myjournal manuscript No.
(will be inserted by the editor)

Theoretical analysis of Adam using hyperparameters close to one without Lipschitz smoothness

Hideaki Iiduka

Received: date / Accepted: date

Abstract Convergence and convergence rate analyses of adaptive methods, such as Adaptive Moment Estimation (Adam) and its variants, have been widely studied for nonconvex optimization. The analyses are based on assumptions that the expected or empirical average loss function is Lipschitz smooth (i.e., its gradient is Lipschitz continuous) and the learning rates depend on the Lipschitz constant of the Lipschitz continuous gradient. Meanwhile, numerical evaluations of Adam and its variants have clarified that using small constant learning rates without depending on the Lipschitz constant and hyperparameters (β_1 and β_2) close to one is advantageous for training deep neural networks. Since computing the Lipschitz constant is NP-hard, the Lipschitz smoothness condition would be unrealistic. This paper provides theoretical analyses of Adam without assuming the Lipschitz smoothness condition in order to bridge the gap between theory and practice. The main contribution is to show theoretical evidence that Adam using small learning rates and hyperparameters close to one performs well, whereas the previous theoretical results were all for hyperparameters close to zero. Our analysis also leads to the finding that Adam performs well with large batch sizes. Moreover, we show that Adam performs well when it uses diminishing learning rates and hyperparameters close to one.

Keywords Adam · adaptive method · batch size · hyperparameters · learning rate · nonconvex optimization

Mathematics Subject Classification (2020) 65K05 · 90C15 · 90C26

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 21K11773.

H. Iiduka
Department of Computer Science, Meiji University
1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, Japan
E-mail: iiduka@cs.meiji.ac.jp

1 Introduction

1.1 Background

One way to train a deep neural network is to find the model parameters of the network that minimize loss functions called the expected risk and empirical risk by using first-order optimization methods [3, Section 4]. The simplest optimizer is mini-batch stochastic gradient descent (SGD) [23,35,19,9,10]. There are many deep learning optimizers to accelerate SGD, such as momentum methods [21,20] and adaptive methods, e.g., Adaptive Gradient (AdaGrad) [7], Root Mean Square Propagation (RMSProp) [27], Adaptive Moment Estimation (Adam) [14], Yogi [31], Adaptive Mean Square Gradient (AMSGrad) [22], Adam with decoupled weight decay (AdamW) [16], and AdaBelief (named for adapting stepsizes by the belief in observed gradients) [34].

Convergence and convergence rate analyses of deep learning optimizers have been widely studied for convex optimization [36,14,22,17,18]. Meanwhile, theoretical investigations on deep learning optimizers for nonconvex optimization are needed so that these optimizers can be practically used for nonconvex optimization in deep learning [30,1,28].

Convergence analyses of SGD for nonconvex optimization were presented in [8, 4,24,15] (see [11,15] for convergence analyses of SGD for two classes of nonconvex optimization problem, quasar-convex and Polyak–Lojasiewicz optimization). For example, Theorem 12 in [24] gave an upper bound of $(1/K)\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$ generated by SGD with a constant learning rate $\alpha = 1/L$ is $O(1/K)+C$, where $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}}$ is the sequence generated by an optimizer, L is the Lipschitz constant of the Lipschitz continuous gradient of the loss function $f\colon \mathbb{R}^d \to \mathbb{R}$, K denotes the number of steps, and $C > 0$ is a constant. Convergence analyses depending on the batch size were presented in [4]. In particular, Theorem 3.2 in [4] indicates that running SGD with a diminishing learning rate $\alpha_k = 1/k$ and a large batch size for sufficiently many steps leads to convergence to a stationary point of the sum of loss functions.

Convergence analyses of adaptive methods for nonconvex optimization were presented in [31,37,6,33,34,5,13]. The previous results are summarized in Table 1. Theorems 1 and 2 in [31] indicate that, if $\alpha_k = O(1/L)$ and a hyperparameter $\beta_2 \geq 1 - O(1/G^2)$, then the upper bounds of $(1/K)\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$ generated by Adam and Yogi are each $O(1/K+1/b)$, where G is the upper bound of the stochastic gradient and b is the batch size. Theorem 2 in [29] indicates that computing the Lipschitz constant L is NP-hard. Hence, using a learning rate depending on the Lipschitz constant L would be unrealistic. Convergence analyses of adaptive methods using diminishing learning rates that do not depend on L were presented in [37,6,34,13], while convergence analyses of adaptive methods using constant learning rates that do not depend on L were presented in [33,5,13]. These studies indicate that, if K is sufficiently large and if β_1 and β_2 are close to 0, then adaptive methods using learning rates that do not depend on L approximate the stationary points of f. For example,

Table 1 Upper bounds of performance measures of optimizers with learning rate α_k and hyperparameters β_1 and β_2 for nonconvex optimization ($C, G > 0$, $s \in (0, 1/2)$, L denotes the Lipschitz constant of the Lipschitz continuous gradient of the loss function, K denotes the number of steps, b is the batch size, and C_1 is a monotone decreasing function. $\beta \approx a$ implies that, if β is close to a, then the upper bounds are small.)

Optimizer	Learning Rate α_k	Parameters β_1, β_2	Upper Bound
SGD (Scaman et al., 2020)	$\dfrac{1}{L}$	—	$O\left(\dfrac{1}{K}\right) + C$
Adam [31]	$O\left(\dfrac{1}{L}\right)$	$\beta_2 \geq 1 - O\left(\dfrac{1}{G^2}\right)$	$O\left(\dfrac{1}{K} + \dfrac{1}{b}\right)$
Yogi [31]	$O\left(\dfrac{1}{L}\right)$	$\beta_2 \geq 1 - O\left(\dfrac{1}{G^2}\right)$	$O\left(\dfrac{1}{K} + \dfrac{1}{b}\right)$
Generic Adam [37]	$O\left(\dfrac{1}{\sqrt{k}}\right)$	$\beta_1 \approx 0$, $\beta_2 = 1 - \dfrac{1}{k} \approx 1$	$O\left(\dfrac{\log K}{\sqrt{K}}\right)$
AdaFom [6]	$\dfrac{1}{\sqrt{k}}$	$\beta_1 \approx 0$	$O\left(\dfrac{\log K}{\sqrt{K}}\right)$
AMSGrad [33]	α	$0 \approx \beta_1 < \sqrt{\beta_2}$	$O\left(\dfrac{1}{K^{\frac{1}{2}-s}}\right)$
AdaBelief [34]	$O\left(\dfrac{1}{\sqrt{k}}\right)$	$\beta_1 \approx 0$, $\beta_2 \approx 0$	$O\left(\dfrac{\log K}{\sqrt{K}}\right)$
Padam [5]	α	$\beta_1 \approx 0$, $\beta_2 \approx 0$	$O\left(\dfrac{1}{K^{\frac{1}{2}-s}}\right)$
Adaptive methods [13]	α	$\beta_1 \approx 0$, $\beta_2 \approx 0$	$O\left(\dfrac{1}{K}\right) + C_1(\alpha, \beta_1)$
Adaptive methods [13]	$\dfrac{1}{\sqrt{k}}$	$\beta_1 \approx 0$, $\beta_2 \approx 0$	$O\left(\dfrac{1}{\sqrt{K}}\right)$

Theorem 3 in [33] indicates that AMSGrad using a constant learning rate α satisfies

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_k)\|^2\right] \leq \frac{M_1}{K\alpha} + \frac{M_2 d}{K} + \frac{\alpha M_3 d}{K^{\frac{1}{2}-s}},$$

where $s \in [0, 1/2]$ and M_i $(i = 1, 2, 3)$ are positive constants, and $M_2 = O(G^3/(1 - \beta_1))$ and $M_3 = O(G^2/(1 - \beta_2))$ depend on β_1 and β_2 and the upper bound G. M_2 and M_3 are small when β_1 and β_2 are small, i.e., $\beta_1, \beta_2 \approx 0$. Hence, using β_1 and β_2 close to 0 is advantageous for AMSGrad.

Meanwhile, numerical evaluations have shown that using β_1 and β_2 such as

$$\beta_1 \in \{0.9, 0.99\} \text{ and } \beta_2 \in \{0.99, 0.999\} \tag{1}$$

is advantageous for training deep neural networks [14,22,31,37,6,34,5]. The practically useful β_1 and β_2 defined by (1) are each close to 1, whereas in contrast the theoretical results in Table 1 show that using β_1 and β_2 close to 0 makes the upper bounds of the performance measures small. Hence, there is a gap between theory ($\beta_1, \beta_2 \approx 0$; see also Table 1) and practice ($\beta_1, \beta_2 \approx 1$; see also (1)) for adaptive methods. In [26], it was numerically shown that using an enormous batch size leads to a reduction in the number of parameter updates and model training time. The theoretical results in [31] showed that using large batch sizes makes the upper bound of $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$ small. Accordingly, the practical results for large batch sizes matches the theoretical ones.

1.2 Motivation

As indicated in Section 1.1, it was numerically shown that the performance of an adaptive method strongly depends on the hyperparameters β_1 and β_2 being close to 1. The motivation of this paper is to show *theoretically* evidence such that Adam performs well when β_1 and β_2 are each set close to 1. Using the Lipschitz constant of the Lipschitz continuous gradient of the loss function would be unrealistic [29]. Hence, it will not be assumed that the loss function f is Lipschitz smooth (i.e., its gradient is Lipschitz continuous). The Lipschitz smoothness condition of f implies that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{2}$$

Almost all of the previous analyses of adaptive methods are based on the descent lemma (2), and hence, they can use the expectation of the squared norm of the full gradient $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$ as the performance measure. Since we do not assume Lipschitz smoothness of the loss function, we cannot use (2). Accordingly, we must use other performance measures that are different from $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$.

1.2.1 Performance measure

This paper considers the Adam optimizer [14], which is defined for all $k \in \mathbb{N}$ by

$$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \frac{\alpha_k}{1 - \beta_1^{k+1}} \mathsf{diag}\left(\hat{v}_{k,i}^{-\frac{1}{2}} \right) \boldsymbol{m}_k, \tag{3}$$

where $\alpha_k > 0$ is the learning rate, $\beta_i \in (0,1)$ $(i = 1,2)$, $\hat{v}_{k,i} := (1 - \beta_2^{k+1})^{-1} v_{k,i}$, $v_{k,i} := \beta_2 v_{k-1,i} + (1 - \beta_2) g_{k,i}^2$, $m_{k,i} := \beta_1 m_{k-1,i} + (1 - \beta_1) g_{k,i}$, $\boldsymbol{m}_k := (m_{k,i})_{i=1}^d$, $\boldsymbol{g}_k := (g_{k,i})_{i=1}^d$ is the stochastic gradient, and $\mathsf{diag}(x_i)$ is a diagonal matrix with diagonal components

$x_1, x_2, \ldots, x_d$ (see also Algorithm 1 for details). We use the following theoretical performance measure to approximate a local minimizer $\boldsymbol{\theta}^\star$ of the nonconvex optimization problem of minimizing the loss function $f\colon \mathbb{R}^d \to \mathbb{R}$ over $\mathbb{R}^d$:

$$\mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \varepsilon, \tag{4}$$

where $\varepsilon > 0$ is the precision. The performance measure (4) is an ε-approximation of the sequence $(\boldsymbol{\theta}_k)_{k \in \mathbb{N}}$ generated by Adam (3) in the sense that the inner product of the vector $\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star$ and the inverse direction $\boldsymbol{m}_k$ of the search direction $-\boldsymbol{m}_k$ is less than or equal to ε. Thanks to the previous theoretical results shown in Table 1, it is guaranteed that Adam and its variants can find a stationary point $\boldsymbol{\theta}^\star$ of f, which implies that, for a sufficiently large step size k, $0 \leq \mathbb{E}[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)]$. Therefore, it is sufficient to check whether or not Adam satisfies (4). If $\mathbb{E}[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)] \approx 0$ for a sufficiently large k, then $\mathbb{E}[\|\boldsymbol{m}_k\|]$ or $\mathbb{E}[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|]$ will be approximately zero. We also use the mean value of $\mathbb{E}[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)]$ ($k \in \{1, 2, \ldots, K\}$), that is,

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \varepsilon. \tag{5}$$

Let $\boldsymbol{\theta}^\star \in \mathbb{R}^d$. $\boldsymbol{\theta}^\star$ is a stationary point of f if and only if $\nabla f(\boldsymbol{\theta}^\star)^\top (\boldsymbol{\theta}^\star - \boldsymbol{\theta}) \leq 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Hence, we also use the following performance measures: for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \varepsilon \tag{6}$$

and

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \varepsilon. \tag{7}$$

The advantage of using (4), (5), (6), and (7) is that we can evaluate the upper bound of the performance measure for Adam without assuming that the loss function f is Lipschitz smooth.

