A Sufficient Condition for Convergences of Adam and RMSProp

Fangyu Zou[†*], Li Shen[‡*], Zequn Jie[‡], Weizhong Zhang[‡], Wei Liu[‡]

[‡]Tencent AI Lab   [†]Stony Brook University

fangyu.zou@stonybrook.edu, mathshenli@gmail.com, zequn.nus@gmail.com,
zhangweizhongzju@gmail.com, wl2223@columbia.edu

Abstract

Adam and RMSProp are two of the most influential adaptive stochastic algorithms for training deep neural networks, which have been pointed out to be divergent even in the convex setting via a few simple counterexamples. Many attempts, such as decreasing an adaptive learning rate, adopting a big batch size, incorporating a temporal decorrelation technique, seeking an analogous surrogate, etc., have been tried to promote Adam/RMSProp-type algorithms to converge. In contrast with existing approaches, we introduce an alternative easy-to-check sufficient condition, which merely depends on the parameters of the base learning rate and combinations of historical second-order moments, to guarantee the global convergence of generic Adam/RMSProp for solving large-scale non-convex stochastic optimization. Moreover, we show that the convergences of several variants of Adam, such as AdamNC, AdaEMA, etc., can be directly implied via the proposed sufficient condition in the non-convex setting. In addition, we illustrate that Adam is essentially a specifically weighted AdaGrad with exponential moving average momentum, which provides a novel perspective for understanding Adam and RMSProp. This observation coupled with this sufficient condition gives much deeper interpretations on their divergences. At last, we validate the sufficient condition by applying Adam and RMSProp to tackle a certain counterexample and train deep neural networks. Numerical results are exactly in accord with our theoretical analysis.

1. Introduction

Large-scale non-convex stochastic optimization [5], covering a slew of applications in statistics and machine learning [13, 5] such as learning a latent variable from massive data whose probability density distribution is unknown, takes the following generic formulation:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \ f(\boldsymbol{x}) = \mathbb{E}_{\xi \sim \mathbb{P}}\left[\widetilde{f}(\boldsymbol{x}, \xi)\right], \qquad (1)$$

---

where $f(\boldsymbol{x})$ is a non-convex function and ξ is a random variable satisfying an unknown distribution $\mathbb{P}$.

Due to the uncertainty of distribution $\mathbb{P}$, the batch gradient descent algorithm [4] is impractical to employ full gradient $\boldsymbol{\nabla} f(\boldsymbol{x})$ to solve problem (1). Alternatively, a compromised approach to handle this difficulty is to use an unbiased stochastic estimate of $\boldsymbol{\nabla} f(\boldsymbol{x})$, denoted as $g(\boldsymbol{x}, \xi)$, which leads to the stochastic gradient descent (SGD) algorithm [21]. Its coordinate-wise version is defined as follows:

$$\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \eta_{t,k}\boldsymbol{g}_{t,k}(\boldsymbol{x}_t, \xi_t), \qquad (2)$$

for $k = 1, 2, \ldots, d$, where $\eta_{t,k} \geq 0$ is the learning rate of the k-th component of stochastic gradient $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$ at the t-th iteration. A sufficient condition [21] to ensure the global convergence of vanilla SGD (2) is to require η_t to meet

$$\sum_{t=1}^{\infty} \|\eta_t\| = \infty \ \text{ and } \ \sum_{t=1}^{\infty} \|\eta_t\|^2 < \infty. \qquad (3)$$

Although the vanilla SGD algorithm with learning rate η_t satisfying condition (3) does converge, its empirical performance could be still stagnating, since it is difficult to tune an effective learning rate η_t via condition (3).

To further improve the empirical performance of SGD, a large variety of adaptive SGD algorithms, including AdaGrad [9], RMSProp [11], Adam [14], Nadam [8], *etc.*, have been proposed to automatically tune the learning rate η_t by using second-order moments of historical stochastic gradients. Let $\boldsymbol{v}_{t,k}$ and $\boldsymbol{m}_{t,k}$ be the linear combinations of the historical second-order moments $(\boldsymbol{g}_{1,k}^2, \boldsymbol{g}_{2,k}^2, \cdots, \boldsymbol{g}_{t,k}^2)$ and stochastic gradient estimates $(\boldsymbol{g}_{1,k}, \boldsymbol{g}_{2,k}, \cdots, \boldsymbol{g}_{t,k})$, respectively. Then, the generic iteration scheme of these adaptive SGD algorithms [20, 6] is summarized as

$$\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \eta_{t,k}\boldsymbol{m}_{t,k}, \ \text{ with } \eta_{t,k} = \alpha_t/\sqrt{\boldsymbol{v}_{t,k}}, \quad (4)$$

for $k = 1, 2, \ldots, d$, where $\alpha_t > 0$ is called base learning rate and it is independent of stochastic gradient estimates $(\boldsymbol{g}_{1,k}, \boldsymbol{g}_{2,k}, \cdots, \boldsymbol{g}_{t,k})$ for all $t \geq 1$. Although RMSProp, Adam, and Nadam work well for solving large-scale convex and non-convex optimization problems such as training deep neural networks, they have been pointed out to be divergent in some scenarios via convex counterexamples [20].

This finding thoroughly destroys the fluke of a direct use of these algorithms without any further assumptions or corrections. Recently, developing sufficient conditions to guarantee global convergences of Adam and RMSProp-type algorithms has attracted much attention from both machine learning and optimization communities. The existing successful attempts can be divided into four categories:

(C1) Decreasing a learning rate. Reddi *et al*. [20] have declared that the core cause of divergences of Adam and RMSProp is largely controlled by the difference between the two adjacent learning rates, *i.e.*,

$$\Gamma_t = 1/\boldsymbol{\eta}_t - 1/\boldsymbol{\eta}_{t-1} = \sqrt{\boldsymbol{v}_t}/\alpha_t - \sqrt{\boldsymbol{v}_{t-1}}/\alpha_{t-1}. \quad (5)$$

Once positive definiteness of Γ_t is violated, Adam and RMSProp may suffer from divergence [20]. Based on this observation, two variants of Adam called AMSGrad and AdamNC have been proposed with convergence guarantees in both the convex [20] and non-convex [6] stochastic settings by requiring $\Gamma_t \succ 0$. In addition, Padam [24] extended from AMSGrad has been proposed to contract the generalization gap in training deep neural networks, whose convergence has been ensured by requiring $\Gamma_t \succ 0$. In the strongly convex stochastic setting, by using the long-term memory technique developed in [20], Huang *et al*. [12] have proposed NosAdam by attaching more weights on historical second-order moments to ensure its convergence. Prior to that, the convergence rate of RMSProp [19] has already been established in the convex stochastic setting by employing similar parameters to those of AdamNC [20].

