IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2019, accepted May 4, 2019, date of publication May 13, 2019, date of current version May 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916341

On the Convergence Proof of AMSGrad and a New Version

TRAN THI PHUONG[1,2,3] **AND LE TRIEU PHONG**[3]

[1]Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[2]Meiji University, Kawasaki 214-8571, Japan
[3]National Institute of Information and Communications Technology (NICT), Tokyo 184-8795, Japan

Corresponding author: Tran Thi Phuong (tranthiphuong@tdtu.edu.vn)

ABSTRACT The adaptive moment estimation algorithm Adam (Kingma and Ba) is a popular optimizer in the training of deep neural networks. However, Reddi *et al.* have recently shown that the convergence proof of Adam is problematic, and they have also proposed a variant of Adam called AMSGrad as a fix. In this paper, we show that the convergence proof of AMSGrad is also problematic. Concretely, the problem in the convergence proof of AMSGrad is in handling the hyper-parameters, treating them as equal while they are not. This is also the neglected issue in the convergence proof of Adam. We provide an explicit counter-example of a simple convex optimization setting to show this neglected issue. Depending on manipulating the hyper-parameters, we present various fixes for this issue. We provide a new convergence proof for AMSGrad as the first fix. We also propose a new version of AMSGrad called AdamX as another fix. Our experiments on the benchmark dataset also support our theoretical results.

INDEX TERMS Optimizer, adaptive moment estimation, Adam, AMSGrad, deep neural networks.

I. INTRODUCTION AND OUR CONTRIBUTIONS

One of the most popular algorithms for training deep neural networks is stochastic gradient descent (SGD) [1] and its variants. Among the various variants of SGD, the algorithm with the adaptive moment estimation Adam [2] is widely used in practice. However, Reddi *et al.* [3] have recently shown that the convergence proof of Adam is problematic and proposed a variant of Adam called AMSGrad to solve this issue.

Our contribution. In this paper, we point out a flaw in the convergence proof of AMSGrad, recalled as Theorem A below. We then fix this flaw by providing a new convergence proof for AMSGrad in the case of special parameters. In addition, in the case of general parameters, we propose a new and slightly modified version of AMSGrad.

To provide more details, let us recall AMSGrad in Algorithm 1, in which the mathematical notation can be fully found in Section II.

The main theorem for the convergence of AMSGrad in [3] is as follows. To simplify the notation, we define $g_t \triangleq \nabla f_t(x_t)$, $g_{t,i}$ as the i^{th} element of g_t and $g_{1:t,i} \in \mathbb{R}^t$ as a vector that contains the i^{th} dimension of the gradients over all iterations up to t, namely, $g_{1:t,i} = [g_{1,i}, g_{2,i}, \ldots, g_{t,i}]$.

The associate editor coordinating the review of this manuscript and approving it for publication was Yanbo Chen.

Algorithm 1 AMSGrad (Reddi *et al.* [3])

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t\}_{t=1}^T$, $\{\beta_{1,t}\}_{t=1}^T$, β_2
Set $m_0 = 0$, $v_0 = 0$, and $\hat{v}_0 = 0$
for $(t = 1; t \le T; t \leftarrow t + 1)$ **do**
  $g_t = \nabla f_t(x_t)$
  $m_t = \beta_{1,t} \cdot m_{t-1} + (1 - \beta_{1,t}) \cdot g_t$
  $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
  $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ and $\hat{V}_t = \text{diag}(\hat{v}_t)$
  $x_{t+1} = \prod_{\mathcal{F}, \sqrt{\hat{V}_t}}(x_t - \alpha_t \cdot m_t / \sqrt{\hat{v}_t})$
end for

Theorem A (Theorem 4 in [3], problematic): *Let x_t and v_t be the sequences obtained from Algorithm 1, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, $\beta_1 = \beta_{1,1}$, $\beta_{1,t} \le \beta_1$ for all $t \in [T]$ and $\frac{\beta_1}{\sqrt{\beta_2}} \le 1$. Assume that $\mathcal{F}$ has bounded diameter D_∞ and $\|\nabla f_t(x)\|_\infty \le G_\infty$ for all $t \in [T]$ and $x \in \mathcal{F}$. For x_t generated using AMSGrad (Algorithm 1), we have the following bound on the regret:*

$$R(T) \le \frac{D_\infty^2 \sqrt{T}}{\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{\hat{v}_{T,i}} + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t} \sqrt{\hat{v}_{t,i}}}{\alpha_t} + \frac{\alpha \sqrt{1 + \ln T}}{(1-\beta_1)^2 (1-\gamma) \sqrt{1-\beta_2}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2.$$

In their proof for Theorem A, Reddi *et al.* resolved an issue on the so-called *telescopic sum* in the convergence proof of Adam ([2, Theorem 10.5]). Specifically, Reddi et al. adjusted $\hat{v}_t$ such that

$$\frac{\sqrt{\hat{v}_{t+1,i}}}{\alpha_{t+1}} \geq \frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t} \tag{1}$$

for all $i \in [d]$. However, there is *another* issue (showed in Section III) in the convergence proof of Adam that AMSGrad unfortunately neglects. The issue affects both the correctness of Reddi et al.'s proof and the upper bound for the regret in Theorem A. To deal with the issue in a general way, we propose to modify Algorithm 1 such that

$$\frac{\sqrt{\hat{v}_{t+1,i}}}{\alpha_{t+1}(1-\boxed{\beta_{1,t+1}})} \geq \frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t(1-\boxed{\beta_{1,t}})}$$

for all $i \in [d]$. The differences with (1) are highlighted in the boxes for clarity.

Paper roadmap. We begin with preliminaries in Section II. We show where the proof of Theorem A becomes invalid in Section III. After that, we suggest two ways to resolve the issue in Sections IV and V.