1.3 Our results and contribution

Numerical evaluations presented in [14,22,31,37,6,34,5] showed that Adam and its variants perform well when they use a small constant learning rate α and hyperparameters β_1 and β_2 with values close to 1. Hence, we would like to show theoretical evidence that Adam performs well when we set a small α and β_1 and β_2 close to 1. In particular, we will show that Adam (3) using $\alpha > 0$, $\beta_1 \in (0,1)$, and $\beta_2 \in [0,1)$

satisfies

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)\right] \leq{}& \underbrace{\frac{D(\boldsymbol{\theta}^\star)M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2}}_{C_1(\beta_1,b)} + \underbrace{\frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b}+G^2\right)}_{C_2(\alpha,\beta_1,\beta_2,b,k)} \\ &+ \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}\end{aligned} \tag{8}$$

and

$$\begin{aligned}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right] \leq{}& \underbrace{\frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2}}_{C_1(\beta_1,b)} + \underbrace{\frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b}+G^2\right)}_{C_2(\alpha,\beta_1,\beta_2,b,k)} \\ &+ \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},\end{aligned}$$

where $\boldsymbol{\theta}^\star$ is a stationary point of f and $\boldsymbol{\theta} \in \mathbb{R}^d$ (see Theorem 1 for the definitions of the parameters). C_1, C_3, C_4, and C_5 are monotone decreasing for $\beta_1 \in (0,1)$ and $b > 0$. Hence, it is desirable to set β_1 close to 1 such as $\beta_1 = 0.9$ and a large batch size b. Although a function $f_k(\beta_1) := 1/(\beta_1(1-\beta_1^{k+1}))$ in C_2 is monotone increasing for β_1 satisfying $\beta_1^{k+1} > 1/(k+2)$, $f_k(\beta_1)$ with a sufficiently large k is monotone decreasing for β_1. $f_{\beta_1}(k) := 1/(\beta_1(1-\beta_1^{k+1}))$ is monotone decreasing for $k \in \mathbb{N}$. Although $f_{\beta_2}(k) := (1-\beta_2^{k+1})^{1/2}$ is monotone increasing for $k \in \mathbb{N}$, $f_{\beta_2}(k) := (1-\beta_2^{k+1})^{1/2} \leq 1$ is small for all β_2 and all $k \in \mathbb{N}$. C_2 is monotone increasing for α and monotone decreasing for β_2. Hence, we must set a small α and β_2 close to 1, such as $\alpha = 10^{-3}$ and $\beta_2 = 0.99$, to make C_2 small. For simplicity, we may evaluate

$$C_2(\alpha,\beta_1,\beta_2,b,k) := \frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b}+G^2\right) \leq \frac{\alpha}{2\sqrt{v_*}\beta_1(1-\beta_1)}\left(\frac{\sigma^2}{b}+G^2\right).$$

Since $1/(\beta_1(1-\beta_1))$ is monotone increasing for $\beta_1 \geq 1/2$, we need to set a small learning rate α to make $\alpha/(\beta_1(1-\beta_1))$ small. Therefore, using a small learning rate α and β_1 and β_2 close to 1 is advantageous for Adam (see Theorem 2 for the results for when Adam uses a diminishing learning rate α_k). Moreover, (8) implies that

$$\limsup_{k \to +\infty} \mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{D(\boldsymbol{\theta}^\star) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_1} \sqrt{\frac{\sigma^2}{b} + G^2}}_{C_1(\beta_1, b)} + \underbrace{\frac{\alpha}{2\sqrt{v_*}\beta_1} \left(\frac{\sigma^2}{b} + G^2\right)}_{\bar{C}_2(\alpha, \beta_1, b)}$$
$$+ \underbrace{\frac{1-\beta_1}{\beta_1} D(\boldsymbol{\theta}^\star) G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1) D(\boldsymbol{\theta}^\star) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right)}_{C_4(\beta_1, b)},$$

which implies that we must set a small learning rate α, β_1 close to 1, and a large batch size b for C_1, $\bar{C}_2$, C_3, and C_4 to be small (see also Corollary 1).

Here, we give numerical evidence showing that using β_1 and β_2 close to 1 is advantageous for Adam. We trained a Residual Network 18 (ResNet-18) on the CIFAR-10 dataset[1] and the MNIST dataset[2]. We set a constant learning rate $\alpha = 10^{-3}$ and batch size $b = 2^7$ and checked how using different hyperparameters β_1 and β_2 affected the performance of Adam. Figure 1 plots the training loss function value versus number of epochs on the CIFAR-10 dataset, and Figure 2 plots the training loss function value versus number of epochs on the MNIST dataset. The figures indicate that Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ minimizes the loss function f faster than Adam with $\beta_1 = \beta_2 = 0.1$ and Adam with $\beta_1 = \beta_2 = 0.5$. Hence, the figures imply that using β_1 and β_2 close to 1 is advantageous in the sense that Adam finds a local minimizer $\boldsymbol{\theta}^\star$ quickly. The numerical results match our theoretical analysis (8) indicating that the upper bound of the performance measure (4) becomes small when β_1 and β_2 are close to 1.

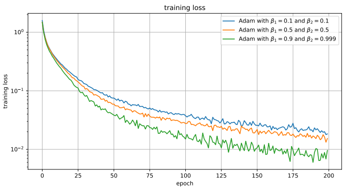

Fig. 1 Training loss function value for Adam with a constant learning rate $\alpha = 10^{-3}$, batch size $b = 2^7$, $\beta_1 \in \{0.1, 0.5, 0.9\}$, and $\beta_2 \in \{0.1, 0.5, 0.999\}$ versus number of epochs on the CIFAR-10 dataset

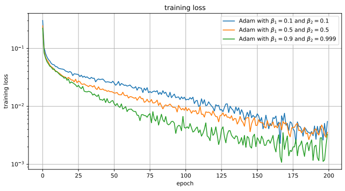

Fig. 2 Training loss function value for Adam with a constant learning rate $\alpha = 10^{-3}$, batch size $b = 2^7$, $\beta_1 \in \{0.1, 0.5, 0.9\}$, and $\beta_2 \in \{0.1, 0.5, 0.999\}$ versus number of epochs on the MNIST dataset

A stochastic convex optimization problem exists such that Adam using $\beta_1 < \sqrt{\beta_2}$ (e.g., $\beta_1 = 0.9$ and $\beta_2 = 0.999$) does not converge to the optimal solution [22, Theorem 3], so it is not guaranteed that Adam can solve every nonconvex optimization

[1] `https://www.cs.toronto.edu/~kriz/cifar.html`

[2] `http://yann.lecun.com/exdb/mnist/`

problem. Reddi et al showed that, if $v_{k,i}$ in (3) satisfies

$$\hat{v}_{k+1,i} \geq \hat{v}_{k,i} \text{ for all } k \in \mathbb{N} \text{ and all } i \in [d], \tag{9}$$

then Adam (AMSGrad) with a diminishing learning rate α_k and β_1 and β_2 defined by

$$\alpha_k = O\left(\frac{1}{\sqrt{k}}\right) \text{ and } \beta_1 < \sqrt{\beta_2} \tag{10}$$

can solve convex optimization problems [22, (2), Algorithm 2, Theorem 4] (see also [34, Theorems 2.1 and 2.2] for the convergence analyses of AdaBelief using (9)). Motivated by the results in [22,34], we decided to study Adam under Condition (9). Our results are summarized in Table 2 (see Theorems 3, 4, 5, and 6 for the details). While the previous results shown in Table 1 used $\beta_1, \beta_2 \approx 0$, our results use $\beta_1, \beta_2 \approx 1$ to make the upper bound of (5) small (see Theorems 3, 4, 5, and 6 for the upper bounds of (7)). Therefore, the results we present in this paper are theoretical confirmation of the numerical evaluations [14,22,31,37,6,34,5] showing that Adam and its variants using β_1 and β_2 close to 1 perform well.

Table 2 Upper bounds of the performance measure (5) of Adam with learning rate α_k and hyperparameters β_1 and β_{2k} for nonconvex optimization ($a > 0$, C_i $(i = 1,2,3)$ are constants, K is the number of steps, b is the batch size, and n is the number of samples. $\beta \approx \gamma$ implies that, if β is close to γ, then the upper bounds are small.)

Optimizer	Learning Rate α_k	Parameters β_1, β_{2k}	Batch b	Upper Bound of (5)
Adam with (9) (Theorem 3)	α	$\beta_1 \approx 1$ $\beta_2 \approx 1$	$b \approx n$	$O\left(\frac{1}{K} + \frac{\alpha}{b} + \frac{1-\beta_1}{\beta_1}\right)$
Adam with (9) (Theorem 4)	α	$\beta_1 \approx 1$ $\beta_{2k} \to 1$	$b \approx n$	$O\left(\frac{1}{\sqrt{K}} + \frac{1-\beta_1}{\beta_1}\right)$
Adam with (9) (Theorem 5)	$\frac{1}{\sqrt{k}}$	$\beta_1 \approx 1$ $\beta_2 \approx 1$	$b \approx n$	$O\left(\frac{1}{\sqrt{K}} + \frac{1-\beta_1}{\beta_1}\right)$
Adam with (9) (Theorem 6)	$\frac{1}{k^a}$	$\beta_1 \approx 1$ $\beta_{2k} \to 1$	$b \approx n$	$O\left(\frac{1}{\sqrt{K}} + \frac{1-\beta_1}{\beta_1}\right)$

Our first contribution is to provide theoretical analyses of Adam without assuming the Lipschitz smoothness condition. Since we cannot use the descent lemma (2) based on this condition, we cannot use the performance measure $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|^2]$. Instead, we use two performance measures (4) and (6): (4) is the inner product of $\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star$ and the inverse direction $\boldsymbol{m}_k$ of the search direction generated by Adam. If the (4) is approximately zero, then Adam can approximate a local minimizer $\boldsymbol{\theta}^\star$ of the problem of minimizing f. (6) is the inner product of $\boldsymbol{\theta}_k - \boldsymbol{\theta}$ and the full gradient $\nabla f(\boldsymbol{\theta}_k)$ generated by Adam. If (6) is approximately zero, then Adam can approximate a local minimizer $\boldsymbol{\theta}^\star$ of the problem of minimizing f. (4) and (6) and the mean values (5) and

(7) of (4) and (6) can be used to evaluate the performance of deep learning optimizers without assuming the Lipschitz smoothness condition.

While the numerical evaluations presented in [14,22,31,37,6,34,5] have shown that adaptive methods using β_1 and β_2 close to 1 are advantageous for training deep neural networks, the theoretical results in Table 1 imply that adaptive methods with β_1 and β_2 close to 0 are good for solving nonconvex optimization problems in deep learning. This implies that there is a large gap between theory and practice for adaptive methods. Our results in Table 2 show that Adam indeed performs well when β_1 and β_2 are set close to 1. Thus, the gap between theory and practice can be bridged for Adam. Our results also show that using a large batch size makes the upper bounds of the performance measures small, which implies that our results match the numerical evaluations in [26].

The remainder of the paper is as follows. First, the mathematical preliminaries are laid out in Section 2, with the definitions of nonconvex optimization and the Adam optimizer. Section 3 describes the theoretical results in detail. Finally, a brief summary and outline of future work are presented in Section 4.

2 Mathematical preliminaries

2.1 Nonconvex optimization

Let $\mathbb{R}^d$ be a d-dimensional Euclidean space with inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y}$ inducing the norm $\|\boldsymbol{x}\|$ and $\mathbb{N}$ the set of nonnegative integers. Let H be a positive-definite matrix denoted by $H \in \mathbb{S}^d_{++}$. The H-inner product of $\mathbb{R}^d$ is defined for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_H := \langle \boldsymbol{x}, H\boldsymbol{y} \rangle = \boldsymbol{x}^\top (H\boldsymbol{y})$, and the H-norm is defined by $\|\boldsymbol{x}\|_H := \sqrt{\langle \boldsymbol{x}, H\boldsymbol{x} \rangle}$. Let $[d] := \{1, 2, \ldots, d\}$ for $d \geq 1$. The mathematical model used in this paper is based on [25]. Given a parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ and given a data point z in a data domain Z, a machine-learning model provides a prediction whose quality is measured by a differentiable nonconvex loss function $\ell(\boldsymbol{\theta}; z)$. We aim to minimize the expected loss defined for all $\boldsymbol{\theta} \in \mathbb{R}^d$ by

$$f(\boldsymbol{\theta}) = \mathbb{E}_{z \sim \mathscr{D}}[\ell(\boldsymbol{\theta}; z)] = \mathbb{E}[\ell_\xi(\boldsymbol{\theta})], \tag{11}$$

where $\mathscr{D}$ is a probability distribution over Z, ξ denotes a random variable with distribution function P, and $\mathbb{E}[\cdot]$ denotes the expectation taken with respect to ξ. A particularly interesting example of (11) is the empirical average loss defined for all $\boldsymbol{\theta} \in \mathbb{R}^d$ by

$$f(\boldsymbol{\theta}; S) = \frac{1}{n} \sum_{i \in [n]} \ell(\boldsymbol{\theta}; z_i) = \frac{1}{n} \sum_{i \in [n]} \ell_i(\boldsymbol{\theta}), \tag{12}$$

where $S = (z_1, z_2, \ldots, z_n)$ denotes the training set, $\ell_i(\cdot) := \ell(\cdot; z_i)$ denotes the loss function corresponding to the i-th training data z_i, and $[n] := \{1, 2, \ldots, n\}$. Our main objective is to find a local minimizer of f over $\mathbb{R}^d$, i.e., a stationary point $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ satisfying $\nabla f(\boldsymbol{\theta}^\star) = \boldsymbol{0}$.