(C2) Adopting a big batch size. Basu *et al*. [2], for the first time, showed that deterministic Adam and RMSProp with original iteration schemes are actually convergent by using full-batch gradient. On the other hand, both Adam and RMSProp can be reshaped as specific signSGD-type algorithms [1, 3] whose $\mathcal{O}(1/\sqrt{T})$ convergence rates have been provided in the non-convex stochastic setting by setting batch size as large as the number of maximum iterations [3]. Recently, Zaheer *et al*. [23] have established $\mathcal{O}(1/\sqrt{T})$ convergence rate of original Adam directly in the non-convex stochastic setting by requiring the batch size to be the same order as the number of maximum iterations. We comment that this type of requirement is impractical when Adam and RMSProp are applied to tackle large-scale problems like (1), since these approaches cost a huge number of computations to estimate big-batch stochastic gradients in each iteration.

(C3) Incorporating a temporal decorrelation. By exploring the structure of the convex counterexample in [20], Zhou *et al*. [25] have pointed out that the divergence of RMSProp is fundamentally caused by the unbalanced learning rate rather than the absence of $\Gamma_t \succ 0$. Based on this viewpoint, Zhou *et al*. [25] have proposed AdaShift by incorporating a temporal decorrelation technique to eliminate the inappropriate correlation between $\boldsymbol{v}_{t,k}$ and the current second-order

moment $\boldsymbol{g}_{t,k}^2$, in which the adaptive learning rate $\eta_{t,k}$ is required to be independent of $\boldsymbol{g}_{t,k}^2$. However, convergence of AdaShift in [25] was merely restricted to RMSProp for solving the convex counterexample in [20].

(C4) Seeking an analogous surrogate. Due to the divergences of Adam and RMSProp [20], Zou *et al*. [26] recently proposed a class of new surrogates called AdaUSM to approximate Adam and RMSProp by integrating weighted AdaGrad with unified heavy ball and Nesterov accelerated gradient momentums. Its $\mathcal{O}(\log(T)/\sqrt{T})$ convergence rate has also been provided in the non-convex stochastic setting by requiring a non-decreasing weighted sequence. Besides, many other adaptive stochastic algorithms without combining momentums, such as AdaGrad [22, 18] and stagewise AdaGrad [7], have been guaranteed to be convergent and work well in the non-convex stochastic setting.

In contrast with the above four types of modifications and restrictions, we introduce an alternative easy-to-check sufficient condition (abbreviated as **(SC)**) to guarantee the global convergences of original Adam and RMSProp. The proposed **(SC)** merely depends on the parameters in estimating $\boldsymbol{v}_{t,k}$ and base learning rate α_t. **(SC)** neither requires the positive definiteness of Γ_t like **(C1)** nor needs the batch size as large as the same order as the number of maximum iterations like **(C2)** in both the convex and non-convex stochastic settings. Thus, it is easier to verify and more practical compared with **(C1)**-**(C3)**. On the other hand, **(SC)** is partially overlapped with **(C1)** since the proposed **(SC)** can cover AdamNC [20], AdaGrad with exponential moving average (AdaEMA) momentum [6], and RMSProp [19] as instances whose convergences are all originally motivated by requiring the positive definiteness of Γ_t. While, based on **(SC)**, we can directly derive their global convergences in the non-convex stochastic setting as byproducts without checking the positive definiteness of Γ_t step by step. Besides, **(SC)** can serve as an alternative explanation on divergences of original Adam and RMSProp, which are possibly due to incorrect parameter settings for accumulating the historical second-order moments rather than the unbalanced learning rate caused by the inappropriate correlation between $\boldsymbol{v}_{t,k}$ and $\boldsymbol{g}_{t,k}^2$ like **(C3)**. In addition, AdamNC and AdaEMA are convergent under **(SC)**, but violate **(C3)** in each iteration.

Moreover, by carefully reshaping the iteration scheme of Adam, we obtain a specific weighted AdaGrad with exponential moving average momentum, which extends the weighted AdaGrad with heavy ball momentum and Nesterov accelerated gradient momentum [26] in two aspects: the new momentum mechanism and the new base learning rate setting provide a new perspective for understanding Adam and RMSProp. At last, we experimentally verify **(SC)** by applying Adam and RMSProp with different parameter settings to solve the counterexample [20] and train deep neural networks including LeNet [16] and ResNet [10]. In summary,

the contributions of this work are five-fold:

(1) We introduce an easy-to-check sufficient condition to ensure the global convergences of original Adam and RMSProp in the non-convex stochastic setting. Moreover, this sufficient condition is distinctive from the existing conditions **(C1)**-**(C4)** and is easier to verify.

(2) We reshape Adam as weighted AdaGrad with exponential moving average momentum, which provides a new perspective for understanding Adam and RMSProp and also complements AdaUSM in [26].

(3) We provide a new explanation on the divergences of original Adam and RMSProp, which are possibly due to an incorrect parameter setting of the combinations of historical second-order moments based on **(SC)**.

(4) We find that the sufficient condition extends the restrictions of RMSProp [19] and covers many convergent variants of Adam, *e.g.*, AdamNC, AdaGrad with momentum, *etc.* Thus, their convergences in the non-convex stochastic setting naturally hold.

(5) We conduct experiments to validate the sufficient condition for the convergences of Adam/RMSProp. The experimental results match our theoretical results.

2. Generic Adam

For readers' convenience, we first clarify a few necessary notations used in the forthcoming Generic Adam. We denote $x_{t,k}$ as the k-th component of $x_t \in \mathbb{R}^d$, and $g_{t,k}$ as the k-th component of the stochastic gradient at the t-th iteration, and call $\alpha_t > 0$ base learning rate and β_t momentum parameter, respectively. Let $\epsilon > 0$ be a sufficiently small constant. Denote $\mathbf{0} = (0, \cdots, 0)^\top \in \mathbb{R}^d$, and $\boldsymbol{\epsilon} = (\epsilon, \cdots, \epsilon)^\top \in \mathbb{R}^d$. All operations, such as multiplying, dividing, and taking square root, are executed in the coordinate-wise sense.