II. PRELIMINARIES

Notation. Given a sequence of vectors $\{x_t\}_{1\leq t\leq T}(1 \leq T \in \mathbb{N})$ in $\mathbb{R}^d$, we denote its i^{th} coordinate by $x_{t,i}$ and use x_t^k to denote the elementwise power of k and $\|x_t\|_2$, resp. $\|x_t\|_\infty$, to denote its ℓ_2-norm, resp. ℓ_∞-norm. Let $\mathcal{F} \subseteq \mathbb{R}^d$ be a feasible set of points such that $\mathcal{F}$ has bounded diameter D_∞, that is, $\|x-y\|_\infty \leq D_\infty$ for all $x, y \in \mathcal{F}$, and $\mathcal{S}_+^d$ denote the set of all positive definite $d \times d$ matrices. For a matrix $A \in \mathcal{S}_+^d$, we denote $A^{1/2}$ for the square root of A. The projection operation $\prod_{\mathcal{F},A}(y)$ for $A \in \mathcal{S}_+^d$ is defined as $\text{argmin}_{x\in\mathcal{F}}\|A^{1/2}(x-y)\|_2$ for all $y \in \mathbb{R}^d$. When $d = 1$ and $\mathcal{F} \subset \mathbb{R}$, the positive definite matrix A is a positive number, so that the projection $\prod_{\mathcal{F},A}(y)$ becomes $\text{argmin}_{x\in\mathcal{F}}|x-y|$. We use $\langle x, y\rangle$ to denote the inner product between x and $y \in \mathbb{R}^d$. The gradient of a function f evaluated at $x \in \mathbb{R}^d$ is denoted by $\nabla f(x)$. For vectors $x, y \in \mathbb{R}^d$, we use $\sqrt{x}$ or $x^{1/2}$ for element-wise square root, x^2 for element-wise square, x/y to denote element-wise division. For an integer $n \in \mathbb{N}$, we denote by $[n]$ the set of integers $\{1, 2, \ldots, n\}$.

Optimization setup. Let $f_1, f_2, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary sequence of convex cost functions and $x_1 \in \mathbb{R}^d$. At each time $t \geq 1$, the goal is to predict the parameter x_t and evaluate it on a previously unknown cost function f_t. Since the nature of the sequence is unknown in advance, the algorithm is evaluated by using the regret, that is, the sum of all the previous differences between the online prediction $f_t(x_t)$ and the best fixed-point parameter $f_t(x^*)$ from a feasible set $\mathcal{F}$ for all the previous steps. Concretely, the regret is defined as

$$R(T) = \sum_{t=1}^{T}[f_t(x_t) - f_t(x^*)],$$

where $x^* = \text{argmin}_{x\in\mathcal{F}}\sum_{t=1}^{T} f_t(x)$.

Definition 2.1: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^d$, and all $\lambda \in [0, 1]$,

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)y).$$

Lemma 2.2: If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then for all $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y-x),$$

where $\nabla f(x)^{\mathsf{T}}$ denotes the transpose of $\nabla f(x)$.

*Lemma 2.3 (**Cauchy–Schwarz inequality**): For all $n \geq 1$, $u_i, v_i \in \mathbb{R}(1 \leq i \leq n)$,*

$$\left(\sum_{i=1}^{n} u_i v_i\right)^2 \leq \left(\sum_{i=1}^{n} u_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right).$$

*Lemma 2.4 (**Taylor series**): For $\alpha \in \mathbb{R}$ and $0 < \alpha < 1$,*

$$\sum_{t\geq 1} \alpha^t = \frac{1}{1-\alpha}$$

and

$$\sum_{t\geq 1} t\alpha^{t-1} = \frac{1}{(1-\alpha)^2}.$$

*Lemma 2.5 (**Upper bound for the harmonic series**): For $N \in \mathbb{N}$,*

$$\sum_{n=1}^{N} \frac{1}{n} \leq \ln N + 1.$$

Lemma 2.6: For $N \in \mathbb{N}$,

$$\sum_{n=1}^{N} \frac{1}{\sqrt{n}} \leq 2\sqrt{N}.$$

Lemma 2.7: For all $n \in \mathbb{N}$ and $a_i, b_i \in \mathbb{R}$ such that $a_i \geq 0$ and $b_i > 0$ for all $i \in [n]$,

$$\frac{\sum_{i=1}^{n} a_i}{\sum_{j=1}^{n} b_j} \leq \sum_{i=1}^{n} \frac{a_i}{b_i}.$$

Lemma 2.8: [4, Lemma 3 in arXiv version] For any $Q \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \subseteq \mathbb{R}^d$, suppose $u_1 = \text{argmin}_{x\in\mathcal{F}}\|Q^{1/2}(x-z_1)\|$ and $u_2 = \text{argmin}_{x\in\mathcal{F}}\|Q^{1/2}(x-z_2)\|$. Then, we have

$$\|Q^{1/2}(u_1-u_2)\| \leq \|Q^{1/2}(z_1-z_2)\|.$$

III. ISSUE IN THE CONVERGENCE PROOF OF AMSGRAD

Before showing the issue in the convergence proof of AMSGrad, let us recall and prove the following inequality, which also appears in [3].

Lemma 3.1: Algorithm 1 achieves the following guarantee, for all $T \geq 1$:

$$\begin{aligned} R(T) \leq & \sum_{i=1}^{d}\sum_{t=1}^{T} \frac{\sqrt{\hat{v}_{t,i}}((x_{t,i}-x_i^*)^2 - (x_{t+1,i}-x_i^*)^2)}{2\alpha_t(1-\beta_{1,t})} \\ & + \sum_{i=1}^{d}\sum_{t=1}^{T} \frac{\alpha_t}{1-\beta_1}\frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\ & + \sum_{i=1}^{d}\sum_{t=2}^{T} \frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_1)}(x_{t,i}-x_i^*)^2. \end{aligned}$$

Proof: We note that

$$x_{t+1} = \prod_{\mathcal{F},\sqrt{\hat{V}_t}} (x_t - \alpha_t \cdot \hat{V}_t^{-1/2} m_t) = \text{argmin}_{x\in\mathcal{F}} \|\hat{V}^{1/4}(x - (x_t - \alpha_t \hat{V}^{-1/2} m_t))\|$$

and $\prod_{\mathcal{F},\sqrt{\hat{V}_t}}(x^*) = x^*$ for all $x^* \in \mathcal{F}$. For all $1 \le t \le T$, put $g_t = \nabla_x f_t(x_t)$. Using Lemma 2.8 with $u_1 = x_{t+1}$ and $u_2 = x^*$, we have

$$\begin{aligned}
&\|\hat{V}^{1/4}(x_{t+1} - x^*)\|^2 \\
&\le \|\hat{V}^{1/4}(x_t - \alpha_t \hat{V}^{-1/2} m_t - x^*)\|^2 \\
&= \|\hat{V}^{1/4}(x_t - x^*)\|^2 + \alpha_t^2 \|\hat{V}^{-1/4} m_t\|^2 - 2\alpha_t \langle m_t, x_t - x^*\rangle \\
&= \|\hat{V}^{1/4}(x_t - x^*)\|^2 + \alpha_t^2 \|\hat{V}^{-1/4} m_t\|^2 \\
&\quad - 2\alpha_t \langle \beta_{1,t} m_{t-1} + (1-\beta_{1,t}) g_t, x_t - x^*\rangle.
\end{aligned}$$