2.2 Adam

We assume that a stochastic first-order oracle (SFO) exists such that, for a given $\boldsymbol{\theta} \in \mathbb{R}^d$, it returns a stochastic gradient $\mathsf{G}_\xi(\boldsymbol{\theta})$ of the function f defined by (11), where a random variable ξ is supported on Ξ independently of $\boldsymbol{\theta}$. Throughout this paper, we will assume the following standard conditions:

(S1) $f \colon \mathbb{R}^d \to \mathbb{R}$ defined by (11) is continuously differentiable;

(S2) Let $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^d$ be the sequence generated by a deep learning optimizer. For each iteration k,

$$\mathbb{E}_{\xi_k}[\mathsf{G}_{\xi_k}(\boldsymbol{\theta}_k)] = \nabla f(\boldsymbol{\theta}_k), \tag{13}$$

where $\xi_0, \xi_1, \ldots$ are independent samples and the random variable ξ_k is independent of $(\boldsymbol{\theta}_l)_{l=0}^k$. There exists a nonnegative constant σ^2 such that

$$\mathbb{E}_{\xi_k}\left[\left\|\mathsf{G}_{\xi_k}(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k)\right\|^2\right] \leq \sigma^2. \tag{14}$$

(S3) For each iteration k, the optimizer samples a batch B_k of size b independently of k and estimates the full gradient ∇f as

$$\nabla f_{B_k}(\boldsymbol{\theta}_k) := \frac{1}{b}\sum_{i\in[b]} \mathsf{G}_{\xi_{k,i}}(\boldsymbol{\theta}_k),$$

where $\xi_{k,i}$ is a random variable generated by the i-th sampling in the k-th iteration.

In the case that f is defined by (12), we have that, for each k, $B_k \subset [n]$ and

$$\nabla f_{B_k}(\boldsymbol{\theta}_k) = \frac{1}{b}\sum_{i\in[b]} \nabla \ell_{\xi_{k,i}}(\boldsymbol{\theta}_k).$$

Algorithm 1 is the Adam optimizer under (S1)–(S3). The symbol $\odot$ in step 6 is defined for all $\boldsymbol{x} = (x_i)_{i=1}^d \in \mathbb{R}^d$, $\boldsymbol{x} \odot \boldsymbol{x} := (x_i^2)_{i=1}^d \in \mathbb{R}^d$. $\mathsf{diag}(x_i)$ in step 8 is a diagonal matrix with diagonal components $x_1, x_2, \ldots, x_d$.

Algorithm 1 Adam [14]

```
Require: $\alpha_k \in (0,+\infty)$, $b \in (0,+\infty)$, $\beta_{1k} \in (0,1)$, $\beta_{2k} \in [0,1)$
 1: $k \leftarrow 0$, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, $\boldsymbol{m}_{-1} := \mathbf{0}$, $\boldsymbol{v}_{-1} := \mathbf{0}$
 2: loop
 3:     $\nabla f_{B_k}(\boldsymbol{\theta}_k) := b^{-1}\sum_{i\in[b]} \mathsf{G}_{\xi_{k,i}}(\boldsymbol{\theta}_k)$
 4:     $\boldsymbol{m}_k := \beta_{1k}\boldsymbol{m}_{k-1} + (1-\beta_{1k})\nabla f_{B_k}(\boldsymbol{\theta}_k)$
 5:     $\hat{\boldsymbol{m}}_k := (1-\beta_{1k}^{k+1})^{-1}\boldsymbol{m}_k$
 6:     $\boldsymbol{v}_k := \beta_{2k}\boldsymbol{v}_{k-1} + (1-\beta_{2k})\nabla f_{B_k}(\boldsymbol{\theta}_k) \odot \nabla f_{B_k}(\boldsymbol{\theta}_k)$
 7:     $\hat{\boldsymbol{v}}_k := (1-\beta_{2k}^{k+1})^{-1}\boldsymbol{v}_k$
 8:     $\mathsf{H}_k := \mathsf{diag}(\sqrt{\hat{v}_{k,i}})$
 9:     $\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \alpha_k \mathsf{H}_k^{-1}\hat{\boldsymbol{m}}_k$
10:     $k \leftarrow k+1$
11: end loop
```

3 Analysis of Adam for Nonconvex Optimization

This section provides our detailed theoretical results for Adam. Table 2 summarizes the results in Theorems 3, 4, 5, and 6.

3.1 Theoretical advantage of setting a small constant learning rate α and hyperparameters β_1 and β_2 close to 1

3.1.1 Constant learning rate rule

Let us consider Adam defined by Algorithm 1 using the following constant learning rate and hyperparameters:

$$\alpha_k := \alpha \in (0, +\infty),\ \beta_{1k} := \beta_1 \in (0,1),\ \text{and}\ \beta_{2k} := \beta_2 \in [0,1). \tag{15}$$

We assume the following conditions that were used in [14, Theorem 4.1]:

(A1) There exist positive numbers G and B such that, for all $k \in \mathbb{N}$, $\|\nabla f(\boldsymbol{\theta}_k)\| \leq G$ and $\|\nabla f_{B_k}(\boldsymbol{\theta}_k)\| \leq B$.

(A2) For all $\boldsymbol{\theta} \in \mathbb{R}^d$, there exists a positive number $D(\boldsymbol{\theta})$ such that, for all $k \in \mathbb{N}$, $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\| \leq D(\boldsymbol{\theta})$.

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then, F is Lipschitz continuous (i.e., $|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq G\|\boldsymbol{x} - \boldsymbol{y}\|$) if and only if $\|\nabla F(\boldsymbol{x})\| \leq G$ ($\boldsymbol{x} \in \mathbb{R}^d$) (see, e.g., Theorem 6.2.2, Corollary 6.1.2, and Exercise 6.1.9(c) in [2]). Let $\boldsymbol{\theta}^*$ be a local minimizer of a Lipschitz continuous function f. The continuity of f ensures that f is convex around $\boldsymbol{\theta}^*$. Hence, for any $\boldsymbol{\theta}$ belonging to a neighborhood $N(\boldsymbol{\theta}^*)$ of $\boldsymbol{\theta}^*$, $\|\nabla f(\boldsymbol{\theta})\| \leq G$. If the sequence $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}}$ generated by Adam approximates $\boldsymbol{\theta}^*$ (see, e.g., [31,37] for a guarantee of convergence of Adam), then $\boldsymbol{\theta}_k \in N(\boldsymbol{\theta}^*)$ for sufficiently large k, i.e., $\|\nabla f(\boldsymbol{\theta}_k)\| \leq G$, that is, (A1) holds.

Instead of assuming (A2), we may modify $\boldsymbol{\theta}_{k+1}$ in Algorithm 1 (step 9) by

$$\boldsymbol{\theta}_{k+1} = P_{C,\mathsf{H}_k}(\boldsymbol{\theta}_k - \alpha_k \mathsf{H}_k^{-1} \hat{\boldsymbol{m}}_k), \tag{16}$$

where $P_{C,H}$ is the projection onto a bounded, closed convex set $C \subset \mathbb{R}^d$ in the sense of the H-norm (see Section 2.1 for the definition of the H-norm). For example, we can set C to be a closed ball with center $\boldsymbol{c} \in \mathbb{R}^d$ and a sufficiently large radius $r > 0$. Accordingly, $P_{C,H}$ is defined by

$$P_{C,H}(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{c} + \frac{r}{\|\boldsymbol{\theta} - \boldsymbol{c}\|_H}(\boldsymbol{\theta} - \boldsymbol{c}) & (\boldsymbol{\theta} \notin C) \\ \boldsymbol{\theta} & (\boldsymbol{\theta} \in C). \end{cases}$$

The sequence $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}}$ generated by (16) satisfies the condition that $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}} \subset C$. Since C is bounded, $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}}$ is also bounded, that is, (A2) holds. Moreover, the sequence $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}} \subset C$ defined by (16) satisfies the condition that there exists $D > 0$ such that, for all $k \in \mathbb{N}$, $\|\boldsymbol{\theta}_k\| \leq D$. Hence, the continuity conditions of ∇f (see (S1)) and the norm imply that there exists $G > 0$ such that $\|\nabla f(\boldsymbol{\theta}_k)\| \leq G$, that is, (A1) holds. A discussion similar to the one showing the theorems in this paper, together with the nonexpansivity condition of the projection ($\|P_{C,H}(\boldsymbol{x}) - P_{C,H}(\boldsymbol{y})\|_H \leq \|\boldsymbol{x}-\boldsymbol{y}\|_H$), leads to versions of the theorems for all $\boldsymbol{\theta}$ belonging to C (see also Remark 1 in the Appendix).

The following is an analysis of Adam using a constant learning rate and hyperparameters (see Appendix A for the proof of Theorem 1).

Theorem 1 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (15) *satisfies that, for all $k \in \mathbb{N}$,*

$$\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{D(\boldsymbol{\theta}^\star)M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2}}_{C_1(\beta_1,b)} + \underbrace{\frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b}+G^2\right)}_{C_2(\alpha,\beta_1,\beta_2,b,k)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}$$

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \underbrace{\frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2}}_{C_1(\beta_1,b)} + \underbrace{\frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b}+G^2\right)}_{C_2(\alpha,\beta_1,\beta_2,b,k)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},$$

where $D(\boldsymbol{\theta})$, G, and B are defined as in (A1) and (A2), $\nabla f_{B_k}(\boldsymbol{\theta}_k) \odot \nabla f_{B_k}(\boldsymbol{\theta}_k) := (g_{k,i}^2) \in \mathbb{R}_+^d$, $M := \sup\{\max_{i\in[d]} g_{k,i}^2 \colon k \in \mathbb{N}\} < +\infty$, and $v_* := \inf\{\min_{i\in[d]} v_{k,i} \colon k \in \mathbb{N}\}$.

We would like to set constant parameters α, β_1, and β_2 so that the upper bounds of the performance measures $\mathbb{E}[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)]$ and $\mathbb{E}[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta})]$, denoted by C_i ($i = 1,2,3,4,5$) in Theorem 1, can be small. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be fixed. We can check that

$$C_3(\beta_1) := \frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G$$

is monotone decreasing for $\beta_1 \in (0,1)$. Moreover,

$$C_1(\beta_1,b) := \frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2},\ C_4(\beta_1,b) := (1-\beta_1)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right),$$
$$C_5(\beta_1,b) := \left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)$$

are monotone decreasing for $\beta_1 \in (0,1)$ and $b > 0$. Therefore, we should set β_1 close to 1 such as $\beta_1 = 0.9$ [14]. Let us consider

$$C_2(\alpha, \beta_1, \beta_2, b, k) := \frac{\alpha\sqrt{1-\beta_2^{k+1}}}{2\sqrt{v_*}\beta_1(1-\beta_1^{k+1})}\left(\frac{\sigma^2}{b} + G^2\right) \leq \frac{\alpha}{2\sqrt{v_*}\beta_1(1-\beta_1)}\left(\frac{\sigma^2}{b} + G^2\right) =: U_2. \tag{17}$$

C_2 is monotone decreasing for $\beta_2 \in [0,1)$ and $b > 0$. Although the function $f_k(\beta_1) := 1/(\beta_1(1-\beta_1^{k+1}))$ in C_2 is monotone increasing for β_1 satisfying $\beta_1^{k+1} > 1/(k+2)$, $f_k(\beta_1)$ with a sufficiently large number k is monotone decreasing for β_1. Moreover, $f_{\beta_1}(k) := 1/(\beta_1(1-\beta_1^{k+1}))$ is monotone decreasing for $k \in \mathbb{N}$. Although a function $f_{\beta_2}(k) := (1-\beta_2^{k+1})^{1/2}$ is monotone increasing for $k \in \mathbb{N}$, $f_{\beta_2}(k) := (1-\beta_2^{k+1})^{1/2} \leq 1$ is small for all β_2 and all $k \in \mathbb{N}$. C_2 is monotone increasing for α and monotone decreasing for β_2. Hence, we must set a small learning rate α and β_2 sufficiently close to 1, for example, $\alpha = 10^{-3}$ and $\beta_2 = 0.999$ [14], so that C_2 will be small when β_1 is close to 1. In a simplistic form, the upper bound U_2 of C_2 depends on $\alpha/(\beta_1(1-\beta_1))$, which is monotone increasing for $\beta_1 \geq 1/2$. Hence, using β_1 close to 1 (e.g., $\beta_1 = 0.9$) implies that U_2 is large. Accordingly, we must set a small learning rate α so that U_2 will be small when β_1 is close to 1.

