\right)_i^2 \sqrt{\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}}{\sqrt{1-\beta_2}\left(\left(\boldsymbol{v}_t\right)_i+\epsilon^2\right)}\right] && (85) \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]+\frac{G}{\sqrt{1-\beta_2}} \mathbb{E}\left[\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\right] && (86)
\end{aligned}$$

The first inequality is derived using the following fact:

$$\forall \lambda>0, x, y \in \mathbb{R}, x y \leq \frac{\lambda}{2} x^2+\frac{y^2}{2 \lambda}. \tag{87}$$

For the second term of Eq. (83), the following holds:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left|\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\left(\boldsymbol{g}_t\right)_i\right|\left(\boldsymbol{g}_t\right)_i^2}{\left(\left(\boldsymbol{v}_t\right)_i+\epsilon^2\right) \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] && (88) \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}} \frac{\left(\boldsymbol{g}_t\right)_i^2}{\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}\right]+\left(1-\beta_2\right) \mathbb{E}\left[\frac{\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}} \frac{\left(\boldsymbol{g}_t\right)_i^4}{\left(\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2\right)^2}\right] && (89) \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]+\mathbb{E}\left[\frac{\sqrt{\mathbb{E}\left[\left(\boldsymbol{g}_t\right)_i^2\right]}\left(\boldsymbol{g}_t\right)_i^2}{\sqrt{1-\beta_2}\left(\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2\right)}\right] && (90) \\
&\leq \frac{1}{\left(1-\beta_2\right)} \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{4 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]+\frac{G}{\sqrt{1-\beta_2}} \mathbb{E}\left[\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\right] && (91)
\end{aligned}$$

The first inequality is derived using Eq. (87).

Putting these inequalities together, the following is derived:

$$\begin{aligned}
&\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^{\top}\left(\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right)\right] \\
&\geq \sum_{i=1}^D \mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{2 \sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right]-2 G \sqrt{1-\beta_2} \mathbb{E}\left[\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\right] && (92) \\
&\geq \frac{1}{2} \mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^{\top}\left(\frac{\boldsymbol{g}_t}{\sqrt{\tilde{\boldsymbol{v}}_t+\epsilon^2}}\right)\right]-2 G \sqrt{1-\beta_2} \mathbb{E}\left[\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right\|^2\right]. && (93)
\end{aligned}$$

□

Lemma 8. *For the RMSprop algorithm, the following holds:*

$$\sum_{t=1}^T \mathbb{E}\left[\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right\|^2\right] \leq D\left(\log \left(1+\frac{\left(1-\beta_2^T\right) G^2}{\epsilon^2}\right)-T \log \beta_2\right) \tag{94}$$

Proof.

$$\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right\|^2=\sum_{i=1}^{D}\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2} \tag{95}$$

$$\begin{align}
\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2} &= \frac{1}{1-\beta_2}\frac{\left(1-\beta_2\right)\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2} \tag{96}\\
&\le -\frac{1}{1-\beta_2}\log\left(1-\frac{\left(1-\beta_2\right)\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\right) \tag{97}\\
&= \frac{1}{1-\beta_2}\log\left(\frac{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}{\beta_2\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}\right) \tag{98}\\
&= \frac{1}{1-\beta_2}\left(\log\left(\frac{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}{\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}\right)+\log\left(\frac{\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}{\beta_2\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}\right)\right) \tag{99}\\
&\le \frac{1}{1-\beta_2}\left(\log\left(\frac{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}{\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}\right)-\log\beta_2\right) \tag{100}
\end{align}$$

$$\begin{align}
\sum_{t=1}^{T}\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2} &\le \frac{1}{1-\beta_2}\left(\log\left(\frac{\left(\boldsymbol{v}_T\right)_i+\epsilon^2}{\epsilon^2}\right)-T\log\beta_2\right) \tag{101}\\
&\le \frac{1}{1-\beta_2}\left(\log\left(1+\frac{\left(1-\beta_2^T\right)G^2}{\epsilon^2}\right)-T\log\beta_2\right) \tag{102}
\end{align}$$