This yields

$$\begin{aligned}
\langle g_t, x_t - x^*\rangle \le{}& \frac{1}{2\alpha_t(1-\beta_{1,t})}\Big[\|\hat{V}^{1/4}(x_t - x^*)\|^2 \\
&- \|\hat{V}^{1/4}(x_{t+1} - x^*)\|^2\Big] \\
&+ \frac{\alpha_t}{2(1-\beta_{1,t})} \|\hat{V}^{-1/4} m_t\|^2 \\
&- \frac{\beta_{1,t}}{1-\beta_{1,t}} \langle m_{t-1}, x_t - x^*\rangle.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\sum_{i=1}^d g_{t,i}(x_{t,i} - x_i^*) \\
&\le \sum_{i=1}^d \frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}\Big((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2\Big) \\
&\quad + \sum_{i=1}^d \frac{\alpha_t}{2(1-\beta_{1,t})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} - \sum_{i=1}^d \frac{\beta_{1,t}}{1-\beta_{1,t}} m_{t-1,i}(x_{t,i} - x_i^*).
\end{aligned} \tag{2}$$

Moreover, by Lemma 2.2, we have $f_t(x^*) - f_t(x_t) \ge g_t^{\mathrm{T}}(x^* - x_t)$, where g_t^{T} denotes the transpose of vector g_t. This means that

$$f_t(x_t) - f_t(x^*) \le g_t^{\mathrm{T}}(x_t - x^*) = \sum_{i=1}^d g_{t,i}(x_{t,i} - x_i^*).$$

Hence,

$$\begin{aligned}
R(T) &= \sum_{t=1}^T [f_t(x_t) - f_t(x^*)] \\
&\le \sum_{t=1}^T g_t^{\mathrm{T}}(x_t - x^*) = \sum_{t=1}^T \sum_{i=1}^d g_{t,i}(x_{t,i} - x_i^*).
\end{aligned} \tag{3}$$

Combining (2) with (3), we obtain

$$\begin{aligned}
R(T) \le{}& \sum_{i=1}^d \sum_{t=1}^T \frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2) \\
&+ \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{2(1-\beta_{1,t})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\
&+ \sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}}{1-\beta_{1,t}} m_{t-1,i}(x_i^* - x_{t,i}).
\end{aligned}$$

where the last term is from the setting that $m_0 = 0$. On the other hand, for all $t \ge 2$, we have

$$\begin{aligned}
m_{t-1,t}(x_i^* - x_{t,i}) &= \frac{(\hat{v}_{t-1,i})^{1/4}}{\sqrt{\alpha_{t-1}}}(x_i^* - x_{t,i})\sqrt{\alpha_{t-1}} \frac{m_{t-1,i}}{(\hat{v}_{t-1,i})^{1/4}} \\
&\le \frac{\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}}(x_{t,i} - x_i^*)^2 + \alpha_{t-1} \frac{m_{t-1,i}^2}{2\sqrt{\hat{v}_{t-1,i}}},
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
R(T) \le{}& \sum_{i=1}^d \sum_{t=1}^T \frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}\Big((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2\Big) \\
&+ \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{2(1-\beta_{1,t})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\
&+ \sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}\alpha_{t-1}}{2(1-\beta_{1,t})} \frac{m_{t-1,i}^2}{\sqrt{\hat{v}_{t-1,i}}} \\
&+ \sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_{1,t})}(x_{t,i} - x_i^*)^2.
\end{aligned} \tag{4}$$

Moreover, we have

$$\begin{aligned}
\sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}\alpha_{t-1}}{2(1-\beta_{1,t})} \frac{m_{t-1,i}^2}{\sqrt{\hat{v}_{t-1,i}}} &= \sum_{i=1}^d \sum_{t=1}^{T-1} \frac{\beta_{1,t+1}\alpha_t}{2(1-\beta_{1,t+1})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\
&\le \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{2(1-\beta_{1,t+1})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\
&\le \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{2(1-\beta_1)} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}},
\end{aligned}$$

where the last inequality is from the assumption that $\beta_{1,t} \le \beta_1 < 1 (1 \le t \le T)$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{2(1-\beta_{1,t})} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} + \sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}\alpha_{t-1}}{2(1-\beta_{1,t})} \frac{m_{t-1,i}^2}{\sqrt{\hat{v}_{t-1,i}}} \\
&\quad \le \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{1-\beta_1} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}}.
\end{aligned} \tag{5}$$

Hence, from (4) and (5) we have

$$\begin{aligned}
R(T) \le{}& \sum_{i=1}^d \sum_{t=1}^T \frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}\Big((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2\Big) \\
&+ \sum_{i=1}^d \sum_{t=1}^T \frac{\alpha_t}{1-\beta_1} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \\
&+ \sum_{i=1}^d \sum_{t=2}^T \frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_1)}(x_{t,i} - x_i^*)^2,
\end{aligned}$$

where the last term is from the property that $\beta_{1,t} \le \beta_1$ $(1 \le t \le T)$. □

Issue in the convergence proof of AMSGrad. We denote the terms on the right hand-side of the upper bound for $R(T)$

in Lemma 3.1 as

$$\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}\left((x_{t,i}-x_i^*)^2-(x_{t+1,i}-x_i^*)^2\right), \quad (6)$$

$$\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\alpha_t}{1-\beta_1}\frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}}, \quad (7)$$

and

$$\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_1)}(x_{t,i}-x_i^*)^2. \quad (8)$$

The issue in the proof of the convergence theorem of AMSGrad [3, Theorem 4] becomes on examining the term (6). Indeed, in [3, page 18], Reddi *et al.* used[1] the property that $\beta_{1,t}\le\beta_1$, and hence

$$\frac{1}{1-\beta_{1,t}}\le\frac{1}{1-\beta_1},$$

to replace all $\beta_{1,t}$ by β_1 as

$$\begin{aligned}(6)&\le\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_1)}\left((x_{t,i}-x_i^*)^2-(x_{t+1,i}-x_i^*)^2\right)\\&\le\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_1)}(x_{1,i}-x_i^*)^2\\&\quad+\frac{1}{2(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t}-\frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}}\right).\end{aligned}$$

However, the first inequality (in red) is not guaranteed because the quantity

$$(x_{t,i}-x_i^*)^2-(x_{t+1,i}-x_i^*)^2$$

in (6) may be *both* negative *and* positive as shown in Counter-example 3.2. This is also a neglected issue in the convergence proofs in Kingma and Ba [2, Theorem 10.5], Luo *et al.* [5, Theorem 4], Bock *et al.* [6, Theorem 4.4], and Chen and Gu [7, Theorem 4.2].