Therefore, Theorem 1 indicates that using a small α and β_1 and β_2 close to 1 is useful to implement Adam defined by Algorithm 1, as shown by the previous numerical results [14,32]. Moreover, C_1, C_2, C_4, and C_5 are monotone decreasing with the batch size b. Hence, we should use a large batch size b to implement Adam defined by Algorithm 1, as shown by the previous numerical results [26].

Theorem 1 leads to the following corollary:

Corollary 1 *Under the assumptions in Theorem 1, Adam defined by Algorithm 1 with* (15) *satisfies*

$$\limsup_{k\to+\infty} \mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{D(\boldsymbol{\theta}^\star)M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b} + G^2}}_{C_1(\beta_1, b)} + \underbrace{\frac{\alpha}{2\sqrt{v_*}\beta_1}\left(\frac{\sigma^2}{b} + G^2\right)}_{\bar{C}_2(\alpha,\beta_1,b)} + \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right)}_{C_4(\beta_1,b)}$$

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\limsup_{k\to+\infty} \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \underbrace{\frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b} + G^2}}_{C_1(\beta_1, b)} + \underbrace{\frac{\alpha}{2\sqrt{v_*}\beta_1}\left(\frac{\sigma^2}{b} + G^2\right)}_{\bar{C}_2(\alpha,\beta_1,b)} + \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1} + 2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right)}_{C_5(\beta_1,b)},$$

where the parameters are defined as in Theorem 1.

The definitions of $C_1, \bar{C}_2, C_3, C_4$, and C_5 in Corollary 1 imply that using a small learning rate α, β_1 close to 1, and a large batch size b is advantageous for Adam.

3.1.2 Diminishing learning rate rule

Next, let us consider Adam defined by Algorithm 1 using the following diminishing learning rate α_k and hyperparameters β_{1k} and β_{2k} which converge to 1: for all $k \geq 1$,

$$\alpha_k := \frac{1}{k^a},\ \beta_{1k} := 1 - \frac{1}{k^{b_1}},\ \text{and}\ \beta_{2k} := \left(1 - \frac{1}{k^{b_2}}\right)^{\frac{1}{k+1}}, \tag{18}$$

where $\alpha_0 = 1$, $\beta_{10} = 0$, $\beta_{20} = 0$, and $a > 0, b_1 > 0$, and $b_2 > 0$ satisfy that

$$a - b_1 + \frac{b_2}{2} > 0.$$

The following is an analysis of Adam for a diminishing learning rate (see Appendix A for the proof of Theorem 2):

Theorem 2 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (18) *satisfies that, for all $k \geq 1$,*

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq\ & \underbrace{\frac{D(\boldsymbol{\theta}^\star)M^{\frac{1}{4}}k^{b_1}}{v_*^{\frac{1}{4}}(k^{b_1}-1)}\sqrt{\frac{\sigma^2}{b}+G^2}}_{D_1(b_1,b,k)} + \underbrace{\frac{1}{2\sqrt{v_*}(k^{b_1}-1)k^{a+\frac{b_2}{2}-2b_1}}\left(\frac{\sigma^2}{b}+G^2\right)}_{D_2(a,b_1,b_2,b,k)}\\ & + \underbrace{\frac{1}{k^{b_1}-1}D(\boldsymbol{\theta}^\star)G}_{D_3(b_1,k)} + \underbrace{\frac{1}{k^{b_1}}D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{D_4(b_1,b,k)}\end{aligned}$$

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $k \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq\ & \underbrace{\frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}k^{b_1}}{v_*^{\frac{1}{4}}(k^{b_1}-1)}\sqrt{\frac{\sigma^2}{b}+G^2}}_{D_1(b_1,b,k)} + \underbrace{\frac{1}{2\sqrt{v_*}(k^{b_1}-1)k^{a+\frac{b_2}{2}-2b_1}}\left(\frac{\sigma^2}{b}+G^2\right)}_{D_2(a,b_1,b_2,b,k)}\\ & + \underbrace{\frac{1}{k^{b_1}-1}D(\boldsymbol{\theta})G}_{D_3(b_1,k)} + \underbrace{\frac{1}{k^{b_1}}D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{D_4(b_1,b,k)}\\ & + \underbrace{\frac{{k^{b_1}}^2+2k^{b_1}-2}{k^{b_1}(k^{b_1}-1)}D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{D_5(b_1,b,k)},\end{aligned}$$

where the parameters are defined as in Theorem 1.

Theorem 2 leads to the following corollary:

Corollary 2 *Under the assumptions in Theorem 2, Adam defined by Algorithm 1 with* (18) *satisfies*

$$\limsup_{k \to +\infty} \mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \frac{D(\boldsymbol{\theta}^\star) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}} \sqrt{\frac{\sigma^2}{b} + G^2}$$

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $k \geq 1$,

$$\limsup_{k \to +\infty} \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}} \sqrt{\frac{\sigma^2}{b} + G^2} + D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2} \right),$$

where the parameters are defined as in Theorem 1.

Corollary 2 implies that the upper bounds of the performance measures of Adam using $(\alpha_k)_{k \in \mathbb{N}}$ converging to 0 and $(\beta_{1k})_{k \in \mathbb{N}}$ and $(\beta_{2k})_{k \in \mathbb{N}}$ converging to 1 are smaller than the ones of Adam using constant parameters α, β_1, and β_2 shown in Corollary 1.

3.2 Adam with Condition (9)

As described in Theorems 1 and 2 and Corollaries 1 and 2, although a small learning rate α_k and hyperparameters β_1 and β_1 close to 1 are useful for implementing Adam, the upper bounds of the performance measures (4) and (6) in Theorems 1 and 2 and Corollaries 1 and 2 would not be small. This result can be attributed to that, in general, Adam does not converge to a local minimizer of the problem of minimizing f over $\mathbb{R}^d$. In fact, there is a stochastic convex optimization problem in which Adam using $\beta_1 < \sqrt{\beta_2}$ (e.g., $\beta_1 = 0.9$ and $\beta_2 = 0.999$) does not converge to the optimal solution [22, Theorem 3].

Theorem 4 in [22] showed that Adam with (9) and (10) can solve convex stochastic optimization problems (see also [34, Theorems 2.1 and 2.2] for the analyses of AdaBelief for convex and nonconvex optimization). Motivated by the results in [22, 34], we will analyze Adam with (9).

3.2.1 Constant learning rate rule

The following is an analysis of Adam using (9), a constant learning rate, and constant hyperparameters (see Appendix A for the proof of Theorem 3).

Theorem 3 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (9) *and* (15) *satisfies that, for* *all* $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}(1-\beta_1^{K+1})}{2\alpha\beta_1\sqrt{1-\beta_2}K}}_{\bar{C}_1(\alpha,\beta_1,\beta_2,K)} + \underbrace{\frac{\alpha}{2\sqrt{v_*}\beta_1(1-\beta_1)}\left(\frac{\sigma^2}{b}+G^2\right)}_{\hat{C}_2(\alpha,\beta_1,b)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ *and all* $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}(1-\beta_1^{K+1})}{2\alpha\beta_1\sqrt{1-\beta_2}K}}_{\bar{C}_1(\alpha,\beta_1,\beta_2,K)} + \underbrace{\frac{\alpha}{2\sqrt{v_*}\beta_1(1-\beta_1)}\left(\frac{\sigma^2}{b}+G^2\right)}_{\hat{C}_2(\alpha,\beta_1,b)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},$$

where the parameters are defined as in Theorem 1 and $\tilde{D}(\boldsymbol{\theta}) := \sup\{\max_{i\in[d]}(\theta_{k,i}-\theta_i)^2 \colon k \in \mathbb{N}\} < +\infty$.

Let us compare Theorem 1 with Theorem 3. A significant difference is

$$C_1(\beta_1,b) := \frac{D(\boldsymbol{\theta}^\star)M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_1}\sqrt{\frac{\sigma^2}{b}+G^2} \text{ and } \bar{C}_1(\alpha,\beta_1,\beta_2,K) := \frac{dD(\boldsymbol{\theta}^\star)\sqrt{M}(1-\beta_1^{K+1})}{2\alpha\beta_1\sqrt{1-\beta_2}K}.$$

Although $\bar{C}_1$ in Theorem 3 is monotone increasing for $\beta_2 \in [0,1)$ and $1/\alpha$, $\bar{C}_1$ becomes small when K is sufficiently large. C_1 in Theorem 1 does not change for any K. $\hat{C}_2$ in Theorem 3 defined by

$$\hat{C}_2(\alpha,\beta_1,b) := \frac{\alpha}{2\sqrt{v_*}\beta_1(1-\beta_1)}\left(\frac{\sigma^2}{b}+G^2\right)$$

is monotone increasing for $\beta_1 \geq 1/2$ (see also the discussion in (17)). Hence, we must set a small learning rate α to make $\hat{C}_2$ small when β_1 is close to 1. From the definition of $\bar{C}_1$, the use of a small learning rate α entails a large number of steps K. Therefore, Theorem 3 indicates that, if K is sufficiently large, then Adam under Condition (9) has a tighter upper bound of $(1/K)\sum_{k=1}^{K}\mathbb{E}[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)]$ than Adam without Condition (9). Theorem 1 implies that there exist C_i ($i=1,2,3$) such that, for all $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)\right] = O\left(\frac{1}{K}\right) + C_1\alpha + C_2\frac{\alpha}{b} + C_3\frac{1-\beta_1}{\beta_1}. \tag{19}$$

Here, we compare Theorem 1 in [31] with Theorem 3 in this paper. Theorem 1 and the proof of Theorem 1 in [31] show that, under the condition that ∇f is Lipschitz continuous with the Lipschitz constant L, Adam using $\alpha = O(1/L)$ and $\beta_1 = 0$ satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_k)\|^2\right] \leq 2\left(\sqrt{\beta_2}G+\varepsilon\right)\left\{\frac{f(\boldsymbol{\theta}_1)-f(\boldsymbol{\theta}^\star)}{\alpha K}+\left(\frac{G\sqrt{1-\beta_2}}{\varepsilon^2}+\frac{L\alpha}{2\varepsilon^2}\right)\frac{\sigma^2}{b}\right\},$$

that is,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_k)\|^2\right] = O\left(\frac{1}{K}\right)+C_4\frac{\alpha}{b}+C_5\left(\sqrt{\beta_2}+1\right)\sqrt{1-\beta_2}, \qquad (20)$$

where $\varepsilon > 0$ is the precision for nonconvex optimization, and C_4 and C_5 are constants. Theorem 1 in [31] assumes that f is Lipschitz smooth, while Theorem 3 in this paper does not assume so. Hence, there is a difference in the performance measure between (19) and (20). However, (19) and (20) show that using a small learning rate α, hyperparameters close to 1, and a large batch size b is advantageous for Adam.

Let us compare the results in [24] with Theorem 3. Theorem 12 in [24] indicates that SGD using a constant learning rate $\alpha = 1/L$, where $L > 0$ is the Lipschitz constant of the Lipschitz continuous gradient ∇f, almost surely satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla f(\boldsymbol{\theta}_k)\|^2 \leq \frac{2L(f(\boldsymbol{\theta}_0)-f^\star)}{K}+\sigma^2,$$

where $f^\star := \min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta})$. Theorem 3 indicates that Adam using a constant learning rate α satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right] \leq \frac{dD(\boldsymbol{\theta})\sqrt{M}(1-\beta_1^{K+1})}{2\alpha\beta_1\sqrt{1-\beta_2}K}+\hat{C}_2(\alpha,\beta_1,b)+C_3(\beta_1)+C_5(\beta_1,b)$$

without assuming that ∇f is Lipschitz continuous. This result indicates that using a small learning rate α, a hyperparameter β_1 close to 1, and a large batch size b is advantageous for Adam.

$\hat{C}_2$ in Theorem 3 is obtained under conditions such that $\alpha_k = \alpha$ and $\beta_{1k} = \beta_1$ and

$$\frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}K}\sum_{k=1}^{K}\frac{\alpha_k\sqrt{1-\beta_{2k}^{k+1}}}{\beta_{1k}(1-\beta_{1k}^{k+1})} \leq \frac{\alpha(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1(1-\beta_1)K}\sum_{k=1}^{K}\sqrt{1-\beta_{2k}^{k+1}} \leq \underbrace{\frac{\alpha(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1(1-\beta_1)}}_{\hat{C}_2(\alpha,\beta_1,b)},$$

where the first inequality comes from $1-\beta_1 \leq 1-\beta_1^{k+1}$ ($k \in \mathbb{N}$) and the second inequality comes from $1-\beta_{2k} \leq 1$ ($k \in \mathbb{N}$). Hence, we can improve $\hat{C}_2$ by using the property of β_{2k}. Let us consider Adam defined by Algorithm 1 using the following constant learning rate and hyperparameters:

$$\alpha_k := \alpha \in (0,+\infty),\ \beta_{1k} := \beta_1 \in (0,1),\ \text{and}\ \beta_{2k} := \left(1-\frac{1}{k^{b_2}}\right)^{\frac{1}{k+1}}, \qquad (21)$$

where $\beta_{20} = 0$ and $b_2 \in (0,2)$.