Algorithm 1 Generic Adam

1: **Parameters:** Choose $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\theta_t\}$. Choose $x_1 \in \mathbb{R}^d$ and set initial values $\boldsymbol{m}_0 = \mathbf{0}$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:     Sample a stochastic gradient $\boldsymbol{g}_t$;
4:     **for** $k = 1, 2, \ldots, d$ **do**
5:         $\boldsymbol{v}_{t,k} = \theta_t \boldsymbol{v}_{t-1,k} + (1 - \theta_t)\boldsymbol{g}_{t,k}^2$;
6:         $\boldsymbol{m}_{t,k} = \beta_t \boldsymbol{m}_{t-1,k} + (1 - \beta_t)\boldsymbol{g}_{t,k}$;
7:         $\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \alpha_t \boldsymbol{m}_{t,k}/\sqrt{\boldsymbol{v}_{t,k}}$;
8:     **end for**
9: **end for**

Generic Adam covers RMSProp by setting $\beta_t = 0$. Moreover, it covers Adam with a bias correction [14] as follows:

Remark 1. *The original Adam with the bias correction [14] takes constant parameters $\beta_t = \beta$ and $\theta_t = \theta$. The iteration scheme is written as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \widehat{\alpha}_t \frac{\widehat{\boldsymbol{m}}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}}$, with $\widehat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1 - \beta^t}$ and $\widehat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1 - \theta^t}$. Let $\alpha_t = \widehat{\alpha}_t \frac{\sqrt{1 - \theta^t}}{1 - \beta^t}$. Then, the above can be rewritten as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t}$. Thus, it is equivalent to taking constant β_t, constant θ_t, and new base learning rate α_t in Generic Adam.*

2.1. Weighted AdaGrad Perspective

Now we show that Generic Adam can be reformulated as a new type of weighted AdaGrad algorithms with exponential moving average momentum (Weighted AdaEMA).

Algorithm 2 Weighted AdaEMA

1: **Parameters:** Choose parameters $\{\alpha_t\}$, momentum factors $\{\beta_t\}$, and weights $\{w_t\}$. Set $W_0 = 1$, $\boldsymbol{m}_0 = \mathbf{0}$, $\boldsymbol{V}_0 = \boldsymbol{\epsilon}$, and $\boldsymbol{x}_1 \in \mathbb{R}^n$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:     Sample a stochastic gradient $\boldsymbol{g}_t$;
4:     $W_t = W_{t-1} + w_t$;
5:     **for** $k = 1, 2, \ldots, d$ **do**
6:         $\boldsymbol{V}_{t,k} = \boldsymbol{V}_{t-1,k} + w_t \boldsymbol{g}_{t,k}^2$;
7:         $\boldsymbol{m}_{t,k} = \beta_t \boldsymbol{m}_{t-1,k} + (1 - \beta_t)\boldsymbol{g}_{t,k}$;
8:         $\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \alpha_t \boldsymbol{m}_{t,k}/\sqrt{\boldsymbol{V}_{t,k}/W_t}$;
9:     **end for**
10: **end for**

Remark 2. *The Weighted AdaEMA is a natural generalization of the AdaGrad algorithm. The classical AdaGrad is to take the weights $w_t = 1$, the momentum factors $\beta_t = 0$, and the parameters $\alpha_t = \eta/\sqrt{t+1}$ for constant η.*

The following proposition states the equivalence between Generic Adam and Weighted AdaEMA.

Proposition 3. *Algorithm 1 and Algorithm 2 are equivalent.*

The divergence issue of Adam/RMSProp. When θ_t is taken constant, *i.e.*, $\theta_t = \theta$, Reddi *et al.* [20] have pointed out that Adam and RMSProp ($\beta_t = 0$) can be divergent even in the convex setting. They conjectured that the divergence is possibly due to the uncertainty of positive definiteness of Γ_t in Eq. (5). This idea has motivated many new convergent variants of Adam by forcing $\Gamma_t \succ 0$. Recently, Zhou *et al.* [25] further argued that the nature of divergences of Adam and RMSProp is possibly due to the unbalanced learning rate η_t caused by the inappropriate correlation between $v_{t,k}$ and $g_{t,k}^2$ by studying the counterexample in [20]. However, this explanation can be violated by many existing convergent Adam-type algorithms such as AdamNC, NosAdam [12], *etc.* So far, there is no satisfactory explanation for the core reason of the divergence issue. We will provide more insights in Section 4 based on our theoretical analysis.

3. Main Results

In this section, we characterize the upper-bound of gradient residual of problem (1) as a function of parameters (θ_t, α_t). Then the convergence rate of Generic Adam is derived directly by specifying appropriate parameters (θ_t, α_t). Below, we state the necessary assumptions that are commonly used for analyzing the convergence of a stochastic algorithm for non-convex problems:

(A1) The minimum value of problem (1) is lower-bounded, *i.e.*, $f^* = \min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) > -\infty$;

(A2) The gradient of f is L-Lipschitz continuous, *i.e.*, $\|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla} f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$;

(A3) The stochastic gradient $\boldsymbol{g}_t$ is an unbiased estimate, *i.e.*, $\mathbb{E}[\boldsymbol{g}_t] = \boldsymbol{\nabla} f_t(\boldsymbol{x}_t)$;

(A4) The second-order moment of stochastic gradient $\boldsymbol{g}_t$ is uniformly upper-bounded, *i.e.*, $\mathbb{E}\|\boldsymbol{g}_t\|^2 \leq G$.

To establish the upper-bound, we also suppose that the parameters $\{\beta_t\}$, $\{\theta_t\}$, and $\{\alpha_t\}$ satisfy the restrictions:

(R1) The parameters $\{\beta_t\}$ satisfy $0 \leq \beta_t \leq \beta < 1$ for all t for some constant β;

(R2) The parameters $\{\theta_t\}$ satisfy $0 < \theta_t < 1$ and θ_t is non-decreasing in t with $\theta := \lim_{t \to \infty} \theta_t > \beta^2$;

(R3) The parameters $\{\alpha_t\}$ satisfy that $\chi_t := \frac{\alpha_t}{\sqrt{1-\theta_t}}$ is "almost" non-increasing in t, by which we mean that there exist a non-increasing sequence $\{a_t\}$ and a positive constant C_0 independent of t such that $a_t \leq \chi_t \leq C_0 a_t$.