Counter-example 3.2 (For AMSGrad Convergence Proof): We use the function in the Synthetic Experiment of Reddi *et al.* [3, Page 6]

$$f_t(x)=\begin{cases}1010x, & t \mod 101=1\\ -10x, & \text{otherwise}\end{cases}$$

with the constraint set $\mathcal{F}=[-1,1]$. The optimal solution is $x^*=-1$. By the proof of [3, Theorem 1], the initial point $x_1=1$. By Algorithm 1, $m_0=0$, $v_0=0$, and $\hat{v}_0=0$. We choose $\beta_1=0.9$, $\beta_{1,t}=\beta_1\lambda^{t-1}$, where $\lambda=0.001$, $\beta_2=0.999$, and $\alpha_t=\alpha/\sqrt{t}$, where $\alpha=0.001$. Under this setting, we have $f_1(x_1)=1010x_1$, $f_2(x_2)=-10x_2$, $f_3(x_3)=-10x_3$ and hence

$$g_1=\nabla f_1(x_1)=1010,$$

$$\begin{aligned}m_1&=\beta_{1,1}m_0+(1-\beta_{1,1})g_1=(1-0.9)1010=101,\\v_1&=\beta_2v_0+(1-\beta_2)g_1^2=(1-0.999)1010^2=1020.1,\\\hat{v}_1&=\max(\hat{v}_0,v_1)=v_1.\end{aligned}$$

Therefore,

$$\begin{aligned}x_1-\alpha_1\,m_1/\sqrt{\hat{v}_1}&=1-(0.001)101/\sqrt{1020.1}\\&=0.9968377223398316.\end{aligned}$$

Since $x_1-\alpha_1\,m_1/\sqrt{\hat{v}_1}>0$, we have

$$\begin{aligned}x_2&=\prod_{\mathcal{F}}(x_1-\alpha_1\,m_1/\sqrt{\hat{v}_1})\\&=\min(1,x_1-\alpha_1\,m_1/\sqrt{\hat{v}_1})\\&=0.9968377223398316.\end{aligned}$$

Hence,

$$(x_1-x^*)^2-(x_2-x^*)^2=0.001264811064067839>0.$$

At $t=2$, we have

$$\begin{aligned}g_2&=-10,\\m_2&=\beta_{1,2}m_1+(1-\beta_{1,2})g_2\\&=(0.9)(0.001)(101)+[1-(0.9)(0.001)](-10)\\&=-9.9001,\\v_2&=\beta_2v_1+(1-\beta_2)g_2^2\\&=(0.999)(1020.1)+(1-0.999)(-10)^2\\&=1019.1799000000001,\\\hat{v}_2&=\max(\hat{v}_1,v_2)=v_1\\&=1020.1.\end{aligned}$$

Therefore,

$$\begin{aligned}x_2-\alpha_2\,m_2/\sqrt{\hat{v}_2}&=0.9968377223398316-\frac{0.001}{\sqrt{2}}\frac{(-9.9001)}{\sqrt{1020.1}}\\&=0.9970569034941291.\end{aligned}$$

Since $x_2-\alpha_2\,m_2/\sqrt{\hat{v}_2}>0$, we obtain

$$\begin{aligned}x_3&=\prod_{\mathcal{F}}(x_2-\alpha_2\,m_2/\sqrt{\hat{v}_2})\\&=\min(1,x_2-\alpha_2\,m_2/\sqrt{\hat{v}_2})\\&=0.9970569034941291.\end{aligned}$$

Hence,

$$(x_2-x^*)^2-(x_3-x^*)^2=-0.0008753864342319062<0.$$

Outline of our solution. Let us rewrite (6) as

$$\begin{aligned}(6)&=\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_{1,1})}(x_{1,i}-x_i^*)^2\\&\quad+\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}(x_{t,i}-x_i^*)^2\\&\quad-\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_{1,t-1})}(x_{t,i}-x_i^*)^2\end{aligned}$$

[1] Concretely, on page 18 of [3], it is stated that "*The [...] inequality use the fact that* $\beta_{1,t}\le\beta_1$."

$$
\begin{aligned}
&- \sum_{i=1}^{d} \frac{\sqrt{\hat{v}_{T,i}}}{2\alpha_T(1-\beta_{1,T})}(x_{T+1,i}-x_i^*)^2 \\
&\le \sum_{i=1}^{d} \frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_{1,1})}(x_{1,i}-x_i^*)^2 \\
&\quad + \sum_{i=1}^{d}\sum_{t=2}^{T} \frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}(x_{t,i}-x_i^*)^2 \\
&\quad - \sum_{i=1}^{d}\sum_{t=2}^{T} \frac{\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_{1,t-1})}(x_{t,i}-x_i^*)^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
(6) &\le \sum_{i=1}^{d} \frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_{1,1})}(x_{1,i}-x_i^*)^2 \\
&\quad + \frac{1}{2}\sum_{i=1}^{d}\sum_{t=2}^{T}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t(1-\boxed{\beta_{1,t}})}\right. \\
&\quad \left. - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}(1-\boxed{\beta_{1,t-1}})}\right), \qquad (9)
\end{aligned}
$$

in which the differences with Reddi *et al.* [3] are highlighted in the boxes, namely, $\boxed{\beta_{1,t}}$ and $\boxed{\beta_{1,t-1}}$ instead of β_1.

We suggest two ways to overcome these differences depending on the setting of $\beta_{1,t}(1 \le t \le T)$:

- **In Section IV:** If either $\beta_{1,t} \triangleq \beta_1\lambda^{t-1}$ or $\beta_{1,t} \triangleq 1/t$, $(1 \le t \le T)$, where $0 \le \beta_1 < 1$ and $0 < \lambda < 1$, then we give a new convergence theorem for AMSGrad in Section IV.
- **In Section V:** If the setting for $\beta_{1,t}(1 \le t \le T)$ is general, as in the statement of Theorem A, then we suggest a new (slightly modified) version for AMSGrad in Section V.

IV. NEW CONVERGENCE THEOREM FOR AMSGRAD

When either $\beta_{1,t} \triangleq \beta_1\lambda^{t-1}$ or $\beta_{1,t} \triangleq 1/t (1 \le t \le T)$, where $0 \le \beta_1 < 1$ and $0 < \lambda < 1$, Theorem A can be fixed as follows, in which the upper bounds of the regret $R(T)$ are changed.