$$\begin{align}
\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t+\epsilon^2}}\right\|^2\right] &\le \sum_{i=1}^{D}\mathbb{E}\left[\sum_{t=1}^{T}\frac{\left(\boldsymbol{g}_t\right)_i^2}{\left(\boldsymbol{v}_t\right)_i+\epsilon^2}\right] \tag{103}\\
&\le \frac{1}{1-\beta_2}\sum_{i=1}^{D}\mathbb{E}\left[\log\left(1+\frac{\left(\boldsymbol{v}_T\right)_i}{\epsilon^2}\right)\right]-\frac{DT\log\beta_2}{1-\beta_2} \tag{104}\\
&\le \sum_{i=1}^{D}\log\left(1+\frac{\mathbb{E}\left[\left(\boldsymbol{v}_T\right)_i\right]}{\epsilon^2}\right)-\frac{DT\log\beta_2}{1-\beta_2} \tag{105}\\
&\le \frac{D}{1-\beta_2}\left(\log\left(1+\frac{\left(1-\beta_2^T\right)G^2}{\epsilon^2}\right)-T\log\beta_2\right) \tag{106}
\end{align}$$

□

Lemma 9. *For the RMSprop algorithm, the following holds:*

$$\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\frac{\boldsymbol{g}_t}{\sqrt{\beta_2\tilde{\boldsymbol{v}}_t+\epsilon^2}}\right)\right]\ge\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\left(1-\beta_2^t\right)G^2+\epsilon^2}} \tag{107}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^{\top}\left(\frac{\boldsymbol{g}_t}{\sqrt{\tilde{\boldsymbol{v}}_t+\epsilon^2}}\right)\right] \\
&=\sum_{i=1}^{D}\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\cdot\left(\boldsymbol{g}_t\right)_i}{\sqrt{\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] \\
&=\sum_{i=1}^{D}\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{\sqrt{\beta_2\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}}\right] \\
&\geq\mathbb{E}\left[\frac{\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^2}{\sqrt{\sum_{i=1}^{D}\left(\tilde{\boldsymbol{v}}_t\right)_i+\epsilon^2}}\right] \\
&\geq\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\mathbb{E}\left[\sum_{i=1}^{D}\left(\tilde{\boldsymbol{v}}_t\right)_i\right]+\epsilon^2}} \\
&\geq\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^{4/3}\right]^{3/2}}{\sqrt{\left(1-\beta_2^t\right)G^2+\epsilon^2}}. \qquad (108)
\end{aligned}$$

The second equality holds due to Assumption 2. The first inequality holds because $\left(\tilde{\boldsymbol{v}}_t\right)_i\geq 0$ for all $i=1,\ldots,D$. The second inequality holds due to the Hölder's inequality. The last inequality holds due to Lemma 6. □

Lemma 10. *For the ADOPT algorithm, the following holds for $t\geq 1$:*

$$\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}=\frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}-\alpha_t\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}, \qquad (109)$$

where we define $\alpha_0=\alpha$.

Proof. For $t=1$, the following holds by definition:

$$\begin{aligned}
\boldsymbol{\phi}_1-\boldsymbol{\phi}_0&=\frac{1}{1-\beta_1}\boldsymbol{\theta}_1-\left(\frac{\beta_1}{1-\beta_1}+1\right)\boldsymbol{\theta}_0 &(110)\\
&=\frac{1}{1-\beta_1}\left(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0\right) &(111)\\
&=-\frac{\alpha_1\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_0+\epsilon^2}}. &(112)
\end{aligned}$$