The restriction (R3) indeed says that χ_t is the product between some non-increasing sequence $\{a_t\}$ and some bounded sequence. This is a slight generalization of χ_t itself being non-decreasing. If χ_t itself is non-increasing, we can then take $a_t = \chi_t$ and $C_0 = 1$. For most of the well-known Adam-type methods, χ_t is indeed non-decreasing, for instance, for AdaGrad with EMA momentum we have $\alpha_t = \eta/\sqrt{t}$ and $\theta_t = 1 - 1/t$, so $\chi_t = \eta$ is constant; for Adam with constant $\theta_t = \theta$ and non-increasing α_t (say $\alpha_t = \eta/\sqrt{t}$ or $\alpha_t = \eta$), $\chi_t = \alpha_t/\sqrt{1-\theta}$ is non-increasing. The motivation, instead of χ_t being decreasing, is that it allows us to deal with the bias correction steps in Adam [14].

We fix a positive constant $\theta' > 0$[1] such that $\beta^2 < \theta' < \theta$. Let $\gamma := \beta^2/\theta' < 1$ and

$$C_1 := \prod_{j=1}^{N} \left(\frac{\theta_j}{\theta'} \right), \qquad (6)$$

where N is the maximum of the indices j with $\theta_j < \theta'$. The finiteness of N is guaranteed by the fact that $\lim_{t \to \infty} \theta_t = \theta > \theta'$. When there are no such indices, *i.e.*, $\theta_1 \geq \theta'$, we take $C_1 = 1$ by convention. In general, $C_1 \leq 1$. Our main results on estimating gradient residual state as follows:

---

[1]In the special case that $\theta_t = \theta$ is constant, we can directly set $\theta' = \theta$.

Theorem 4. *Let $\{\boldsymbol{x}_t\}$ be a sequence generated by Generic Adam for initial values $\boldsymbol{x}_1$, $\boldsymbol{m}_0 = \boldsymbol{0}$, and $\boldsymbol{v}_0 = \epsilon$. Assume that f and stochastic gradients $\boldsymbol{g}_t$ satisfy assumptions (A1)-(A4). Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. We have*

$$\left(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^{4/3} \right] \right)^{3/2} \leq \frac{C + C' \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t}}{T\alpha_T},$$

where $C' = 2C_0^2 C_3 d\sqrt{G^2 + \epsilon d}/[(1-\beta)\theta_1]$ and

$$C = \frac{2C_0\sqrt{G^2 + \epsilon d}}{1 - \beta}\left[(C_4 + C_3 C_0 d\chi_1 \log\left(1 + \frac{G^2}{\epsilon d}\right)) \right],$$

where C_4 and C_3 are defined as $C_4 = f(x_1) - f^*$ and $C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left[\frac{C_0^2 \chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2 G \right]$.

Theorem 5. *Suppose the same setting and hypothesis as Theorem 4. Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. Then for any $\delta > 0$, the following bound holds with probability at least $1 - \delta^{2/3}$:*

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^2 \leq \frac{C + C' \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t}}{\delta T\alpha_T} := Bound(T),$$

where C and C' are defined as those in Theorem 4.

Remark 6. **(i)** *The constants C and C' depend on apriori known constants $C_0, C_1, \beta, \theta', G, L, \epsilon, d, f^*$ and $\theta_1, \alpha_1, \boldsymbol{x}_1$.* **(ii)** *Convergence in expectation in Theorem 4 is slightly stronger than convergence in probability in Theorem 5. Convergence in expectation is on the term $(\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^{\frac{3}{2}}])^{\frac{4}{3}}$, which is slightly weaker than $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$. The latter is adopted for most SGD variants with global learning rates, namely, the learning rate for each coordinate is the same. This is due to that $\frac{1}{\sum_{t=1}^{T} \alpha_t} \mathbb{E}\sum_{t=1}^{T} \alpha_t \|\nabla f(\boldsymbol{x}_t)\|^2$ is exactly $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$ if τ is randomly selected via distribution $\mathbb{P}(\tau = k) = \frac{\alpha_k}{\sum_{t=1}^{T} \alpha_t}$. This does not apply to coordinate-wise adaptive methods because the learning rate for each coordinate is different, and hence unable to randomly select an index according to some distribution uniform for each coordinate. On the other hand, the proofs of AMSGrad and AdaEMA [6] are able to achieve the bound for $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$. This is due to the strong assumption $\|g_t\| \leq G$ which results in a uniform lower bound for each coordinate of the adaptive learning rate $\eta_{t,k} \geq \alpha_t/G$. Thus, the proof of AMSGrad [6] can be dealt with in a way similar to the case of global learning rate. In our paper we use a coordinate-wise adaptive learning rate and assume a weaker assumption $\mathbb{E}[\|g_t\|^2] \leq G$ instead of $\|g_t\|^2 \leq G$. To separate the term $\|\nabla f(\boldsymbol{x}_t)\|$ from $\|\nabla f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2$, we can only apply the Hölder theorem to obtain a bound for $(\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^{\frac{3}{2}}])^{\frac{4}{3}}$.*

Corollary 7. *Take $\alpha_t = \eta/t^s$ with $0 \le s < 1$. Suppose $\lim_{t\to\infty} \theta_t = \theta < 1$. Then the $Bound(T)$ in Theorem 5 is bounded from below by constants*

$$Bound(T) \ge \frac{C'\sqrt{1-\theta}}{\delta}. \qquad (7)$$

In particular, when $\theta_t = \theta < 1$, we have the following more subtle estimate on lower and upper-bounds for $Bound(T)$

$$\frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta} \le Bound(T) \le \frac{C}{\delta\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta(1-s)}.$$

Remark 8. (i) *Corollary 7 shows that if $\lim_{t\to\infty} \theta_t = \theta < 1$, the bound in Theorem 5 is only $\mathcal{O}(1)$, hence not guaranteeing convergence. This result is not surprising as Adam with constant θ_t has already shown to be divergent [20]. Hence, $\mathcal{O}(1)$ is its best convergence rate we can expect. We will discuss this case in more details in Section 4.*
(ii) *Corollary 7 also indicates that in order to guarantee convergence, the parameter has to satisfy $\lim_{t\to\infty} \theta_t = 1$. Although we do not assume this in our restrictions (R1)-(R3), it turns out to be the consequence from our analysis. Note that if $\beta < 1$ in (R1) and $\lim_{t\to\infty} \theta_t = 1$, then the restriction $\lim_{t\to\infty} \theta_t > \beta^2$ is automatically satisfied in (R2).*

We are now ready to give the Sufficient Condition (**SC**) for convergence of Generic Adam based on Theorem 5.