Theorem 4.1 (Fixes for Theorem A): Let x_t and v_t be the sequences obtained from Algorithm 1, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, either $\beta_{1,t} = \beta_1\lambda^{t-1}$, where $\lambda \in (0,1)$, or $\beta_{1,t} = \frac{\beta_1}{t}$ for all $t \in [T]$ and $\gamma = \frac{\beta_1}{\sqrt{\beta_2}} \le 1$. Assume that $\mathcal{F}$ has bounded diameter D_∞ and $\|\nabla f_t(x)\|_\infty \le G_\infty$ for all $t \in [T]$ and $x \in \mathcal{F}$. For x_t generated using AMSGrad (Algorithm 1), we have the following bound on the regret. Then, there is some $1 \le t_0 \le T$ such that AMSGrad achieves the following guarantee for all $T \ge 1$:

$$
\begin{aligned}
R(T) &\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\left(\sum_{t=1}^{t_0}\sqrt{t}+\sqrt{T}\right) \\
&\quad + \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)(1-\lambda)^2} \\
&\quad + \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2,
\end{aligned}
$$

provided $\beta_{1,t} = \beta_1\lambda^{t-1}$, and

$$
\begin{aligned}
R(T) &\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\left(\sum_{t=1}^{t_0}\sqrt{t}+\sqrt{T}\right) \\
&\quad + \frac{dD_\infty^2 G_\infty\sqrt{T}}{\alpha(1-\beta_1)} \\
&\quad + \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2,
\end{aligned}
$$

provided $\beta_{1,t} = \frac{\beta_1}{t}$.

To prove Theorem 4.1, we need the following Lemmas 4.2, 4.3, and 4.4.

Lemma 4.2: $\sqrt{\hat{v}_t} \le G_\infty$.

Proof: From the definition of $\hat{v}_t$ in AMSGrad's algorithm, it is implied that $\hat{v}_t = \max\{v_1,\ldots,v_t\}$. Therefore, there is some $1 \le s \le t$ such that $\hat{v}_t = v_s$. Hence,

$$
\begin{aligned}
\sqrt{\hat{v}_t} &= \sqrt{v_s} \\
&= \sqrt{1-\beta_2}\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}g_k^2} \\
&\le \sqrt{1-\beta_2}\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}(\max_{1\le j\le s}|g_j|)^2} \\
&= G_\infty\sqrt{1-\beta_2}\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}} \\
&\le G_\infty\sqrt{1-\beta_2}\frac{1}{\sqrt{1-\beta_2}} \\
&= G_\infty,
\end{aligned}
$$

where the first inequality is by the fact that $g_k \le \max_{1\le j\le s}|g_j|$, $k \in [s]$, and the last inequality is by Lemma 2.4. □

Lemma 4.3: If either $\beta_{1,t} = \beta_1\lambda^{t-1}$ or $\beta_{1,t} = \beta_1/t$, then there exists some t_0 such that for every $t > t_0$,

$$
\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} \ge \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}.
$$

Proof: Since $\hat{v}_{t,i} \ge \hat{v}_{t-1,i}$ owing to $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ in Algorithm 1, it is sufficient to prove that there exists some t_0 such that for every $t > t_0$,

$$
\frac{\sqrt{t}}{1-\beta_{1,t}} \ge \frac{\sqrt{t-1}}{1-\beta_{1,t-1}}.
$$

In other word,

$$
1 - \frac{\beta_{1,t-1}-\beta_{1,t}}{1-\beta_{1,t}} \ge \sqrt{1-\frac{1}{t}}. \qquad (10)
$$

When $\beta_{1,t} = \beta_1/t$, from (10) we have

$$
1 - \frac{\beta_1}{(t-1)(t-\beta_1)} \ge \sqrt{1-\frac{1}{t}}. \qquad (11)
$$

When $\beta_{1,t} = \beta_1\lambda^{t-1}$, (10) have the following form

$$
1 - \frac{(1-\lambda)\beta_1\lambda^{t-2}}{1-\beta_1\lambda^{t-1}} = \frac{1-\beta_1\lambda^{t-2}}{1-\beta_1\lambda^{t-1}} \ge \sqrt{1-\frac{1}{t}}. \qquad (12)
$$

Since β_1 and λ are smaller than 1, it is easy to see that when t is sufficiently large, meaning that $t > t_0$ for some t_0, the left-hand side of (11) is $1 - O(1/t^2)$ and the left-hand side of (12) is larger than $1 - \beta_1\lambda^{t-2} = 1 - O(\lambda^{t-2})$. Therefore, (11) and (12) hold when t is sufficiently large. □

Lemma 4.4: For the parameter settings and conditions assumed in Theorem 4.1, we have

$$\sum_{t=1}^{T} \frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \le \frac{\sqrt{\ln T + 1}}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)} \|g_{1:T,i}\|_2.$$

Proof: The proof is almost identical to that of [3, Lemma 2]. Owing to $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ in Algorithm 1, we have $\hat{v}_{t,i} \ge v_{t,i}$ for all $t \ge 1$. Therefore

$$\begin{aligned}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\le \frac{m_{t,i}^2}{\sqrt{tv_{t,i}}} \\ &= \frac{[\sum_{k=1}^{t}(1-\beta_{1,k})(\prod_{j=k+1}^{t}\beta_{1,j})g_{k,i}]^2}{\sqrt{(1-\beta_2)t\sum_{k=1}^{t}\beta_2^{t-k}g_{k,i}^2}}. \end{aligned} \tag{13}$$

Moreover, by Lemma 2.3 we have

$$\begin{aligned}&\left(\sum_{k=1}^{t}(1-\beta_{1,k})\Big(\prod_{j=k+1}^{t}\beta_{1,j}\Big)g_{k,i}\right)^2 \\ &\le \left(\sum_{k=1}^{t}(1-\beta_{1,k})^2\Big(\prod_{j=k+1}^{t}\beta_{1,j}\Big)\right)\left(\sum_{k=1}^{t}\Big(\prod_{j=k+1}^{t}\beta_{1,j}\Big)g_{k,i}^2\right).\end{aligned}$$