For $t\geq 2$, the following holds:

$$\begin{aligned}
\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}&=\frac{1}{1-\beta_1}\left(\boldsymbol{\theta}_t-\boldsymbol{\theta}_{t-1}\right)-\frac{\beta_1}{1-\beta_1}\left(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}_{t-2}\right) &(113)\\
&=\frac{1}{1-\beta_1}\left(\alpha_{t-1}\beta_1\boldsymbol{m}_{t-1}-\alpha_t\boldsymbol{m}_t\right) &(114)\\
&=\frac{1}{1-\beta_1}\left(\alpha_{t-1}\beta_1\boldsymbol{m}_{t-1}-\alpha_t\left(\beta_1\boldsymbol{m}_{t-1}+\left(1-\beta_1\right)\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right)\right) &(115)\\
&=\frac{1}{1-\beta_1}\left(\left(\alpha_{t-1}-\alpha_t\right)\beta_1\boldsymbol{m}_{t-1}-\alpha_t\left(1-\beta_1\right)\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right) &(116)\\
&=\frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}-\alpha_t\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}} &(117)
\end{aligned}$$

□

Lemma 11. *For the ADOPT algorithm, the following holds for* $t \geq 1$*:*

$$\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}_{t-1} = -\frac{\alpha_{t-1}\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}. \tag{118}$$

Proof. For $t = 1$, Eq. (118) holds obviously because $\boldsymbol{\phi}_0 = \boldsymbol{\theta}_0$ and $\boldsymbol{m}_0 = \mathbf{0}$. For $t \geq 2$, the following holds:

$$\begin{aligned}
\boldsymbol{\phi}_{t-1} - \boldsymbol{\theta}_{t-1} &= \left(\frac{1}{1-\beta_1} - 1\right)\boldsymbol{\theta}_{t-1} - \frac{\beta_1}{1-\beta_1}\boldsymbol{\theta}_{t-2} && (119)\\
&= \frac{\beta_1}{1-\beta_1}\left(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-2}\right) && (120)\\
&= -\frac{\alpha_{t-1}\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}. && (121)
\end{aligned}$$

□

Lemma 12. *For the ADOPT algorithm, the following holds for* $t \geq 1$*:*

$$\begin{aligned}
&\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\right]\\
&\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)\beta_1\left(1-\beta_1^{t-1}\right)G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-2}+\epsilon^2}} - \alpha_t \frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} + \epsilon^2}}. && (122)
\end{aligned}$$

Proof.

$$\begin{aligned}
&\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right)\\
&= \frac{\left(\alpha_{t-1} - \alpha_t\right)\beta_1}{1-\beta_1}\nabla f\left(\boldsymbol{\theta}_{t-1}\right))^\top\boldsymbol{m}_{t-1} - \alpha_t\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}} && (123)\\
&\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)\beta_1}{1-\beta_1}\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right))\right\|\left\|\boldsymbol{m}_{t-1}\right\| - \alpha_t\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}} && (124)\\
&\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)\beta_1 G}{1-\beta_1}\left\|\boldsymbol{m}_{t-1}\right\| - \alpha_t\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^\top\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}. && (125)
\end{aligned}$$

By taking the expectation for both sides, the following holds:

$$\begin{aligned}
&\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^{\top}\cdot\left(\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}\right)\right] \\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\mathbb{E}\left[\nabla f\left(\boldsymbol{\theta}_{t-1}\right)^{\top}\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right] && (126)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\sum_{i=1}^{D}\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i\cdot\left(\boldsymbol{g}_t\right)_i}{\sqrt{\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}}\right] && (127)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\sum_{i=1}^{D}\mathbb{E}\left[\frac{\left(\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right)_i^2}{\sqrt{\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}}\right] && (128)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\mathbb{E}\left[\frac{\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|^2}{\sqrt{\sum_{i=1}^{D}\left(\boldsymbol{v}_{t-1}\right)_i+\epsilon^2}}\right] && (129)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_{t-1}\right)_i\right]+\epsilon^2}} && (130)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1 G}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\right]-\alpha_t\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T,1\right\}+\epsilon^2}} && (131)\\
&\leq \frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1\left(1-\beta_1^{t-1}\right)G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-2}+\epsilon^2}}-\alpha_t\frac{\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_{t-1}\right)\right\|_i^{4/3}\right]^{3/2}}{\sqrt{\max\left\{G^2+(1-G^2)\beta_2^T,1\right\}+\epsilon^2}}. && (132)
\end{aligned}$$