Corollary 9 (Sufficient Condition(**SC**)). *Generic Adam is convergent if the parameters $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\theta_t\}$ satisfy*

1. *$\beta_t \le \beta < 1$;*
2. *$0 < \theta_t < 1$ and θ_t is non-decreasing in t;*
3. *$\chi_t := \alpha_t/\sqrt{1-\theta_t}$ is "almost" non-increasing;*
4. *$\left(\sum_{t=1}^T \alpha_t \sqrt{1-\theta_t}\right)/\left(T\alpha_T\right) = o(1)$.*

3.1. Convergence Rate of Generic Adam

We now provide the convergence rate of Generic Adam with a specific class of parameters $\{(\theta_t, \alpha_t)\}$, *i.e.*,

$$\alpha_t = \eta/t^s \text{ and } \theta_t = \begin{cases} 1 - \alpha/K^r, & t < K, \\ 1 - \alpha/t^r, & t \ge K, \end{cases} \quad (8)$$

for positive constants α, η, K, where K is taken such that $\alpha/K^r < 1$. Note that α can be taken bigger than 1. When $\alpha < 1$, we can take $K = 1$ and then $\theta_t = 1 - \alpha/t^r, t \ge 1$. To guarantee (R3), we require $r \le 2s$. For such a family of parameters we have the following corollary.

Corollary 10. *Generic Adam with the above family of parameters converges as long as $0 < r \le 2s < 2$, and its non-asymptotic convergence rate is given by*

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|^2 \le \begin{cases} \mathcal{O}(T^{-r/2}), & r/2 + s < 1 \\ \mathcal{O}(\log(T)/T^{1-s}), & r/2 + s = 1 \\ \mathcal{O}(1/T^{1-s}), & r/2 + s > 1 \end{cases}.$$

Remark 11. *Corollary 10 recovers and extends the results of some well-known algorithms below:*

- ***AdaGrad with exponential moving average (EMA).*** *When $\theta_t = 1 - 1/t$, $\alpha_t = \eta/\sqrt{t}$, and $\beta_t = \beta < 1$, Generic Adam is exactly AdaGrad with EMA momentum (AdaEMA) [6]. In particular, if $\beta = 0$, this is the vanilla coordinate-wise AdaGrad. It corresponds to taking $r = 1$ and $s = 1/2$ in Corollary 10. Hence, AdaEMA has convergence rate $\log(T)/\sqrt{T}$.*

- ***AdamNC.*** *Taking $\theta_t = 1 - 1/t$, $\alpha_t = \eta/\sqrt{t}$, and $\beta_t = \beta\lambda^t$ in Generic Adam, where $\lambda < 1$ is the decay factor for the momentum factors β_t, we recover AdamNC [20]. Its $\mathcal{O}(\log(T)/\sqrt{T})$ convergence rate can be directly derived via Corollary 10.*

- ***RMSProp.*** *Mukkamala and Hein [19] have reached the same $\mathcal{O}(\log(T)/\sqrt{T})$ convergence rate for RMSprop with $\theta_t = 1 - \alpha/t$, when $0 < \alpha \le 1$ and $\alpha_t = \eta/\sqrt{t}$ under the convex assumption. Since RMSProp is essentially Generic Adam with all momentum factors $\beta_t = 0$, we recover Mukkamala and Hein's results by taking $r = 1$ and $s = 1/2$ in Corollary 10. Moreover, our result generalizes to the non-convex stochastic setting, and it holds for all $\alpha > 0$ rather than only $0 < \alpha \le 1$.*

As Weighted AdaEMA is equivalent to Generic Adam, we present its convergence rate with specific polynomial growth weights in the following corollary.

Corollary 12. *Suppose in Weighted AdaEMA the weights $w_t = t^r$ for $r \ge 0$, and $\alpha_t = \eta/\sqrt{t}$. Then Weighted AdaEMA has the $\mathcal{O}(\log(T)/\sqrt{T})$ non-asymptotic convergence rate.*

Remark 13. *Zou et al. [26] proposed weighted AdaGrad with a unified momentum form which incorporates Heavy Ball (HB) momentum and Nesterov Accelerated Gradients (NAG) momentum. The same convergence rate was established for weights with polynomial growth. Our result complements [26] by showing that the same convergence rate also holds for exponential moving average momentum.*

Remark 14. (i) *Huang et al. [12] proposed Nostalgic Adam (NosAdam) which corresponds to taking the learning rate $\alpha_t = \eta/\sqrt{t}$ and $\theta_t = B_{t-1}/B_t$ with $B_t = \sum_{i=1}^t b_i$ for $b_i > 0$, $i \ge 0$, and $B_0 > 0$[2] in Generic Adam. The idea of NosAdam is to guarantee $\Gamma_t \succ 0$ by laying more weights on the historical second-order moments. A special case of NosAdam is NosAdam-HH which takes $B_t = \sum_{i=1}^t i^{-r}$ for $r \ge 0$ as the hyper-harmonic series. Its $\mathcal{O}(1/\sqrt{T})$ convergence rate is established in the strongly convex stochastic setting. NosAdam-HH can be viewed as the Weighted AdaEMA taking $\alpha_t = \eta/\sqrt{t}$ and $w_t = t^{-r}$ for $r \ge 0$.*

---

[2]We directly use B_t and b_i along with the notations of NosAdam [12].