And hence,

$$\begin{aligned}&\left(\sum_{k=1}^{t}(1-\beta_{1,k})\Big(\prod_{j=k+1}^{t}\beta_{1,j}\Big)g_{k,i}\right)^2 \\ &\qquad\le \left(\sum_{k=1}^{t}\beta_1^{t-k}\right)\left(\sum_{k=1}^{t}\beta_1^{t-k}g_{k,i}^2\right),\end{aligned} \tag{14}$$

since $\beta_{1,k} \le 1$ and $\beta_{1,k} \le \beta_1$ for all $1 \le k \le T$. Combining (13) and (14) we obtain

$$\begin{aligned}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\le \frac{\left(\sum_{k=1}^{t}\beta_1^{t-k}\right)\left(\sum_{k=1}^{t}\beta_1^{t-k}g_{k,i}^2\right)}{\sqrt{(1-\beta_2)t\sum_{k=1}^{t}\beta_2^{t-k}g_{k,i}^2}} \\ &\le \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}}\frac{\sum_{k=1}^{t}\beta_1^{t-k}g_{k,i}^2}{\sqrt{t\sum_{k=1}^{t}\beta_2^{t-k}g_{k,i}^2}},\end{aligned}$$

where the last inequality is obtained by applying Lemma 2.4 to $\sum_{k=1}^{t}\beta_1^{t-k}$. Therefore,

$$\begin{aligned}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\le \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}\sqrt{t}}\frac{\sum_{k=1}^{t}\beta_1^{t-k}g_{k,i}^2}{\sqrt{\sum_{k=1}^{t}\beta_2^{t-k}g_{k,i}^2}} \\ &\le \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}\sqrt{t}}\sum_{k=1}^{t}\frac{\beta_1^{t-k}g_{k,i}^2}{\sqrt{\beta_2^{t-k}g_{k,i}^2}} \\ &\le \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}\sqrt{t}}\sum_{k=1}^{t}\frac{\beta_1^{t-k}}{\sqrt{\beta_2^{t-k}}}|g_{k,i}| \\ &= \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}|,\end{aligned}$$

where the second inequality is by Lemma 2.7. Therefore

$$\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \le \frac{1}{(1-\beta_1)\sqrt{1-\beta_2}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}|. \tag{15}$$

It is sufficient to consider $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}|$. Firstly, $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}|$ can be expanded as

$$\begin{aligned}&\frac{1}{\sqrt{2}}\Big(\gamma^1|g_{1,i}| + \gamma^0|g_{2,i}|\Big) \\ &\quad + \frac{1}{\sqrt{3}}\Big(\gamma^2|g_{1,i}| + \gamma^1|g_{2,i}| + \gamma^0|g_{3,i}|\Big) \\ &\quad + \cdots \\ &\quad + \frac{1}{\sqrt{T}}\Big(\gamma^{T-1}|g_{1,i}| + \gamma^{T-2}|g_{2,i}| + \ldots + \gamma^0|g_{T,i}|\Big).\end{aligned}$$

Changing the role of $|g_{1,i}|$ as the common factor, we have $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}|$ is equal to

$$\begin{aligned}&|g_{1,i}|\Big(\gamma^0 + \frac{1}{\sqrt{2}}\gamma^1 + \frac{1}{\sqrt{3}}\gamma^2 + \ldots + \frac{1}{\sqrt{T}}\gamma^{T-1})\Big) \\ &\quad + |g_{2,i}|\Big(\frac{1}{\sqrt{2}}\gamma^0 + \frac{1}{\sqrt{3}}\gamma^1 + \ldots + \frac{1}{\sqrt{T}}\gamma^{T-2}\Big) \\ &\quad + |g_{3,i}|\Big(\frac{1}{\sqrt{3}}\gamma^0 + \frac{1}{\sqrt{4}}\gamma^1 + \ldots + \frac{1}{\sqrt{T}}\gamma^{T-3}\Big) \\ &\quad + \cdots \\ &\quad + |g_{T,i}|\frac{1}{\sqrt{T}}\gamma^0.\end{aligned}$$

In other words,

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}| = \sum_{t=1}^{T}|g_{t,i}|\sum_{k=t}^{T}\frac{1}{\sqrt{k}}\gamma^{k-t}$$

Moreover, since $\sum_{k=t}^{T}\frac{1}{\sqrt{k}}\gamma^{k-t} \le \sum_{k=t}^{T}\frac{1}{\sqrt{t}}\gamma^{k-t} = \frac{1}{\sqrt{t}}\sum_{k=t}^{T}\gamma^{k-t} = \frac{1}{\sqrt{t}}\sum_{k=0}^{T-t}\gamma^{k} \le \frac{1}{\sqrt{t}}\left(\frac{1}{1-\gamma}\right)$, where the last inequality is by Lemma 2.4, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}| &\le \sum_{t=1}^{T}|g_{t,i}|\frac{1}{\sqrt{t}}\left(\frac{1}{1-\gamma}\right) \\ &= \frac{1}{1-\gamma}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}|g_{t,i}|.\end{aligned}$$

Furthermore, since

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}|g_{t,i}| &= \sqrt{\left(\sum_{t=1}^{T}\frac{1}{\sqrt{t}}|g_{t,i}|\right)^2} \\ &\le \sqrt{\sum_{t=1}^{T}\frac{1}{t}}\sqrt{\sum_{t=1}^{T}g_{t,i}^2} \\ &\le (\sqrt{\ln T + 1})\|g_{1:T,i}\|_2,\end{aligned}$$

where the first inequality is by Lemma 2.3 and the last inequality is by Lemma 2.5, we obtain

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\gamma^{t-k}|g_{k,i}| \le \frac{\sqrt{\ln T+1}}{1-\gamma}\|g_{1:T,i}\|_2.$$

Hence, by (15),

$$\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \le \frac{\sqrt{\ln T+1}}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)}\|g_{1:T,i}\|_2,$$

which ends the proof. □

Let us now prove Theorem 4.1.

Proof of Theorem 4.1: To prove Theorem 4.1, by Lemma 3.1, we need to bound the terms (6), (7), and (8). First, we consider (7). We have

$$\begin{aligned}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\alpha_t}{1-\beta_1}\frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} &= \frac{\alpha}{1-\beta_1}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}}\\ &\le \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\\ &\quad\times\sum_{i=1}^{d}\|g_{1:T,i}\|_2, \end{aligned}\tag{16}$$

where the equality is by the assumption that $\alpha_t=\alpha/\sqrt{t}$ and the last inequality is by Lemma 4.4. Next, we consider (8). The bound for (8) depends on either $\beta_{1,t}=\beta_1\lambda^{t-1}(0<\lambda<1)$ or $\beta_{1,t}=\frac{\beta_1}{t}$. Recall that by assumption, $\|x_m-x_n\|_\infty\le D_\infty$ for any $m,n\in\{1,2,\ldots,T\}$, $\alpha_t=\alpha/\sqrt{t}$. If $\beta_{1,t}=\beta_1\lambda^{t-1}(0<\lambda<1)$, then,