Theorem 4 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (9) *and* (21) *satisfies that, for all $K \geq 1$,*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\alpha\beta_1 K^{1-\frac{b_2}{2}}}}_{\hat{C}_1(\alpha,\beta_1,\beta_2,K)} + \underbrace{\frac{\alpha}{\sqrt{v_*}\beta_1(1-\beta_1)K^{\frac{b_2}{2}}}\left(\frac{\sigma^2}{b}+G^2\right)}_{\tilde{C}_2(\alpha,\beta_1,b,K)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}}{2\alpha\beta_1 K^{1-\frac{b_2}{2}}}}_{\hat{C}_1(\alpha,\beta_1,\beta_2,K)} + \underbrace{\frac{\alpha}{\sqrt{v_*}\beta_1(1-\beta_1)K^{\frac{b_2}{2}}}\left(\frac{\sigma^2}{b}+G^2\right)}_{\tilde{C}_2(\alpha,\beta_1,b,K)}$$
$$+\underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},$$

where the parameters are defined as in Theorem 3.

The definitions of $\hat{C}_1$ and $\tilde{C}_2$ imply that the best setting of b_2 is 1, since

$$0 < \min\left\{1-\frac{b_2}{2}, \frac{b_2}{2}\right\}$$

attains the maximum value $1/2$ when $b_2 = 1$. Hence, Adam using (9) and (21) with $b_2 = 1$ satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] = O\left(\frac{1}{\sqrt{K}} + \frac{1-\beta_1}{\beta_1}\right).$$

3.2.2 Diminishing learning rate rule

Let us consider Adam defined by Algorithm 1 under Condition (9) and the following diminishing learning rate α_k and constant hyperparameters β_1 and β_2: for all $k \geq 1$,

$$\alpha_k := \frac{1}{k^a},\ \beta_{1k} := \beta_1 \in (0,1),\ \text{and}\ \beta_{2k} := \beta_2 \in [0,1), \tag{22}$$

where $\alpha_0 = 1$ and $a \in (0,1)$.

Theorem 5 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (9) *and* (22) *satisfies that, for all $K \geq 1$,*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1\sqrt{1-\beta_2}K^{1-a}}}_{\bar{D}_1(\beta_1,\beta_2,K)} + \underbrace{\frac{1}{\sqrt{v_*}\beta_1(1-\beta_1)K^a}\left(\frac{\sigma^2}{b}+G^2\right)}_{\bar{D}_2(\beta_1,b,K)}$$
$$+ \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta})\right] \leq \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}}{2\beta_1\sqrt{1-\beta_2}K^{1-a}}}_{\bar{D}_1(\beta_1,\beta_2,K)} + \underbrace{\frac{1}{\sqrt{v_*}\beta_1(1-\beta_1)K^a}\left(\frac{\sigma^2}{b}+G^2\right)}_{\bar{D}_2(\beta_1,b,K)}$$
$$+ \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},$$

where the parameters are defined as in Theorem 3.

The definitions of $\bar{D}_1$ and $\bar{D}_2$ imply that the best setting of a is $1/2$ since

$$\min\{1-a, a\}$$

attains the maximum value $1/2$ when $a = 1/2$. Hence, Adam using (9) and (22) with $a = 1/2$ satisfies

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\boldsymbol{m}_k^\top (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] = O\left(\frac{1}{\sqrt{K}} + \frac{1-\beta_1}{\beta_1}\right).$$

Motivated by (21), we consider Adam defined by Algorithm 1 using the following constant learning rate and hyperparameters:

$$\alpha_k := \frac{1}{k^a},\ \beta_{1k} := \beta_1 \in (0,1),\ \text{and}\ \beta_{2k} := \left(1 - \frac{1}{k^{b_2}}\right)^{\frac{1}{k+1}}, \tag{23}$$

where $\alpha_0 = 0$, $\beta_{20} = 0$, and $a \in (0,1)$ and $b_2 \in (0,2)$ satisfy

$$1 - a - \frac{b_2}{2} > 0.$$

Theorem 6 *Suppose that (S1)–(S3) and (A1)–(A2) hold and $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is a stationary point of f. Then, Adam defined by Algorithm 1 using* (9) *and* (23) *satisfies that, for all $K \geq 1$,*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)\right] \leq{} & \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1 K^{1-(a+\frac{b_2}{2})}}}_{\hat{D}_1(\beta_1,\beta_2,K)} + \underbrace{\frac{(\sigma^2 b^{-1}+G^2)}{\sqrt{v_*}\beta_1(1-\beta_1)K^{a+\frac{b_2}{2}}}}_{\hat{D}_2(\beta_1,b,K)} \\ & + \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta}^\star)G}_{C_3(\beta_1)} + \underbrace{(1-\beta_1)D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_4(\beta_1,b)}\end{aligned}$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and all $K \geq 1$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right] \leq{} & \underbrace{\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}}{2\beta_1 K^{1-(a+\frac{b_2}{2})}}}_{\hat{D}_1(\beta_1,\beta_2,K)} + \underbrace{\frac{(\sigma^2 b^{-1}+G^2)}{\sqrt{v_*}\beta_1(1-\beta_1)K^{a+\frac{b_2}{2}}}}_{\hat{D}_2(\beta_1,b,K)} \\ & + \underbrace{\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G}_{C_3(\beta_1)} + \underbrace{\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)}_{C_5(\beta_1,b)},\end{aligned}$$

where the parameters are defined as in Theorem 3.

The definitions of $\hat{D}_1$ and $\hat{D}_2$ imply that

$$\min\left\{1-\left(a+\frac{b_2}{2}\right), a+\frac{b_2}{2}\right\}$$

attains the maximum value $1/2$ when $a+b_2/2 = 1/2$, e.g., $a = 1/4$ and $b_2 = 1/2$. Hence, Adam using (9) and (23) with $a = 1/2$ satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\boldsymbol{m}_k^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)\right] = O\left(\frac{1}{\sqrt{K}}+\frac{1-\beta_1}{\beta_1}\right).$$

4 Conclusion and Future Work

This paper presented theoretical analyses of the Adam optimizer without assuming the Lipschitz smoothness condition for nonconvex optimization in deep learning. The analyzes indicated that Adam performs well when it uses hyperparameters close to one and not only a small learning rate but also a diminishing learning rate. Hence, our results are theoretical evidence supporting numerical evaluations showing that small constant learning rates and hyperparameters close to one are advantageous for training deep neural networks.

This paper focused on convergence analyses of Adam for nonconvex optimization. In the future, we should consider developing convergence analyses of Adam's variants for nonconvex optimization and show theoretically that adaptive methods, such as Yogi, AMSGrad, AdaBelief, Padam, and AdamW, using hyperparameters close to one perform well.

A Appendix A

Unless stated otherwise, all relationships between random variables are supported to hold almost surely.

A.1 Lemmas

Lemma 1 *Suppose that (S1), (S2)*(13)*, and (S3) hold. Then, Adam defined by Algorithm 1 satisfies the following: for all $k \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right] &= \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right] + \alpha_k^2 \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] \\ &\quad + 2\alpha_k \left\{ \frac{\beta_{1k}}{\tilde{\beta}_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \boldsymbol{m}_{k-1}\right] + \frac{\hat{\beta}_{1k}}{\tilde{\beta}_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f(\boldsymbol{\theta}_k)\right] \right\},\end{aligned}$$

where $\mathsf{d}_k := -\mathsf{H}_k^{-1}\hat{\boldsymbol{m}}_k$*,* $\hat{\beta}_{1k} := 1 - \beta_{1k}$*, and* $\tilde{\beta}_{1k} := 1 - \beta_{1k}^{k+1}$*.*

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and $k \in \mathbb{N}$. The definition of $\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k + \alpha_k \mathsf{d}_k$ implies that

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 = \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 + 2\alpha_k \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}, \mathsf{d}_k \rangle_{\mathsf{H}_k} + \alpha_k^2 \|\mathsf{d}_k\|_{\mathsf{H}_k}^2.$$

Moreover, the definitions of d_k, $\boldsymbol{m}_k$, and $\hat{\boldsymbol{m}}_k$ ensure that

$$\begin{aligned}\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}, \mathsf{d}_k \rangle_{\mathsf{H}_k} = \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}, \mathsf{H}_k \mathsf{d}_k \rangle = \langle \boldsymbol{\theta} - \boldsymbol{\theta}_k, \hat{\boldsymbol{m}}_k \rangle &= \frac{1}{\tilde{\beta}_{1k}} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \boldsymbol{m}_k \\ &= \frac{\beta_{1k}}{\tilde{\beta}_{1k}} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \boldsymbol{m}_{k-1} + \frac{\hat{\beta}_{1k}}{\tilde{\beta}_{1k}} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f_{B_k}(\boldsymbol{\theta}_k).\end{aligned}$$

Hence,

$$\begin{aligned}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 &= \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 + \alpha_k^2 \|\mathsf{d}_k\|_{\mathsf{H}_k}^2 \\ &\quad + 2\alpha_k \left\{ \frac{\beta_{1k}}{\tilde{\beta}_{1k}} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \boldsymbol{m}_{k-1} + \frac{\hat{\beta}_{1k}}{\tilde{\beta}_{1k}} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f_{B_k}(\boldsymbol{\theta}_k) \right\}.\end{aligned} \tag{24}$$

Conditions (13) and (S3) guarantee that

$$\mathbb{E}\left[\mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f_{B_k}(\boldsymbol{\theta}_k) \middle| \boldsymbol{\theta}_k\right]\right] = \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \mathbb{E}\left[\nabla f_{B_k}(\boldsymbol{\theta}_k) \middle| \boldsymbol{\theta}_k\right]\right] = \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f(\boldsymbol{\theta}_k)\right].$$

Therefore, the lemma follows by taking the expectation on both sides of (24). This completes the proof. □

Remark 1 Let us consider (16); that is,

$$\boldsymbol{\theta}_{k+1} = P_{C,\mathsf{H}_k}(\boldsymbol{\theta}_k + \alpha_k \mathsf{d}_k).$$

Let $k \in \mathbb{N}$ and $\boldsymbol{\theta} \in C$ (i.e., $\boldsymbol{\theta} = P_{C,\mathsf{H}_k}(\boldsymbol{\theta})$). The nonexpansivity condition of P_{C,H_k} ensures that

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k} = \|P_{C,\mathsf{H}_k}(\boldsymbol{\theta}_k + \alpha_k \mathsf{d}_k) - P_{C,\mathsf{H}_k}(\boldsymbol{\theta})\|_{\mathsf{H}_k} \leq \|(\boldsymbol{\theta}_k + \alpha_k \mathsf{d}_k) - \boldsymbol{\theta}\|_{\mathsf{H}_k}.$$

Hence, we have that

$$\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 \leq \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2 + 2\alpha_k \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}, \mathsf{d}_k \rangle_{\mathsf{H}_k} + \alpha_k^2 \|\mathsf{d}_k\|_{\mathsf{H}_k}^2.$$

Accordingly, a discussion similar to the one showing Lemma 1 ensures that, for all $\boldsymbol{\theta} \in C$ and all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right] &\leq \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right] + \alpha_k^2 \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] \\ &\quad + 2\alpha_k \left\{ \frac{\beta_{1k}}{\tilde{\beta}_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \boldsymbol{m}_{k-1}\right] + \frac{\hat{\beta}_{1k}}{\tilde{\beta}_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f(\boldsymbol{\theta}_k)\right] \right\}.\end{aligned} \tag{25}$$

We may assume without loss of generality that the assertion in Lemma 1 holds for all $\boldsymbol{\theta} \in C$ and all $k \in \mathbb{N}$, since the theorems in this paper evaluate the upper bounds of (4), (5), (6), and (7). A discussion similar to the one showing the theorems in this paper (see the following lemmas and the proof of theorems) leads to versions of the theorems for all $\boldsymbol{\theta}$ belonging to C; that is, the sequence $(\boldsymbol{\theta}_k)_{k \in \mathbb{N}}$ generated by (16) satisfies the assertions in the theorems for all $\boldsymbol{\theta} \in C$.

Lemma 2 *Adam defined by Algorithm 1 satisfies that, under (S2)*(13), (14)*, and (A1), for all* $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{m}_k\|^2\right] \leq \frac{\sigma^2}{b} + G^2, \quad \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] \leq \frac{\sqrt{\tilde{\beta}_{2k}}}{\tilde{\beta}_{1k}^2 \sqrt{v_*}} \left(\frac{\sigma^2}{b} + G^2 \right),$$

where $v_* := \inf\{\min_{i \in [d]} v_{k,i} \colon k \in \mathbb{N}\}$, $\tilde{\beta}_{1k} := 1 - \beta_{1k}^{k+1}$*, and* $\tilde{\beta}_{2k} := 1 - \beta_{2k}^{k+1}$.