(ii) *Corollary 12 differs from the motivation of NosAdam as the weights we consider are $w_t = t^r$ for $r \geq 0$. Note that in both cases when $r = 0$, this is the AdaGrad algorithm, which corresponds to assigning equal weights to the past squares of gradients. Hence, we are actually in the opposite direction of NosAdam. We are more interested in the case of assigning more weights to the recent stochastic gradients. This can actually be viewed as a situation between AdaGrad and the original Adam with constant θ_t's.*

Comparison between (SC) and (C1). Most of the convergent modifications of original Adam, such as AMSGrad, AdamNC, and NosAdam, all require $\Gamma_t \succ 0$ in Eq. (5), which is equivalent to decreasing the adaptive learning rate η_t step by step. Since the term Γ_t (or adaptive learning rate η_t) involves the past stochastic gradients (hence not deterministic), the modification to guarantee $\Gamma_t \succ 0$ either needs to change the iteration scheme of Adam (like AMS-Grad) or needs to impose some strong restrictions on the base learning rate α_t and θ_t (like AdamNC). Our sufficient condition provides an easy-to-check criterion for the convergence of Generic Adam in Corollary 9. It is not necessary to require $\Gamma_t \succ 0$. Moreover, we use exactly the same iteration scheme as original Adam without any modifications. Our work shows that the positive definiteness of Γ_t may not be the essential issue for divergence of original Adam. It is probably due to that the parameters are not set correctly.

4. Constant θ_t case: insights for divergence

The currently most popular RMSProp and Adam's parameter setting takes constant θ_t, *i.e.*, $\theta_t = \theta < 1$. The motivation behind is to use the exponential moving average of squares of past stochastic gradients. In practice, parameter θ is recommended to be set very close to 1. For instance, a commonly adopted θ is taken as 0.999.

Although great performance in practice has been observed, such a constant parameter setting has the serious flaw that there is no convergence guarantee even for convex optimization, as proved by the counterexamples in [20]. Ever since much work has been done to analyze the divergence issue of Adam and to propose modifications with convergence guarantees, as summarized in the introduction section. However, there is still not a satisfactory explanation that touches the fundamental reason of the divergence. In this section, we try to provide more insights for the divergence issue of Adam/RMSProp with constant parameter θ_t, based on our analysis of the sufficient condition for convergence.

From the sufficient condition perspective. Let $\alpha_t = \eta/t^s$ for $0 \leq s < 1$ and $\theta_t = \theta < 1$. According to Corollary 7, $Bound(T)$ in Theorem 5 has the following estimate:

$$\frac{C}{\delta \eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta} \leq Bound(T) \leq \frac{C}{\delta \eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{\delta(1-s)}.$$

The bounds tell us some points on Adam with constant θ_t:

1. $Bound(T) = \mathcal{O}(1)$, so the convergence is not guaranteed. This result coincides with the divergence issue demonstrated in [20]. Indeed, since in this case Adam is not convergent, this is the best bound we can have.

2. Consider the dependence on parameter s. The bound is decreasing in s. The best bound in this case is when $s = 0$, *i.e.*, the base learning rate is taken constant. This explains why in practice taking a more aggressive constant base learning rate often leads to even better performance, comparing with taking a decaying one.

3. Consider the dependence on parameter θ. Note that the constants C and C' depend on θ_1 instead of the whole sequence θ_t. We can always set $\theta_t = \theta$ for $t \geq 2$ while fix $\theta_1 < \theta$, by which we can take C and C' independent of constant θ. Then the principal term of $Bound(T)$ is linear in $\sqrt{1-\theta}$, so decreases to zero as $\theta \to 1$. This explains why setting θ close to 1 in practice often results in better performance in practice.

Moreover, Corollary 10 shows us how the convergence rate continuously changes when we continuously verify parameters θ_t. Let us fix $\alpha_t = 1/\sqrt{t}$ and consider the following continuous family of parameters $\{\theta_t^{(r)}\}$ with $r \in [0,1]$:

$$\theta_t^{(r)} = 1 - \alpha^{(r)}/t^r, \quad \text{where} \quad \alpha^{(r)} = r\bar{\theta} + (1-\bar{\theta}), \ 0 < \bar{\theta} < 1.$$

Note that when $r = 1$, then $\theta_t = 1 - 1/t$, this is the AdaEMA, which has the convergence rate $\mathcal{O}(\log T/\sqrt{T})$; when $r = 0$, then $\theta_t = \bar{\theta} < 1$, this is the original Adam with constant θ_t, which only has the $\mathcal{O}(1)$ bound; when $0 < r < 1$, by Corollary 10, the algorithm has the $\mathcal{O}(T^{-r/2})$ convergence rate. Along this continuous family of parameters, we observe that the theoretical convergence rate continuously deteriorates as the real parameter r decreases from 1 to 0, namely, as we gradually shift from AdaEMA to Adam with constant θ_t. In the limiting case, the latter is not guaranteed with convergence any more. This phenomenon is empirically verified by the Synthetic Counterexample in Section 5.

From the Weighted AdaEMA perspective. Since Generic Adam is equivalent to Weighted AdaEMA, we can examine Adam with $\theta_t = \theta < 1$ in terms of Weighted AdaEMA. In this case, we find that the associated sequence of weights $w_t = (1-\theta)\theta^{-t}$ is growing in an exponential order. Corollary 12 shows that as long as the weights are in polynomial growth, Weighted AdaEMA is convergent and its convergence rate is $\mathcal{O}(\log T/\sqrt{T})$. This indicates that the exponential-moving-average technique in the estimate of second-order moments may assign a too aggressive weight to the current gradient, which leads to the divergence.
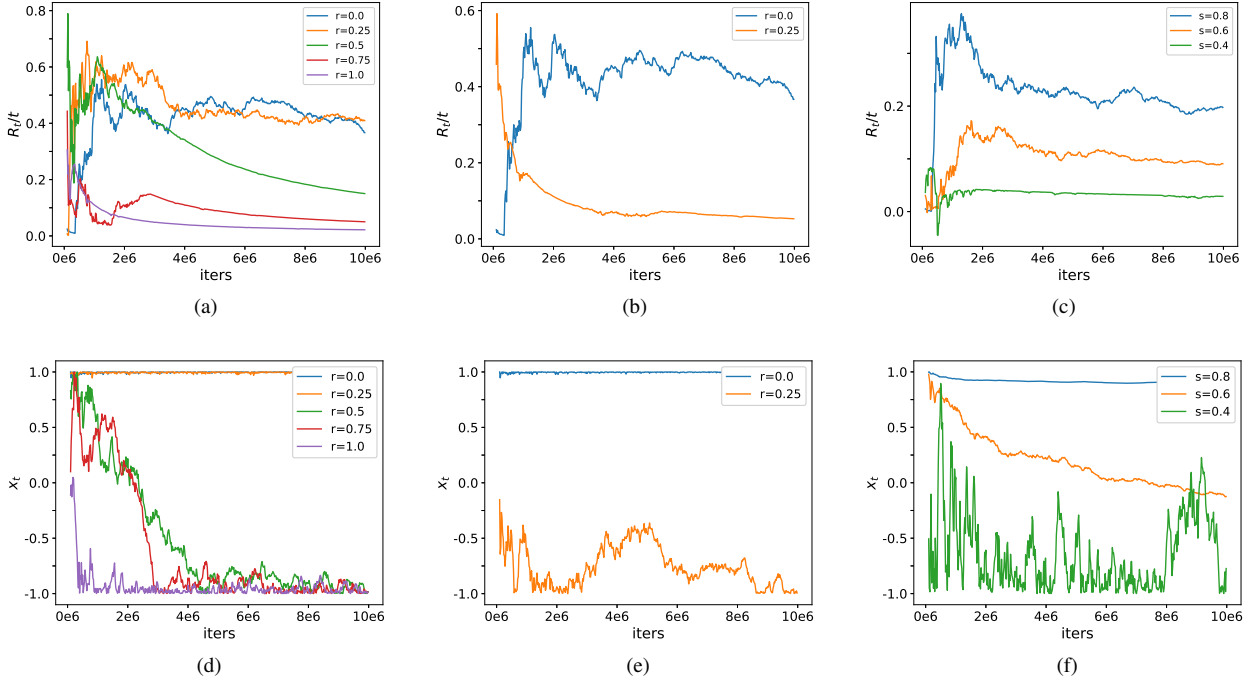
Figure 1. The above figures are for average regret and x values with different r and s values, respectively. Figures (a) and (d) plot the performance profiles of Generic Adam with different r values. Figures (b) and (e) plot the performance profiles of Generic Adam with $\theta_t^{(r)} = 1 - \frac{0.01}{t^r}$ and $r = 0$ and 0.25. Figures (c) and (f) plot the performance profiles of Generic Adam with different s values.