$$\begin{aligned}&\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_1)}(x_{t,i}-x_i^*)^2\\ &\quad=\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_1\lambda^{t-1}\sqrt{(t-1)}\sqrt{\hat{v}_{t-1,i}}}{2\alpha(1-\beta_1)}(x_{t,i}-x_i^*)^2\\ &\quad\le\frac{D_\infty^2G_\infty}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}\sqrt{(t-1)}\lambda^{t-1}\\ &\quad\le\frac{D_\infty^2G_\infty}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}t\lambda^{t-1}\\ &\quad\le\frac{D_\infty^2G_\infty}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\frac{1}{(1-\lambda)^2}\\ &\quad=\frac{dD_\infty^2G_\infty}{2\alpha(1-\beta_1)(1-\lambda)^2},\end{aligned}\tag{17}$$

where the first inequality is from Lemma 4.2 and the assumption that $\beta_1\le 1$, the last inequality is by Lemma 2.4. If $\beta_{1,t}=\frac{\beta_1}{t}$, then,

$$\begin{aligned}&\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_1)}(x_{t,i}-x_i^*)^2\\ &\quad=\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_1\sqrt{(t-1)}\sqrt{\hat{v}_{t-1,i}}}{2\alpha(1-\beta_1)t}(x_{t,i}-x_i^*)^2\\ &\quad\le\frac{D_\infty^2G_\infty}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\sqrt{(t-1)}}{t}\\ &\quad\le\frac{D_\infty^2G_\infty}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{1}{\sqrt{t}}\\ &\quad=\frac{dD_\infty^2G_\infty\sqrt{T}}{\alpha(1-\beta_1)},\end{aligned}\tag{18}$$

where the first inequality is from Lemma 4.2 and the assumption that $\beta_1\le 1$, and the last inequality is by Lemma 2.6.

Finally, we will bound (6). From (9) and replacing α_t with $\frac{\alpha}{\sqrt{t}}(1\le t\le T)$, we obtain

$$\begin{aligned}(6)&\le\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha(1-\beta_1)}(x_{1,i}-x_i^*)^2\\ &\quad+\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{T}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right).\end{aligned}$$

By Lemma 4.3, there is some $t_0(1\le t_0\le T)$ such that $\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}\ge\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}$ for all $t>t_0$. Therefore,

$$\begin{aligned}(6)&\le\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_{1,1})}(x_{1,i}-x_i^*)^2\\ &\quad+\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{t_0}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right)\\ &\quad+\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=t_0+1}^{T}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right)\\ &\le\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{1-\beta_{1,1}}\\ &\quad+\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{t_0}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right)\\ &\quad+\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\sum_{t=t_0+1}^{T}\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right).\end{aligned}$$

Since

$$\begin{aligned}&\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\sum_{t=t_0+1}^{T}\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right)\\ &\quad=\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{T\hat{v}_{T,i}}}{1-\beta_{1,T}}-\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{t_0\hat{v}_{t_0,i}}}{1-\beta_{1,t_0}}\\ &\quad\le\frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{T\hat{v}_{T,i}}}{1-\beta_{1,T}},\end{aligned}$$

we have

$$
\begin{aligned}
(6) &\le \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{1-\beta_{1,1}} + \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{T\hat{v}_{T,i}}}{1-\beta_{1,T}} \\
&\quad + \frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{t_0}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right) \\
&\le \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{1-\beta_{1,1}} + \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{T\hat{v}_{T,i}}}{1-\beta_{1,T}} \\
&\quad + \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{t_0}\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} \\
&\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\left(\sum_{t=1}^{t_0}\sqrt{t}+\sqrt{T}\right), \qquad (19)
\end{aligned}
$$

where the second inequality is obtained by omitting the term $\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{t_0}(x_{t,i}-x_i^*)^2\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}$, and the last inequality is by Lemma 4.2 and the assumption that $\beta_{1,t} \le \beta_1$ $(1 \le t \le T)$. Summing up, if $\beta_{1,t} = \beta_1\lambda^{t-1}$, then, from (16), (17), and (19), we obtain

$$
\begin{aligned}
R(T) &\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\left(\sum_{t=1}^{t_0}\sqrt{t}+\sqrt{T}\right) \\
&\quad + \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)(1-\lambda)^2} \\
&\quad + \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2.
\end{aligned}
$$

If $\beta_{1,t} = \frac{\beta_1}{t}$, then, from (16), (18), and (19), we obtain

$$
\begin{aligned}
R(T) &\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\left(\sum_{t=1}^{t_0}\sqrt{t}+\sqrt{T}\right) \\
&\quad + \frac{dD_\infty^2 G_\infty\sqrt{T}}{\alpha(1-\beta_1)} \\
&\quad + \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2,
\end{aligned}
$$

which ends the proof. □

The following corollary shows that, when either $\beta_{1,t} = \beta_1\lambda^{t-1}$ or $\beta_{1,t} = 1/t (1 \le t \le T)$, where $0 \le \beta_1 < 1$ and $0 < \lambda < 1$, the average regret of AMSGrad converges.

Corollary 4.5: With the same assumption as in Theorem 4.1, AMSGrad achieves the following guarantee:

$$\lim_{T\to\infty}\frac{R(T)}{T} = 0.$$

Proof: The result is obtained by using Theorem IV and the following fact:

$$
\begin{aligned}
\sum_{i=1}^{d}\|g_{1:T,i}\|_2 &= \sum_{i=1}^{d}\sqrt{g_{1,i}^2+g_{2,i}^2,\ldots+g_{T,i}^2} \\
&\le \sum_{i=1}^{d}\sqrt{TG_\infty^2} \\
&= dG_\infty\sqrt{T},
\end{aligned}
$$

where the inequality is from the assumption that $\|g_t\|_\infty \le G_\infty$ for all $t \in [T]$. □

V. NEW VERSION OF AMSGRAD OPTIMIZER: ADAMX

Let $f_1, f_2, \ldots, f_T : \mathcal{F} \to \mathbb{R}$ be an arbitrary sequence of convex cost functions. If the system $\{\beta_{1,t}\}_{1\le t\le T}$ is kept arbitrary, as in the setting of Theorem A, to ensure that the regret $R(T)$ satisfies $R(T)/T \to 0$, we suggest a new algorithm as follows.