□

Lemma 13. *For the ADOPT algorithm, the following holds for* $t \geq 0$*:*

$$\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right]\leq\max\left\{G^2+(1-G^2)\beta_2^T,1\right\}. \tag{133}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right] &= \mathbb{E}\left[\beta_2^t+\left(1-\beta_2\right)\sum_{i=1}^{D}\sum_{k=1}^{t}\beta_2^{t-k}\left(\boldsymbol{g}_{k-1}\right)_i^2\right] && (134)\\
&\leq \beta_2^t+\left(1-\beta_2\right)G^2\sum_{k=1}^{t}\beta_2^{t-k} && (135)\\
&= \beta_2^t+\left(1-\beta_2^t\right)G^2 && (136)\\
&= G^2+(1-G^2)\beta_2^t. && (137)
\end{aligned}$$

When $G<1$, the following holds:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right] &\leq G^2+(1-G^2)\beta_2^t && (138)\\
&\leq 1. && (139)
\end{aligned}$$

When $G \geq 1$, the following holds:

$$\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right] \leq G^2 + (1-G^2)\beta_2^t \tag{140}$$

$$\leq G^2 + (1-G^2)\beta_2^T \tag{141}$$

$$\leq G^2. \tag{142}$$

Putting the two together yields:

$$\mathbb{E}\left[\sum_{i=1}^{D}\left(\boldsymbol{v}_t\right)_i\right] \leq \max\left\{G^2 + (1-G^2)\beta_2^T, 1\right\} \tag{143}$$

$$\leq \max\left\{G^2, 1\right\}. \tag{144}$$

□

Lemma 14. *For the ADOPT algorithm, the following holds for* $0 \leq t \leq T$.

$$\mathbb{E}\left[\|\boldsymbol{m}_t\|^2\right] \leq \frac{G^2}{\epsilon^2}. \tag{145}$$

Proof.

$$\mathbb{E}\left[\|\boldsymbol{m}_t\|^2\right] = \mathbb{E}\left[\left\|\beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|^2\right] \tag{146}$$

$$= \mathbb{E}\left[\beta_1^2\|\boldsymbol{m}_{t-1}\|^2 + (1-\beta_1)^2\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|^2 + 2\beta_1(1-\beta_1)\boldsymbol{m}_{t-1}^\top\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right] \tag{147}$$

$$\leq \mathbb{E}\left[\beta_1\|\boldsymbol{m}_{t-1}\|^2 + (1-\beta_1)\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|^2\right] \tag{148}$$

$$\leq \mathbb{E}\left[\beta_1\|\boldsymbol{m}_{t-1}\|^2 + \frac{1-\beta_1}{\beta_2^{t-1}+\epsilon^2}\|\boldsymbol{g}_t\|^2\right] \tag{149}$$

$$\leq \mathbb{E}\left[\frac{1-\beta_1}{\beta_2^{t-1}+\epsilon^2}\sum_{k=1}^{t}\beta_1^{t-k}\|\boldsymbol{g}_k\|^2\right] \tag{150}$$

$$\leq \frac{(1-\beta_1)G^2}{\beta_2^{t-1}+\epsilon^2}\sum_{k=1}^{t}\beta_1^{t-k} \tag{151}$$

$$\leq \frac{(1-\beta_1^t)G^2}{\beta_2^{t-1}+\epsilon^2} \tag{152}$$

$$\leq \frac{G^2}{\epsilon^2}. \tag{153}$$

□

Lemma 15. *For the ADOPT algorithm, the following holds for* $t \geq 0$.

$$\mathbb{E}\left[\|\boldsymbol{m}_t\|\right] \leq \frac{G}{\epsilon} \tag{154}$$