5. Experiments

In this section, we experimentally validate the proposed sufficient condition by applying Generic Adam and RM-SProp to solve the counterexample [20] and to train LeNet [16] on the MNIST dataset [17] and ResNet [10] on the CIFAR-100 dataset [15], respectively.

5.1. Synthetic Counterexample

In this experiment, we verify the phenomenon described in Section 4 that how the convergence rate of Generic Adam gradually changes along a continuous path of families of parameters on the one-dimensional counterexample in [20]:

$$\mathcal{R}(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in [-1,1]} \sum_{t=1}^{T} f_t(x), \qquad (9)$$

where T is the number of maximum iterations, $f_t(x) = 1010x$ with probability 0.01, and $f_t(x) = 10x$ with probability 0.99.

Sensitivity of parameter r. We set $T = 10^7$, $\alpha_t = 0.5/\sqrt{t}$, $\beta = 0.9$, and θ_t as $\theta_t^{(r)} = 1 - (0.01 + 0.99r)/t^r$ with $r \in \{0, 0.25, 0.5, 0.75, 1.0\}$, respectively. Note that when $r = 0$, Generic Adam reduces to the originally divergent Adam [14] with $(\beta, \bar{\theta}) = (0.9, 0.99)$. When $r = 1$, Generic Adam reduces to the AdaEMA [6] with $\beta = 0.9$.

The experimental results are shown in Figures 1(a) and 1(d). We can see that for $r = 1.0, 0.75$, and 0.5, Generic

Adam is convergent. Moreover, the convergence becomes slower when r decreases, which exactly matches Corollary 10. On the other hand, for $r = 0$ and 0.25, Figure 1(d) shows that they do not converge. It seems that the divergence for $r = 0.25$ contradicts our theory. However, this is because when r is very small, the $\mathcal{O}(T^{-r/2})$ convergence rate is so slow that we may not see a convergent trend in even 10^7 iterations. Indeed, for $r = 0.25$, we actually have

$$\theta_t^{(0.25)} \le 1 - (0.01 + 0.25 * 0.99)/10^{7*0.25} \approx 0.9954,$$

which is not sufficiently close to 1. As a complementary experiment, we fix the numerator and only change r when r is small. We take α_t and β_t as the same, while $\theta_t^{(r)} = 1 - \frac{0.01}{t^r}$ for $r = 0$ and 0.25, respectively. The result is shown in Figures 1(b) and 1(e). We can see that Generic Adam with $r = 0.25$ is indeed convergent in this situation.

Sensitivity of parameter s. Now, we show the sensitivity of s of the sufficient condition (**SC**) by fixing $r = 0.8$ and selecting s from the collection $s = \{0.4, 0.6, 0.8\}$. Figures 1(c) and 1(e) illustrate the sensitivity of parameter s when Generic Adam is applied to solve the counterexample (9). The performance shows that when s is fixed, smaller r can lead to a faster and better convergence speed, which also coincides with the convergence results in Corollary 10.
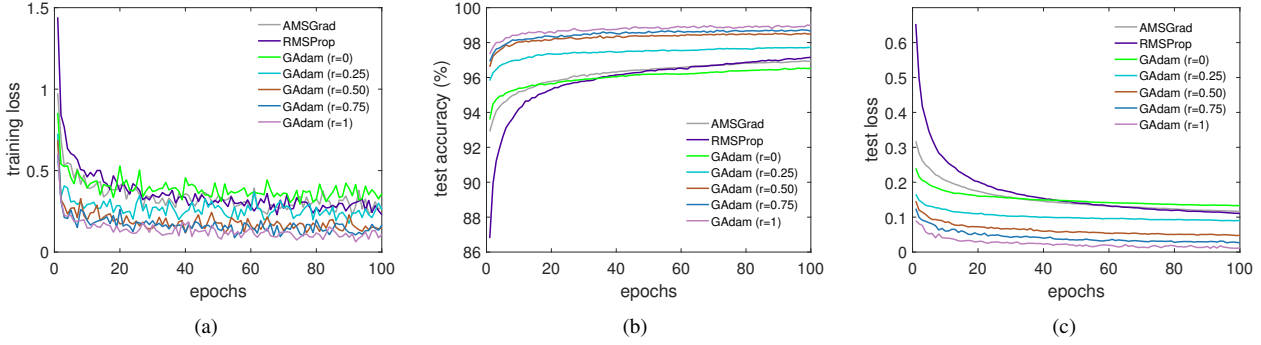
Figure 2. Performance profiles of Generic Adam with $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad for training LeNet on the MNIST dataset. Figures (a), (b), and (c) illustrate training loss vs. epochs, test accuracy vs. epochs, and test loss vs. epochs, respectively.
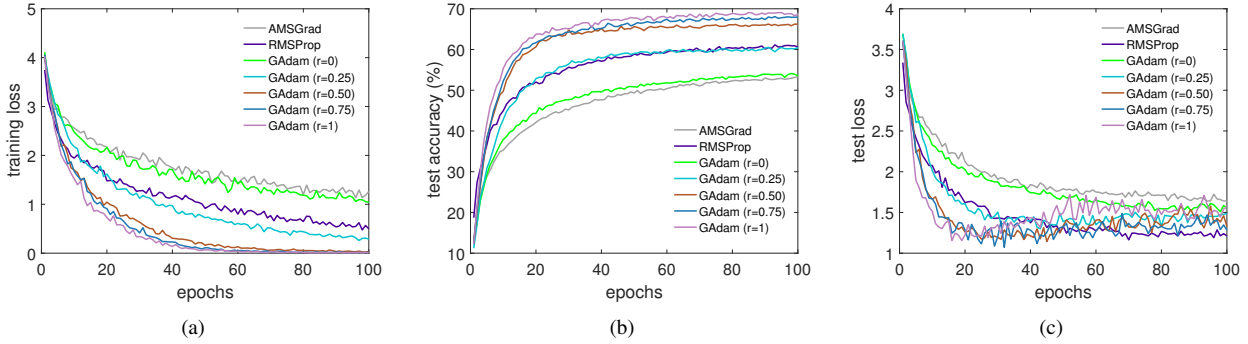