Proof Assumption (S2)(13) implies that

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k)\right\|^2 \middle| \boldsymbol{\theta}_k\right] &= \mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k) + \nabla f(\boldsymbol{\theta}_k)\right\|^2 \middle| \boldsymbol{\theta}_k\right] \\ &= \mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k)\right\|^2 \middle| \boldsymbol{\theta}_k\right] + \mathbb{E}\left[\left\|\nabla f(\boldsymbol{\theta}_k)\right\|^2 \middle| \boldsymbol{\theta}_k\right] \\ &\quad + 2\mathbb{E}\left[(\nabla f_{B_k}(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k))^\top \nabla f(\boldsymbol{\theta}_k) \middle| \boldsymbol{\theta}_k\right] \\ &= \mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k)\right\|^2 \middle| \boldsymbol{\theta}_k\right] + \|\nabla f(\boldsymbol{\theta}_k)\|^2,\end{aligned} \tag{26}$$

which, together with (S2)(14) and (A1), implies that

$$\mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k)\right\|^2\right] \leq \frac{\sigma^2}{b} + G^2. \tag{27}$$

The convexity of $\|\cdot\|^2$, together with the definition of $\boldsymbol{m}_k$ and (27), guarantees that, for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{m}_k\|^2\right] &\leq \beta_{1k} \mathbb{E}\left[\|\boldsymbol{m}_{k-1}\|^2\right] + \hat{\beta}_{1k} \mathbb{E}\left[\left\|\nabla f_{B_k}(\boldsymbol{\theta}_k)\right\|^2\right] \\ &\leq \beta_{1k} \mathbb{E}\left[\|\boldsymbol{m}_{k-1}\|^2\right] + \hat{\beta}_{1k} \left(\frac{\sigma^2}{b} + G^2 \right).\end{aligned}$$

Induction thus ensures that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{m}_k\|^2\right] \leq \max\left\{ \|\boldsymbol{m}_{-1}\|^2, \frac{\sigma^2}{b} + G^2 \right\} = \frac{\sigma^2}{b} + G^2, \tag{28}$$

where $\boldsymbol{m}_{-1} = \mathbf{0}$. For $k \in \mathbb{N}$, $\mathsf{H}_k \in \mathbb{S}_{++}^d$ guarantees the existence of a unique matrix $\overline{\mathsf{H}}_k \in \mathbb{S}_{++}^d$ such that $\mathsf{H}_k = \overline{\mathsf{H}}_k^2$ [12, Theorem 7.2.6]. We have that, for all $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_{\mathsf{H}_k}^2 = \|\overline{\mathsf{H}}_k \boldsymbol{x}\|^2$. Accordingly, the definitions of d_k and $\hat{\boldsymbol{m}}_k$ imply that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] = \mathbb{E}\left[\left\|\overline{\mathsf{H}}_k^{-1} \mathsf{H}_k \mathsf{d}_k\right\|^2\right] \leq \frac{1}{\tilde{\beta}_{1k}^2} \mathbb{E}\left[\left\|\overline{\mathsf{H}}_k^{-1}\right\|^2 \|\boldsymbol{m}_k\|^2\right],$$

where

$$\left\|\overline{\mathsf{H}}_k^{-1}\right\| = \left\|\mathsf{diag}\left(\hat{v}_{k,i}^{-\frac{1}{4}}\right)\right\| = \max_{i\in[d]} \hat{v}_{k,i}^{-\frac{1}{4}} = \max_{i\in[d]} \left(\frac{v_{k,i}}{\tilde{\beta}_{2k}}\right)^{-\frac{1}{4}} =: \left(\frac{v_{k,i^*}}{\tilde{\beta}_{2k}}\right)^{-\frac{1}{4}}.$$

Moreover, the definition of

$$v_* := \inf\left\{v_{k,i^*} : k \in \mathbb{N}\right\}$$

and (28) imply that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] \leq \frac{\tilde{\beta}_{2k}^{\frac{1}{2}}}{\tilde{\beta}_{1k}^2 v_*^{\frac{1}{2}}} \left(\frac{\sigma^2}{b} + G^2\right),$$

completing the proof. □

Lemma 3 *Suppose that (S1)–(S3) and (A1)–(A2) hold. Then, Adam defined by Algorithm 1 satisfies the following: for all $k \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] \leq \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}},$$

where $\nabla f_{B_k}(\boldsymbol{\theta}_k) \odot \nabla f_{B_k}(\boldsymbol{\theta}_k) := (g_{k,i}^2) \in \mathbb{R}_+^d$, $M := \sup\{\max_{i\in[d]} g_{k,i}^2 : k \in \mathbb{N}\} < +\infty$, $\hat{\beta}_{1k} := 1 - \beta_{1k}$, $\tilde{\beta}_{1k} := 1 - \beta_{1k}^{k+1}$, $\tilde{\beta}_{2k} := 1 - \beta_{2k}^{k+1}$, v_* is defined as in Lemma 2, and $D(\boldsymbol{\theta})$ and G are defined as in Assumptions (A1) and (A2).

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$. Lemma 1 guarantees that for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] = &\underbrace{\frac{\tilde{\beta}_{1k}}{2\alpha_k \beta_{1k}} \left\{\mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right] - \mathbb{E}\left[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}^2\right]\right\}}_{a_k} + \underbrace{\frac{\alpha_k \tilde{\beta}_{1k}}{2\beta_{1k}} \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right]}_{b_k} \\ &+ \underbrace{\frac{\hat{\beta}_{1k}}{\beta_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f(\boldsymbol{\theta}_k)\right]}_{c_k}.\end{aligned} \tag{29}$$

The triangle inequality and the definition of $\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k + \alpha_k \mathsf{d}_k$ ensure that

$$\begin{aligned} a_k &= \frac{\tilde{\beta}_{1k}}{2\alpha_k \beta_{1k}} \mathbb{E}\left[\left(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k} + \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}\right)\left(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k} - \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}\right)\right] \\ &\leq \frac{\tilde{\beta}_{1k}}{2\alpha_k \beta_{1k}} \mathbb{E}\left[\left(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k} + \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}\right)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|_{\mathsf{H}_k}\right] \\ &= \frac{\tilde{\beta}_{1k}}{2\beta_{1k}} \mathbb{E}\left[\left(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k} + \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k}\right)\|\mathsf{d}_k\|_{\mathsf{H}_k}\right]. \end{aligned} \tag{30}$$

Let $\nabla f_{B_k}(\boldsymbol{\theta}_k) \odot \nabla f_{B_k}(\boldsymbol{\theta}_k) := (g_{k,i}^2) \in \mathbb{R}_+^d$. Assumption (A1) ensures that there exists $M \in \mathbb{R}$ such that, for all $k \in \mathbb{N}$, $\max_{i\in[d]} g_{k,i}^2 \leq M$. The definition of $\boldsymbol{v}_k$ guarantees that, for all $i \in [d]$ and all $k \in \mathbb{N}$,

$$v_{k,i} = \beta_{2k} v_{k-1,i} + \hat{\beta}_{2k} g_{k,i}^2.$$

Induction thus ensures that, for all $i \in [d]$ and all $k \in \mathbb{N}$,

$$v_{k,i} \leq \max\{v_{0,i}, M\} = M,$$

where $\boldsymbol{v}_0 = (v_{0,i}) = \mathbf{0}$. From the definition of $\hat{\boldsymbol{v}}_k$, we have that, for all $i \in [d]$ and all $k \in \mathbb{N}$,

$$\hat{v}_{k,i} = \frac{v_{k,i}}{\tilde{\beta}_{2k}} \leq \frac{M}{\tilde{\beta}_{2k}}, \tag{31}$$

which implies that

$$\|\overline{\mathsf{H}}_k\| = \left\|\mathsf{diag}\left(\hat{v}_{k,i}^{\frac{1}{4}}\right)\right\| = \max_{i \in [d]} \hat{v}_{k,i}^{\frac{1}{4}} \leq \left(\frac{M}{\tilde{\beta}_{2k}}\right)^{\frac{1}{4}}.$$

Hence, (A2) implies that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\mathsf{H}_k} &= \|\overline{\mathsf{H}}_k(\boldsymbol{\theta}_k - \boldsymbol{\theta})\| \leq \|\overline{\mathsf{H}}_k\| \, \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\| \leq D(\boldsymbol{\theta}) \left(\frac{M}{\tilde{\beta}_{2k}}\right)^{\frac{1}{4}}, \\
\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\|_{\mathsf{H}_k} &= \|\overline{\mathsf{H}}_k(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta})\| \leq \|\overline{\mathsf{H}}_k\| \, \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}\| \leq D(\boldsymbol{\theta}) \left(\frac{M}{\tilde{\beta}_{2k}}\right)^{\frac{1}{4}}.
\end{aligned}$$

Lemma 2, Jensen's inequality, and (30) ensure that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
a_k &\leq \frac{\tilde{\beta}_{1k}}{2\beta_{1k}} 2D(\boldsymbol{\theta}) \left(\frac{M}{\tilde{\beta}_{2k}}\right)^{\frac{1}{4}} \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}\right] \leq \frac{\tilde{\beta}_{1k}}{\beta_{1k}} D(\boldsymbol{\theta}) \frac{M^{\frac{1}{4}}}{\tilde{\beta}_{2k}^{\frac{1}{4}}} \frac{\tilde{\beta}_{2k}^{\frac{1}{4}}}{\tilde{\beta}_{1k} v_*^{\frac{1}{4}}} \sqrt{\frac{\sigma^2}{b} + G^2} \\
&= \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2}.
\end{aligned} \tag{32}$$

Lemma 2 guarantees that, for all $k \in \mathbb{N}$,

$$b_k = \frac{\alpha_k \tilde{\beta}_{1k}}{2\beta_{1k}} \mathbb{E}\left[\|\mathsf{d}_k\|_{\mathsf{H}_k}^2\right] \leq \frac{\alpha_k \tilde{\beta}_{1k}}{2\beta_{1k}} \frac{\sqrt{\tilde{\beta}_{2k}}}{\tilde{\beta}_{1k}^2 \sqrt{v_*}} \left(\frac{\sigma^2}{b} + G^2\right) = \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right). \tag{33}$$

The Cauchy–Schwarz inequality and Assumption (A2) imply that, for all $k \in \mathbb{N}$,

$$c_k = \frac{\hat{\beta}_{1k}}{\beta_{1k}} \mathbb{E}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}}. \tag{34}$$

Therefore, (29), (32), (33), and (34) ensure that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] \leq \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}},$$

which completes the proof. □

Lemma 4 *Suppose that (S1)–(S3) and (A1)–(A2) hold. Then, Adam defined by Algorithm 1 satisfies the following: for all* $k \in \mathbb{N}$ *and all* $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} \\
&+ \hat{\beta}_{1k} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),
\end{aligned}$$

where the parameters are defined as in Lemma 3.

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and $k \in \mathbb{N}$. The definition of $\boldsymbol{m}_k$ implies that

$$\begin{aligned}(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k &= (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1} + (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\boldsymbol{m}_k - \boldsymbol{m}_{k-1}) \\ &= (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1} + \hat{\beta}_{1k} (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\nabla f_{B_k}(\boldsymbol{\theta}_k) - \boldsymbol{m}_{k-1}),\end{aligned}$$

which, together with the Cauchy–Schwarz inequality, the triangle inequality, and Assumptions (A1) and (A2), implies that

$$\begin{aligned}(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k &\leq (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1} + \hat{\beta}_{1k} D(\boldsymbol{\theta}) \|\nabla f_{B_k}(\boldsymbol{\theta}_k) - \boldsymbol{m}_{k-1}\| \\ &\leq (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1} + \hat{\beta}_{1k} D(\boldsymbol{\theta}) (B + \|\boldsymbol{m}_{k-1}\|).\end{aligned}$$

Lemma 2 and Jensen's inequality guarantee that

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] + \hat{\beta}_{1k} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right). \tag{35}$$

Hence, Lemma 3 implies that

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*} \beta_{1k} \tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} \\ &+ \hat{\beta}_{1k} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),\end{aligned}$$

which completes the proof. □

Lemma 5 *Suppose that (S1)–(S3) and (A1)–(A2) hold. Then, Adam defined by Algorithm 1 satisfies the following: for all $k \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*} \beta_{1k} \tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} \\ &+ D(\boldsymbol{\theta}) \left(\frac{1}{\beta_{1k}} + 2\hat{\beta}_{1k}\right) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),\end{aligned}$$

where the parameters are defined as in Lemma 3.