Proof.

$$\mathbb{E}\left[\|\boldsymbol{m}_t\|\right] = \mathbb{E}\left[\left\|(1-\beta_1)\sum_{k=1}^{t}\beta_1^{t-k}\frac{\boldsymbol{g}_k}{\sqrt{\boldsymbol{v}_{k-1}+\epsilon^2}}\right\|\right] \tag{155}$$

$$\leq (1-\beta_1)\sum_{k=1}^{t}\beta_1^{t-k}\mathbb{E}\left[\left\|\frac{\boldsymbol{g}_k}{\sqrt{\boldsymbol{v}_{k-1}+\epsilon^2}}\right\|\right] \tag{156}$$

$$\leq (1-\beta_1)\sum_{k=1}^{t}\frac{\beta_1^{t-k}}{\sqrt{\beta_2^{k-1}+\epsilon^2}}\mathbb{E}\left[\|\boldsymbol{g}_k\|\right] \tag{157}$$

$$\leq \frac{1-\beta_1}{\sqrt{\beta_2^{t-1}+\epsilon^2}}\sum_{k=1}^{t}\beta_1^{t-k}\mathbb{E}\left[\|\boldsymbol{g}_k\|^2\right]^{1/2} \tag{158}$$

$$\leq \frac{(1-\beta_1)\,G}{\sqrt{\beta_2^{t-1}+\epsilon^2}}\sum_{k=1}^{t}\beta_1^{t-k} \tag{159}$$

$$= \frac{(1-\beta_1^t)\,G}{\sqrt{\beta_2^{t-1}+\epsilon^2}} \tag{160}$$

$$\leq \frac{G}{\epsilon}. \tag{161}$$

□

Lemma 16. *For the ADOPT algorithm, the following holds for* $t \geq 1$*:*

$$\mathbb{E}\left[\|\boldsymbol{\phi}_{t-1}-\boldsymbol{\theta}_{t-1}\|\,\|\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}\|\right]$$
$$\leq \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2\left(1-\beta_1^{t-1}\right)G^2}{\left(\beta_2^{t-2}+\epsilon^2\right)\left(1-\beta_1\right)^2} + \frac{\alpha_t\alpha_{t-1}\beta_1\sqrt{1-\beta_1^{t-1}}G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-1}+\epsilon^2}\sqrt{\beta_2^{t-2}+\epsilon^2}}. \tag{162}$$

Proof.

$$\|\boldsymbol{\phi}_{t-1}-\boldsymbol{\theta}_{t-1}\|\,\|\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}\|$$
$$= \left\|-\frac{\alpha_{t-1}\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}\right\|\left\|\frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1}{1-\beta_1}\boldsymbol{m}_{t-1}-\alpha_t\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\| \tag{163}$$

$$\leq \frac{\alpha_{t-1}\beta_1}{1-\beta_1}\|\boldsymbol{m}_{t-1}\|\left(\frac{\left(\alpha_{t-1}-\alpha_t\right)\beta_1}{1-\beta_1}\|\boldsymbol{m}_{t-1}\|+\alpha_t\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|\right) \tag{164}$$

$$\leq \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2}{\left(1-\beta_1\right)^2}\|\boldsymbol{m}_{t-1}\|^2+\frac{\alpha_t\alpha_{t-1}\beta_1}{1-\beta_1}\|\boldsymbol{m}_{t-1}\|\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|. \tag{165}$$