Figure 3. Performance profiles of Generic Adam with $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad for training ResNet on the CIFAR-100 dataset. Figures (a), (b), and (c) illustrate training loss vs. epochs, test accuracy vs. epochs, and test loss vs. epochs, respectively.

5.2. LeNet on MNIST and ResNet-18 on CIFAR-100

In this subsection, we apply Generic Adam to train LeNet on the MNIST dataset and ResNet-18 on the CIFAR-100 dataset, respectively, in order to validate the convergence rates in Corollary 10. Meanwhile, the comparisons between Generic Adam and AMSGrad [20, 6] are also provided to distinguish their differences in training deep neural networks. We illustrate the performance profiles in three aspects: training loss vs. epochs, test loss vs. epochs, and test accuracy vs. epochs, respectively. Besides, the architectures of LeNet and ResNet-18, and the statistics of the MNIST and CIFAR-100 datasets are described in the **supplementary material**.

In the experiments, for Generic Adam, we set $\theta_t^{(r)} = 1 - (0.001 + 0.999r)/t^r$ with $r \in \{0, 0.25, 0.5, 0.75, 1\}$ and $\beta_t = 0.9$, respectively; for RMSProp, we set $\beta_t = 0$ and $\theta_t = 1 - \frac{1}{t}$ along with the parameter settings in [19]. For fairness, the base learning rates α_t in Generic Adam, RMSProp, and AMSGrad are all set as $0.001/\sqrt{t}$. Figures 2 and 3 illustrate the results of Generic Adam with different r, RMSProp, and AMSGrad for training LeNet on MNIST and training ResNet-18 on CIFAR-100, respectively. We can see that AMSGrad and Adam (Generic Adam with $r = 0$) decrease the training loss slowest and show the worst test accuracy among the compared optimizers. One possible

reason is due to the use of constant θ in AMSGrad and original Adam. By Figures 2 and 3, we can observe that the convergences of Generic Adam are extremely sensitive to the choice of parameter θ_t. Larger r can contribute to a faster convergence rate of Generic Adam, which corroborates the theoretical result in Corollary 10. Additionally, the test accuracies in Figures 2(b) and 3(b) indicate that a smaller training loss can contribute to a higher test accuracy for Generic Adam.

6. Conclusions

In this work, we delved into the convergences of Adam and RMSProp, and presented an easy-to-check sufficient condition to guarantee their convergences in the non-convex stochastic setting. This sufficient condition merely depends on the base learning rate α_t and the linear combination parameter θ_t of second-order moments. Relying on this sufficient condition, we found that the divergences of Adam and RMSProp are possibly due to the incorrect parameter settings of α_t and θ_t. In addition, we reformulated Adam as weighted AdaGrad with exponential moving average momentum, which provides a novel perspective for understanding Adam and RMSProp. At last, the correctness of theoretical results was also verified via the counterexample in [20] and training deep neural networks on real-world datasets.

References

[1] Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413, 2018. 2

[2] Amitabh Basu, Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to nesterov acceleration on autoencoders. *arXiv preprint arXiv:1807.06766*, 2018. 2

[3] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569, 2018. 2

[4] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999. 1

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. 1

[6] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018. 1, 2, 4, 5, 7, 8

[7] Zaiyi Chen, Tianbao Yang, Jinfeng Yi, Bowen Zhou, and Enhong Chen. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *arXiv preprint arXiv:1808.06296*, 2018. 2

[8] Timothy Dozat. Incorporating Nesterov momentum into Adam. *International Conference on Learning Representations Workshop*, 2016. 1

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 7

[11] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. page 14, 2012. 1

[12] Haiwen Huang, Chang Wang, and Bin Dong. Nostalgic Adam: Weighing more of the past gradients when designing the adaptive learning rate. *arXiv preprint arXiv:1805.07557*, 2018. 2, 3, 5

[13] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 3, 4, 7

[15] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 7

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 7

[17] Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. 2010. *URL http://yann. lecun. com/exdb/mnist*, 2010. 7

[18] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv preprint arXiv:1805.08114*, 2018. 2

[19] Mahesh Chandra Mukkamala and Matthias Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In *International Conference on Machine Learning*, pages 2545–2553, 2017. 2, 3, 5, 8

[20] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 5, 6, 7, 8

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985. 1

[22] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018. 2

[23] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018. 2

[24] Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018. 2

[25] Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. AdaShift: Decorrelation and convergence of adaptive learning rate methods. *arXiv preprint arXiv:1810.00143*, 2018. 2, 3

[26] Fangyu Zou and Li Shen. On the convergence of AdaGrad with momentum for training deep neural networks. *arXiv preprint arXiv:1808.03408*, 2018. 2, 3, 5