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and $k \in \mathbb{N}$. The definition of $\boldsymbol{m}_k$ ensures that

$$\begin{aligned}&(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f_{B_k}(\boldsymbol{\theta}_k) \\ &= (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k + (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\nabla f_{B_k}(\boldsymbol{\theta}_k) - \boldsymbol{m}_{k-1}) + (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\boldsymbol{m}_{k-1} - \boldsymbol{m}_k) \\ &= (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k + \frac{1}{\beta_{1k}} (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\nabla f_{B_k}(\boldsymbol{\theta}_k) - \boldsymbol{m}_k) + \hat{\beta}_{1k} (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top (\boldsymbol{m}_{k-1} - \nabla f_{B_k}(\boldsymbol{\theta}_k)),\end{aligned}$$

which, together with the Cauchy–Schwarz inequality, the triangle inequality, and Assumptions (A1) and (A2), implies that

$$(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f_{B_k}(\boldsymbol{\theta}_k) \leq (\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k + \frac{1}{\beta_{1k}} D(\boldsymbol{\theta}) (B + \|\boldsymbol{m}_k\|) + \hat{\beta}_{1k} D(\boldsymbol{\theta}) (B + \|\boldsymbol{m}_{k-1}\|).$$

Lemma 2 and Jensen's inequality guarantee that

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] + \left(\frac{1}{\beta_{1k}} + \hat{\beta}_{1k}\right) D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right), \tag{36}$$

which, together with Lemma 4, implies that

$$\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq \frac{D(\boldsymbol{\theta})M^{\frac{1}{4}}}{v_*^{\frac{1}{4}}\beta_{1k}}\sqrt{\frac{\sigma^2}{b}+G^2}+\frac{\alpha_k\sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}}\left(\frac{\sigma^2}{b}+G^2\right)+D(\boldsymbol{\theta})G\frac{\hat{\beta}_{1k}}{\beta_{1k}}$$
$$+D(\boldsymbol{\theta})\left(\frac{1}{\beta_{1k}}+2\hat{\beta}_{1k}\right)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right),$$

which completes the proof. □

Lemma 6 *Suppose that (S1)–(S3) and (A1)–(A2) hold, $\beta_{1k} := \beta_1 \in (0,1)$, and $(\alpha_k)_{k\in\mathbb{N}}$ is monotone decreasing. Then, Adam defined by Algorithm 1 with* (9) *satisfies the following: for all $K \geq 1$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] \leq \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha_K\sqrt{\tilde{\beta}_{2K}}K}+\frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\alpha_k\sqrt{\tilde{\beta}_{2k}}+D(\boldsymbol{\theta})G\frac{\hat{\beta}_1}{\beta_1},$$

where the parameters are defined as in Lemma 3 and $\tilde{D}(\boldsymbol{\theta}) := \sup\{\max_{i\in[d]}(\theta_{k,i}-\theta_i)^2 \colon k\in\mathbb{N}\} < +\infty$.

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$ and

$$\gamma_k := \frac{\tilde{\beta}_{1k}}{2\beta_1\alpha_k}$$

for all $k\in\mathbb{N}$. Since $(\alpha_k)_{k\in\mathbb{N}}$ is monotone decreasing and $\tilde{\beta}_{1k} = 1-\beta_1^{k+1} \leq 1-\beta_1^{k+2} = \tilde{\beta}_{1,k+1}$, $(\gamma_k)_{k\in\mathbb{N}}$ is monotone increasing. From the definition of a_k in (29), we have that, for all $K\geq 1$,

$$\sum_{k=1}^{K}a_k = \gamma_1\mathbb{E}\left[\|\boldsymbol{\theta}_1-\boldsymbol{\theta}\|^2_{\mathsf{H}_1}\right]+\underbrace{\sum_{k=2}^{K}\left\{\gamma_k\mathbb{E}\left[\|\boldsymbol{\theta}_k-\boldsymbol{\theta}\|^2_{\mathsf{H}_k}\right]-\gamma_{k-1}\mathbb{E}\left[\|\boldsymbol{\theta}_k-\boldsymbol{\theta}\|^2_{\mathsf{H}_{k-1}}\right]\right\}}_{\Gamma_K}$$
$$-\gamma_K\mathbb{E}\left[\|\boldsymbol{\theta}_{K+1}-\boldsymbol{\theta}\|^2_{\mathsf{H}_K}\right]. \tag{37}$$

Since $\overline{\mathsf{H}}_k \in \mathbb{S}^d_{++}$ exists such that $\mathsf{H}_k = \overline{\mathsf{H}}_k^2$, we have $\|\boldsymbol{x}\|^2_{\mathsf{H}_k} = \|\overline{\mathsf{H}}_k\boldsymbol{x}\|^2$ for all $\boldsymbol{x}\in\mathbb{R}^d$. Accordingly, we have

$$\Gamma_K = \mathbb{E}\left[\sum_{k=2}^{K}\left\{\gamma_k\left\|\overline{\mathsf{H}}_k(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right\|^2-\gamma_{k-1}\left\|\overline{\mathsf{H}}_{k-1}(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right\|^2\right\}\right].$$

From $\overline{\mathsf{H}}_k = \mathsf{diag}(\hat{v}_{k,i}^{1/4})$, we have that, for all $\boldsymbol{x} = (x_i)_{i=1}^d \in \mathbb{R}^d$, $\|\overline{\mathsf{H}}_k\boldsymbol{x}\|^2 = \sum_{i=1}^d \sqrt{\hat{v}_{k,i}}x_i^2$. Hence, for all $K\geq 2$,

$$\Gamma_K = \mathbb{E}\left[\sum_{k=2}^{K}\sum_{i=1}^{d}\left(\gamma_k\sqrt{\hat{v}_{k,i}}-\gamma_{k-1}\sqrt{\hat{v}_{k-1,i}}\right)(\theta_{k,i}-\theta_i)^2\right]. \tag{38}$$

Condition (9) and $\gamma_k \geq \gamma_{k-1}$ $(k\geq 1)$ imply that, for all $k\geq 1$ and all $i\in[d]$,

$$\gamma_k\sqrt{\hat{v}_{k,i}}-\gamma_{k-1}\sqrt{\hat{v}_{k-1,i}} \geq 0.$$

Moreover, (A2) ensures that $\tilde{D}(\boldsymbol{\theta}) := \sup\{\max_{i\in[d]}(\theta_{k,i}-\theta_i)^2 \colon k\in\mathbb{N}\} < +\infty$. Accordingly, for all $K\geq 2$,

$$\Gamma_K \leq \tilde{D}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{k=2}^{K}\sum_{i=1}^{d}\left(\gamma_k\sqrt{\hat{v}_{k,i}}-\gamma_{k-1}\sqrt{\hat{v}_{k-1,i}}\right)\right] = \tilde{D}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^{d}\left(\gamma_K\sqrt{\hat{v}_{K,i}}-\gamma_1\sqrt{\hat{v}_{1,i}}\right)\right].$$

Therefore, (37), $\mathbb{E}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|^2_{\mathsf{H}_1}] \leq \tilde{D}(\boldsymbol{\theta})\mathbb{E}[\sum_{i=1}^d \sqrt{\hat{v}_{1,i}}]$, and (31) imply, for all $K \geq 1$,

$$\begin{aligned}\sum_{k=1}^K a_k &\leq \gamma_1 \tilde{D}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^d \sqrt{\hat{v}_{1,i}}\right] + \tilde{D}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^d \left(\gamma_K \sqrt{\hat{v}_{K,i}} - \gamma_1 \sqrt{\hat{v}_{1,i}}\right)\right] \\ &= \gamma_K \tilde{D}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^d \sqrt{\hat{v}_{K,i}}\right] \\ &\leq \gamma_K \tilde{D}(\boldsymbol{\theta}) \sum_{i=1}^d \sqrt{\frac{M}{\tilde{\beta}_{2K}}} \\ &\leq \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1 \alpha_K \sqrt{\tilde{\beta}_{2K}}}.\end{aligned} \tag{39}$$

Inequality (33) with $\beta_{1k} = \beta_1$ and $\tilde{\beta}_{1k} := 1 - \beta_1^{k+1} \geq 1 - \beta_1 =: \hat{\beta}_1$ implies that

$$b_k \leq \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) \leq \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_1 \hat{\beta}_1} \left(\frac{\sigma^2}{b} + G^2\right). \tag{40}$$

Inequality (34) with $\beta_{1k} = \beta_1$ implies that

$$c_k \leq D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} = D(\boldsymbol{\theta}) G \frac{\hat{\beta}_1}{\beta_1}. \tag{41}$$

Hence, (29), (39), (40), and (41) ensure that, for all $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] \leq \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1 \alpha_K \sqrt{\tilde{\beta}_{2K}} K} + \frac{(\sigma^2 b^{-1} + G^2)}{2\sqrt{v_*}\beta_1 \hat{\beta}_1 K} \sum_{k=1}^K \alpha_k \sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_1}{\beta_1},$$

which completes the proof. □

Lemma 7 *Suppose that (S1)–(S3) and (A1)–(A2) hold, $\beta_{1k} := \beta_1 \in (0,1)$, and $(\alpha_k)_{k\in\mathbb{N}}$ is monotone decreasing. Then, Adam defined by Algorithm 1 with* (9) *satisfies the following: for all $K \geq 1$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq &\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1 \alpha_K \sqrt{\tilde{\beta}_{2K}} K} + \frac{(\sigma^2 b^{-1} + G^2)}{2\sqrt{v_*}\beta_1 \hat{\beta}_1 K} \sum_{k=1}^K \alpha_k \sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_1}{\beta_1} \\ &+ \hat{\beta}_1 D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),\end{aligned}$$

where the parameters are defined as in Lemma 6.

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$. Inequality (35) with $\beta_{1k} = \beta_1$ implies that, for all $K \geq 1$,

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_{k-1}\right] + \hat{\beta}_1 D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right).$$

Hence, Lemma 6 leads to Lemma 7. □

Lemma 8 *Suppose that (S1)–(S3) and (A1)–(A2) hold, $\beta_{1k} := \beta_1 \in (0,1)$, and $(\alpha_k)_{k\in\mathbb{N}}$ is monotone decreasing. Then, Adam defined by Algorithm 1 with* (9) *satisfies the following: for all $K \geq 1$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq &\frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1 \alpha_K \sqrt{\tilde{\beta}_{2K}} K} + \frac{(\sigma^2 b^{-1} + G^2)}{2\sqrt{v_*}\beta_1 \hat{\beta}_1 K} \sum_{k=1}^K \alpha_k \sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_1}{\beta_1} \\ &+ \left(\frac{1}{\beta_1} + 2\hat{\beta}_1\right) D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),\end{aligned}$$

where the parameters are defined as in Lemma 6.

Proof Let $\boldsymbol{\theta} \in \mathbb{R}^d$. Inequality (36) with $\beta_{1k} = \beta_1$ implies that, for all $K \geq 1$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \\
&\leq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] + \left(\frac{1}{\beta_1} + \hat{\beta}_1\right) D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),
\end{aligned}$$

which, together with Lemma 7, shows that Lemma 8 holds. □

A.2 Proof of Theorem 1

Proof Lemmas 4 and 5 with

$$\alpha_k = \alpha,\ \beta_{1k} = \beta_1,\ \beta_{2k} = \beta_2,\ \tilde{\beta}_{1k} = 1 - \beta_1^{k+1},\ \tilde{\beta}_{2k} = 1 - \beta_2^{k+1},\ \hat{\beta}_1 = 1 - \beta_1$$

imply that

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] &\leq \frac{D(\boldsymbol{\theta}^\star) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_1} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_1 \tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}^\star) G \frac{\hat{\beta}_1}{\beta_1} \\
&\quad + \hat{\beta}_1 D(\boldsymbol{\theta}^\star) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right), \\
\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla L(\boldsymbol{\theta}_k)\right] &\leq \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_1} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_1 \tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) \\
&\quad + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_1}{\beta_1} + D(\boldsymbol{\theta}) \left(\frac{1}{\beta_1} + 2\hat{\beta}_1\right) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),
\end{aligned}$$

which completes the proof. □

A.3 Proof of Corollary 1

Proof The sequences $(\tilde{\beta}_{1k})_{k\in\mathbb{N}}$ and $(\tilde{\beta}_{2k})_{k\in\mathbb{N}}$ converge to 1. Theorem 1 thus leads to Corollary 1. □