Taking the expectation yields:

$$\mathbb{E}\left[\|\boldsymbol{\phi}_{t-1}-\boldsymbol{\theta}_{t-1}\|\,\|\boldsymbol{\phi}_t-\boldsymbol{\phi}_{t-1}\|\right]$$
$$\leq \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2}{\left(1-\beta_1\right)^2}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^2\right]+\frac{\alpha_t\alpha_{t-1}\beta_1}{1-\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\left\|\frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1}+\epsilon^2}}\right\|\right] \tag{166}$$

$$\leq \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2}{\left(1-\beta_1\right)^2}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|^2\right]+\frac{\alpha_t\alpha_{t-1}\beta_1}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-1}+\epsilon^2}}\mathbb{E}\left[\|\boldsymbol{m}_{t-1}\|\,\|\boldsymbol{g}_t\|\right] \tag{167}$$

$$\leq \frac{\alpha_{t-1}\left(\alpha_{t-1}-\alpha_t\right)\beta_1^2\left(1-\beta_1^{t-1}\right)G^2}{\left(\beta_2^{t-2}+\epsilon^2\right)\left(1-\beta_1\right)^2}+\frac{\alpha_t\alpha_{t-1}\beta_1\sqrt{1-\beta_1^{t-1}}G^2}{\left(1-\beta_1\right)\sqrt{\beta_2^{t-1}+\epsilon^2}\sqrt{\beta_2^{t-2}+\epsilon^2}}. \tag{168}$$

□

Lemma 17. *For the ADOPT algorithm, the following holds for* $t \geq 1$*:*

$$\mathbb{E}\left[\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2\right] \leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2 \left(1 - \beta_1^{t-1}\right) G^2}{\left(1 - \beta_1\right)^2 \left(\beta_2^{t-2} + \epsilon^2\right)} + \frac{\alpha_t^2 G^2}{\beta_2^{t-1} + \epsilon^2} + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1 \sqrt{1 - \beta_1^{t-1}} G^2}{\left(1 - \beta_1\right) \sqrt{\beta_2^{t-1} + \epsilon^2} \sqrt{\beta_2^{t-2} + \epsilon^2}}. \tag{169}$$

Proof.

$$\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2$$

$$= \left\| \frac{\left(\alpha_{t-1} - \alpha_t\right) \beta_1}{1 - \beta_1} \boldsymbol{m}_{t-1} - \alpha_t \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t + \epsilon^2}} \right\|^2 \tag{170}$$

$$= \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2}{\left(1 - \beta_1\right)^2} \left\|\boldsymbol{m}_{t-1}\right\|^2 + \alpha_t^2 \left\| \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}} \right\|^2 - \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1}{1 - \beta_1} \boldsymbol{m}_{t-1}^\top \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}} \tag{171}$$

$$\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2}{\left(1 - \beta_1\right)^2} \left\|\boldsymbol{m}_{t-1}\right\|^2 + \frac{\alpha_t^2}{\beta_2^{t-1} + \epsilon^2} \left\|\boldsymbol{g}_t\right\|^2 + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1}{1 - \beta_1} \left\|\boldsymbol{m}_{t-1}\right\| \left\| \frac{\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_{t-1} + \epsilon^2}} \right\| \tag{172}$$

$$\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2}{\left(1 - \beta_1\right)^2} \left\|\boldsymbol{m}_{t-1}\right\|^2 + \frac{\alpha_t^2}{\beta_2^{t-1} + \epsilon^2} \left\|\boldsymbol{g}_t\right\|^2 + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1}{\left(1 - \beta_1\right) \sqrt{\beta_2^{t-1} + \epsilon^2}} \left\|\boldsymbol{m}_{t-1}\right\| \left\|\boldsymbol{g}_t\right\|. \tag{173}$$

Taking the expectation yields:

$$\mathbb{E}\left[\left\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\right\|^2\right]$$

$$\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2}{\left(1 - \beta_1\right)^2} \mathbb{E}\left[\left\|\boldsymbol{m}_{t-1}\right\|^2\right] + \frac{\alpha_t^2}{\beta_2^{t-1} + \epsilon^2} \mathbb{E}\left[\left\|\boldsymbol{g}_t\right\|^2\right] + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1}{\left(1 - \beta_1\right) \sqrt{\beta_2^{t-1} + \epsilon^2}} \mathbb{E}\left[\left\|\boldsymbol{m}_{t-1}\right\| \left\|\boldsymbol{g}_t\right\|\right] \tag{174}$$

$$\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2 \left(1 - \beta_1^{t-1}\right) G^2}{\left(1 - \beta_1\right)^2 \left(\beta_2^{t-2} + \epsilon^2\right)} + \frac{\alpha_t^2 G^2}{\beta_2^{t-1} + \epsilon^2} + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1}{\left(1 - \beta_1\right) \sqrt{\beta_2^{t-1} + \epsilon^2}} \mathbb{E}\left[\left\|\boldsymbol{m}_{t-1}\right\| \left\|\boldsymbol{g}_t\right\|\right] \tag{175}$$

$$\leq \frac{\left(\alpha_{t-1} - \alpha_t\right)^2 \beta_1^2 \left(1 - \beta_1^{t-1}\right) G^2}{\left(1 - \beta_1\right)^2 \left(\beta_2^{t-2} + \epsilon^2\right)} + \frac{\alpha_t^2 G^2}{\beta_2^{t-1} + \epsilon^2} + \frac{\alpha_t \left(\alpha_{t-1} - \alpha_t\right) \beta_1 \sqrt{1 - \beta_1^{t-1}} G^2}{\left(1 - \beta_1\right) \sqrt{\beta_2^{t-1} + \epsilon^2} \sqrt{\beta_2^{t-2} + \epsilon^2}}. \tag{176}$$

□

H Details of Experimental Setups

H.1 Code

Our implementation for the experiment is available at `https://anonymous.4open.science/r/adopt-iclr2024-submission-C7BE`.

H.2 Total amount of compute

We run our experiments mainly on cloud GPU instances with $8\times$ A100. It took approximately 320 hours for our experiments in total.

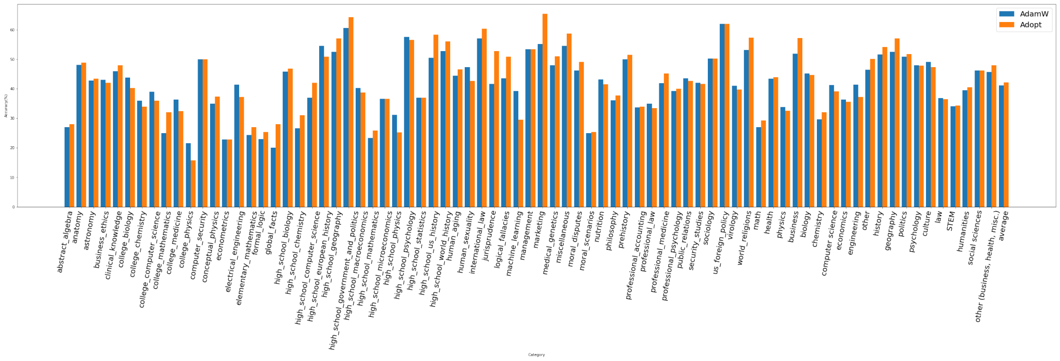

Figure 5: Comparison of MMLU scores for LLaMA-7B finetuned via instruction following using AdamW and ADOPT.

H.3 License of Assets

Datasets: The MNIST database is downloaded from http://yann.lecun.com/exdb/mnist, which is license-free. The terms of access for the ImageNet database is provided at https://www.image-net.org/download. The dataset of Stanford Alpaca is CC BY NC 4.0 (allowing only non-commercial use).

Pretrained models: The pretrained model of LLaMA is provided under GNU General Public License v3.0.

Simulator: MuJoCo is provided under Apache License 2.0.

Code: Our implementation of ImageNet classification is based on the Torchvision's official training recipe provided at https://github.com/UiPath/torchvision/tree/master/references/classification. Torchvision is provided under BSD 3-Clause License. We use the official implementation of NVAE provided at https://github.com/NVlabs/NVAE, whose license is described at https://github.com/NVlabs/NVAE/blob/master/LICENSE.