A.4 Proof of Theorem 2

Proof Lemmas 4 and 5 with

$$\begin{aligned}
&\alpha_k = \frac{1}{k^a},\ \beta_{1k} = 1 - \frac{1}{k^{b_1}},\ \beta_{2k} = \left(1 - \frac{1}{k^{b_2}}\right)^{\frac{1}{k+1}},\ \tilde{\beta}_{1k} = 1 - \beta_{1k}^{k+1} \geq 1 - \beta_{1k},\ \tilde{\beta}_{2k} = 1 - \beta_{2k}^{k+1}, \\
&\hat{\beta}_{1k} = 1 - \beta_{1k}
\end{aligned}$$

imply that

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \boldsymbol{m}_k\right] \leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} \\
&+ \hat{\beta}_{1k} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right) \\
\leq{}& \frac{D(\boldsymbol{\theta}^\star) M^{\frac{1}{4}} k^{b_1}}{v_*^{\frac{1}{4}} (k^{b_1} - 1)} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{1}{2\sqrt{v_*}(k^{b_1} - 1) k^{a + \frac{b_2}{2} - 2b_1}} \left(\frac{\sigma^2}{b} + G^2\right) \\
&+ \frac{1}{k^{b_1} - 1} D(\boldsymbol{\theta}^\star) G + \frac{1}{k^{b_1}} D(\boldsymbol{\theta}^\star) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] \leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}}}{v_*^{\frac{1}{4}} \beta_{1k}} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{\alpha_k \sqrt{\tilde{\beta}_{2k}}}{2\sqrt{v_*}\beta_{1k}\tilde{\beta}_{1k}} \left(\frac{\sigma^2}{b} + G^2\right) + D(\boldsymbol{\theta}) G \frac{\hat{\beta}_{1k}}{\beta_{1k}} \\
&+ D(\boldsymbol{\theta}) \left(\frac{1}{\beta_{1k}} + 2\hat{\beta}_{1k}\right) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right) \\
\leq{}& \frac{D(\boldsymbol{\theta}) M^{\frac{1}{4}} k^{b_1}}{v_*^{\frac{1}{4}} (k^{b_1} - 1)} \sqrt{\frac{\sigma^2}{b} + G^2} + \frac{1}{2\sqrt{v_*}(k^{b_1} - 1) k^{a + \frac{b_2}{2} - 2b_1}} \left(\frac{\sigma^2}{b} + G^2\right) \\
&+ \frac{1}{k^{b_1} - 1} D(\boldsymbol{\theta}) G + \frac{1}{k^{b_1}} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right) \\
&+ \frac{{k^{b_1}}^2 + 2k^{b_1} - 2}{k^{b_1}(k^{b_1} - 1)} D(\boldsymbol{\theta}) \left(B + \sqrt{\frac{\sigma^2}{b} + G^2}\right),
\end{aligned}
$$

which completes the proof. □

A.5 Proof of Corollary 2

Proof Since $a - b_1 + b_2/2 > 0$, we have that

$$
\frac{1}{(k^{b_1} - 1) k^{a + \frac{b_2}{2} - 2b_1}} = \frac{1}{k^{a + \frac{b_2}{2} - b_1} - k^{a + \frac{b_2}{2} - 2b_1}} = \frac{1}{k^{a + \frac{b_2}{2} - b_1}} \left(1 - \frac{1}{k^{b_1}}\right)^{-1} \to 0.
$$

Theorem 2 thus leads to Corollary 2. □

A.6 Proof of Theorem 3

Proof Lemmas 7 and 8 with

$$
\alpha_k = \alpha,\ \beta_{1k} = \beta_1,\ \beta_{2k} = \beta_2,\ \tilde{\beta}_{1k} = 1 - \beta_1^{k+1},\ \tilde{\beta}_{2k} = 1 - \beta_2^{k+1} \leq 1,\ \hat{\beta}_1 = 1 - \beta_1
$$

ensure that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] &\leq \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\alpha\sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)\\ &\leq \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1}\alpha + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)\end{aligned}$$

and that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta})^\top \nabla f(\boldsymbol{\theta}_k)\right] &\leq \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\alpha\sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta})G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \left(\frac{1}{\beta_1}+2\hat{\beta}_1\right) D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)\\ &\leq \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1}\alpha + D(\boldsymbol{\theta})G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \left(\frac{1}{\beta_1}+2\hat{\beta}_1\right) D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right),\end{aligned}$$

which completes the proof. □

A.7 Proof of Theorem 4

Proof Let

$$\alpha_k = \alpha,\ \beta_{1k} = \beta_1,\ \beta_{2k} = \left(1-\frac{1}{k^{b_2}}\right)^{\frac{1}{k+1}},\ \tilde{\beta}_{1k} = 1-\beta_1^{k+1} \leq 1,\ \tilde{\beta}_{2k} = 1-\beta_{2k}^{k+1},\ \hat{\beta}_1 = 1-\beta_1,$$

where $b_2 \in (0,2)$. We have that

$$\sqrt{\tilde{\beta}_{2k}} = \sqrt{1-\beta_{2k}^{k+1}} = \sqrt{\frac{1}{k^{b_2}}}.$$

Lemma 7 ensures that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] &\leq \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\alpha\sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right)\\ &\leq \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1\alpha K^{1-\frac{b_2}{2}}} + \frac{(\sigma^2 b^{-1}+G^2)\alpha}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\frac{1}{k^{\frac{b_2}{2}}} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1}\\ &\quad + \hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right).\end{aligned}$$

We also have that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{k^{\frac{b_2}{2}}} &\leq \frac{1}{K}\left(1+\int_{1}^{K}\frac{\mathrm{d}t}{t^{\frac{b_2}{2}}}\right) = \frac{1}{K}\left\{1+\left[\left(1-\frac{b_2}{2}\right)t^{1-\frac{b_2}{2}}\right]_1^K\right\} \\ &\leq \frac{1}{K}\left\{1+\left(1-\frac{b_2}{2}\right)K^{1-\frac{b_2}{2}}\right\} \leq \frac{2}{K}K^{1-\frac{b_2}{2}} = \frac{2}{K^{\frac{b_2}{2}}}.\end{aligned} \tag{42}$$

Hence,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] \leq{}& \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1\alpha K^{1-\frac{b_2}{2}}} + \frac{(\sigma^2 b^{-1}+G^2)\alpha}{\sqrt{v_*}\beta_1\hat{\beta}_1 K^{\frac{b_2}{2}}} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1} \\ &+\hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right).\end{aligned}$$

A discussion similar to the one showing the above inequality and Lemma 8 imply that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\nabla f(\boldsymbol{\theta}_k)^\top(\boldsymbol{\theta}_k-\boldsymbol{\theta})\right] \leq{}& \frac{d\tilde{D}(\boldsymbol{\theta})\sqrt{M}}{2\alpha\beta_1 K^{1-\frac{b_2}{2}}} + \frac{\alpha}{\sqrt{v_*}\beta_1(1-\beta_1)K^{\frac{b_2}{2}}}\left(\frac{\sigma^2}{b}+G^2\right) \\ &+\frac{1-\beta_1}{\beta_1}D(\boldsymbol{\theta})G+\left(\frac{1}{\beta_1}+2(1-\beta_1)\right)D(\boldsymbol{\theta})\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right),\end{aligned}$$

which completes the proof. □

A.8 Proof of Theorem 5

Proof Let

$$\alpha_k = \frac{1}{k^a},\ \beta_{1k}=\beta_1,\ \beta_{2k}=\beta_2,\ \tilde{\beta}_{1k}=1-\beta_1^{k+1}\leq 1,\ \tilde{\beta}_{2k}=1-\beta_2^{k+1}\leq 1,\ \hat{\beta}_1=1-\beta_1.$$

We have that $\tilde{\beta}_{2k}=1-\beta_{2k}^{k+1}\geq 1-\beta_2$. Lemma 7 ensures that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] \leq{}& \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}\tilde{\beta}_{1K}}{2\beta_1\alpha_K\sqrt{\tilde{\beta}_{2K}}K} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\alpha_k\sqrt{\tilde{\beta}_{2k}} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1} \\ &+\hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right) \\ \leq{}& \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1\sqrt{1-\beta_2}K^{1-a}} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K}\sum_{k=1}^{K}\frac{1}{k^a} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1} \\ &+\hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right),\end{aligned}$$

which, together with (42), implies that

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{\theta}_k-\boldsymbol{\theta}^\star)^\top \boldsymbol{m}_k\right] \leq{}& \frac{d\tilde{D}(\boldsymbol{\theta}^\star)\sqrt{M}}{2\beta_1\sqrt{1-\beta_2}K^{1-a}} + \frac{(\sigma^2 b^{-1}+G^2)}{2\sqrt{v_*}\beta_1\hat{\beta}_1 K^a} + D(\boldsymbol{\theta}^\star)G\frac{\hat{\beta}_1}{\beta_1} \\ &+\hat{\beta}_1 D(\boldsymbol{\theta}^\star)\left(B+\sqrt{\frac{\sigma^2}{b}+G^2}\right).\end{aligned}$$

A discussion similar to the one showing the above inequality and Lemma 8 implies the second assertion in Theorem 5. □

A.9 Proof of Theorem 6

Proof The proofs of Theorems 4 and 5 lead to Theorem 6. □

Declarations

Ethical Approval:

Not Applicable

Availability of supporting data:

Not Applicable

Competing interests:

Not Applicable

Funding:

This work was supported by JSPS KAKENHI Grant Number 21K11773.

Authors' contributions:

H.I. wrote the manuscript.

Acknowledgments:

I am sincerely grateful to Editor Claude Brezinski, Associate Editor, and the anonymous reviewers for helping me improve the original manuscript. I also thank Naoki Sato for his input on the numerical examples.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN `https://arxiv.org/pdf/1701.07875.pdf` (2017)
2. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples. Springer, New York (2000)
3. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Review **60**, 223–311 (2018)
4. Chen, H., Zheng, L., AL Kontar, R., Raskutti, G.: Stochastic gradient descent in correlated settings: A study on Gaussian processes. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
5. Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., Gu, Q.: Closing the generalization gap of adaptive gradient methods in training deep neural network. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, vol. 452, pp. 3267–3275 (2021)
6. Chen, X., Liu, S., Sun, R., Hong, M.: On the convergence of a class of Adam-type algorithms for non-convex optimization. In: Proceedings of The International Conference on Learning Representations (2019)

7. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research **12**, 2121–2159 (2011)
8. Fehrman, B., Gess, B., Jentzen, A.: Convergence rates for the stochastic gradient descent method for non-convex objective functions. Journal of Machine Learning Research **21**, 1–48 (2020)
9. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. SIAM Journal on Optimization **22**, 1469–1492 (2012)
10. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. SIAM Journal on Optimization **23**, 2061–2089 (2013)
11. Gower, R.M., Sebbouh, O., Loizou, N.: SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, vol. 130 (2021)
12. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
13. Iiduka, H.: Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. IEEE Transactions on Cybernetics **52**(12), 13250–13261 (2022)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of The International Conference on Learning Representations (2015)
15. Loizou, N., Vaswani, S., Laradji, I., Lacoste-Julien, S.: Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, vol. 130 (2021)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
17. Luo, L., Xiong, Y., Liu, Y., Sun, X.: Adaptive gradient methods with dynamic bound of learning rate. In: Proceedings of The International Conference on Learning Representations (2019)
18. Mendler-Dünner, C., Perdomo, J.C., Zrnic, T., Hardt, M.: Stochastic optimization for performative prediction. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
19. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization **19**, 1574–1609 (2009)
20. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Doklady AN USSR **269**, 543–547 (1983)
21. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics **4**, 1–17 (1964)
22. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. In: Proceedings of The International Conference on Learning Representations (2018)
23. Robbins, H., Monro, H.: A stochastic approximation method. The Annals of Mathematical Statistics **22**, 400–407 (1951)
24. Scaman, K., Malherbe, C.: Robustness analysis of non-convex stochastic gradient descent using biased expectations. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
25. Shallue, C.J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., Dahl, G.E.: Measuring the effects of data parallelism on neural network training. Journal of Machine Learning Research **20**, 1–49 (2019)
26. Smith, S.L., Kindermans, P.J., Le, Q.V.: Don't decay the learning rate, increase the batch size. In: Proceedings of The International Conference on Learning Representations (2018)
27. Tieleman, T., Hinton, G.: RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4**, 26–31 (2012)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
29. Virmaux, A., Scaman, K.: Lipschitz regularity of deep neural networks: analysis and efficient estimation. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
30. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 2048–2057 (2015)
31. Zaheer, M., Reddi, S., Sachan, D., Kale, S., Kumar, S.: Adaptive methods for nonconvex optimization. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018)
32. Zhang, G., Li, L., Nado, Z., Martens, J., Sachdeva, S., Dahl, G.E., Shallue, C.J., Grosse, R.: Which algorithmic choices matter at which batch sizes? Insights from a noisy quadratic model. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

33. Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., Gu, Q.: On the convergence of adaptive gradient methods for nonconvex optimization. In: 12th Annual Workshop on Optimization for Machine Learning (2020)
34. Zhuang, J., Tang, T., Ding, Y., Tatikonda, S., Dvornek, N., Papademetris, X., Duncan, J.S.: AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
35. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the 20th International Conference on Machine Learning, pp. 928–936 (2003)
36. Zinkevich, M., Weimer, M., Li, L., Smola, A.: Parallelized stochastic gradient descent. In: Advances in Neural Information Processing Systems, vol. 23 (2010)
37. Zou, F., Shen, L., Jie, Z., Zhang Weizhong, W.L.: A sufficient condition for convergences of Adam and RMSProp. In: Computer Vision and Pattern Recognition Conference, pp. 11127–11135 (2019)