Algorithm 2 AdamX: A New Variant of Adam and AMSGrad

Input: $x_1 \in \mathbb{R}^d$, step size $\{\alpha_t\}_{t=1}^T$, $\{\beta_{1,t}\}_{t=1}^T$, β_2
Set $m_0 = 0$, $v_0 = 0$, and $\hat{v}_0 = 0$
for $(t = 1; t \le T; t \leftarrow t+1)$ **do**
$g_t = \nabla f_t(x_t)$
$m_t = \beta_{1,t}\cdot m_{t-1} + (1-\beta_{1,t})\cdot g_t$
$v_t = \beta_2\cdot v_{t-1} + (1-\beta_2)\cdot g_t^2$

$$\hat{v}_t = \begin{cases} v_t & \text{if } t = 1 \\ \max\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1}, v_t\} & \text{if } t \ge 2 \end{cases}$$

$x_{t+1} = \prod_{\mathcal{F},\sqrt{\hat{V}_t}}(x_t - \alpha_t\cdot m_t/\sqrt{\hat{v}_t})$,
where $\hat{V}_t = \mathrm{diag}(\hat{v}_t)$
end for
Output: x_{T+1}

With this Algorithm 2, the regret is bounded as follows.

Theorem 5.1: Let x_t and v_t be the sequences obtained from Algorithm 2, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, $\beta_1 = \beta_{1,1}$, $\beta_{1,t} \le \beta_1$ for all $t \in [T]$ and $\frac{\beta_1}{\sqrt{\beta_2}} \le 1$. Assume that $\mathcal{F}$ has bounded diameter D_∞ and $\|\nabla f_t(x)\|_\infty \le G_\infty$ for all $t \in [T]$ and $x \in \mathcal{F}$. For x_t generated using the AdamX (Algorithm 2), we have the following bound on the regret:

$$
\begin{aligned}
R(T) &\le \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\sqrt{T} \\
&\quad + \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)}\sum_{t=2}^{T}\beta_{1,t}\sqrt{(t-1)} \\
&\quad + \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2.
\end{aligned}
$$

To prove Theorem 5.1, we need the following Lemmas 5.2, 5.3, and 5.4.

Lemma 5.2: For all $t \ge 1$, we have

$$\hat{v}_t = \max\left\{\tfrac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\le s\le t-1}, \qquad (20)$$

where $\hat{v}_t$ is in Algorithm 2.

Proof: We will prove (20) by induction on t. Recall that by the update rule on $\hat{v}_t$, we have $\hat{v}_1 \triangleq v_1$ and

$\hat{v}_t \triangleq \max\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1}, v_t\}$ if $t \geq 2$. Therefore,

$$\begin{aligned}\hat{v}_2 &\triangleq \max\left\{\frac{(1-\beta_{1,2})^2}{(1-\beta_{1,1})^2}\hat{v}_1, v_2\right\}\\ &= \max\left\{\frac{(1-\beta_{1,2})^2}{(1-\beta_{1,1})^2}v_1, v_2\right\}\\ &= \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq 2}.\end{aligned}$$

Assume that

$$\hat{v}_{t-1} = \max\left\{\frac{(1-\beta_{1,t-1})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t-1}$$

and the (20) holds for all $1 \leq j \leq t-1$. Since

$$\hat{v}_t \triangleq \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1}, v_t\right\},$$

we have

$$\begin{aligned}&\hat{v}_t\\ &= \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\left(\max\{\frac{(1-\beta_{1,t-1})^2}{(1-\beta_{1,s})^2}\hat{v}_s\}_{1\leq s\leq t-1}\right), v_t\right\}\\ &= \max\left\{\max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\frac{(1-\beta_{1,t-1})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t-1},\right.\\ &\quad\left.\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}v_t\right\}\\ &= \max\left\{\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t-1}, \frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}v_t\right\}\\ &= \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t-1},\end{aligned}$$

which ends the proof. □

Lemma 5.3: For all $t \geq 1$, we have $\sqrt{\hat{v}_t} \leq \frac{G_\infty}{1-\beta_1}$, where $\hat{v}_t$ is in Algorithm 2.

Proof: By Lemma 5.2,

$$\hat{v}_t = \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t}.$$

Therefore, there is some s such that $1 \leq s \leq t$ and $\hat{v}_t = \frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s$. Hence,

$$\begin{aligned}\sqrt{\hat{v}_t} &= \sqrt{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s}\\ &= \sqrt{1-\beta_2}\left(\frac{1-\beta_{1,t}}{1-\beta_{1,s}}\right)\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}g_k^2}\\ &\leq \sqrt{1-\beta_2}\left(\frac{1-\beta_{1,t}}{1-\beta_{1,s}}\right)\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}(\max_{1\leq k\leq s}|g_k|)^2}\\ &= G_\infty\sqrt{1-\beta_2}\left(\frac{1-\beta_{1,t}}{1-\beta_{1,s}}\right)\sqrt{\sum_{k=1}^{s}\beta_2^{s-k}}\\ &\leq G_\infty\sqrt{1-\beta_2}\left(\frac{1-\beta_{1,t}}{1-\beta_{1,s}}\right)\frac{1}{\sqrt{1-\beta_2}}\\ &= \left(\frac{1-\beta_{1,t}}{1-\beta_{1,s}}\right)G_\infty\\ &\leq \frac{G_\infty}{1-\beta_1},\end{aligned}$$

which ends the proof. □

Lemma 5.4: For the parameter settings and conditions assumed in Theorem 5.1, we have

$$\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{\sqrt{\ln T+1}}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)}\|g_{1:T,i}\|_2.$$

Proof: Since for all $t \geq 1$

$$\hat{v}_{t,i} = \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,s})^2}v_s\right\}_{1\leq s\leq t},$$

by Lemma 5.2, we have $\hat{v}_{t,i} \geq v_{t,i}$, and hence the proof is the same as that of Lemma 4.4. □

Proof of Theorem 5.1: Similarly to the proof of Theorem 4.1, we need to bound (6), (7), and (8). By using Lemma 5.4, we obtain the same bound for (7) as in the proof of Theorem 4.1, that is,

$$\begin{aligned}(7) &= \sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\alpha_t}{1-\beta_1}\frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}}\\ &= \frac{\alpha}{1-\beta_1}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{m_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}}\\ &\leq \frac{\alpha\sqrt{\ln T+1}}{(1-\beta_1)^2\sqrt{1-\beta_2}(1-\gamma)}\sum_{i=1}^{d}\|g_{1:T,i}\|_2,\end{aligned}$$

where the last inequality is by Lemma 5.4. Now we bound (8). By the assumption that $\|x_m - x_n\|_\infty \leq D_\infty$ for any $m, n \in \{1, \ldots, T\}$, $\alpha_t = \alpha/\sqrt{t}$, and $\beta_{1,t} = \beta_1\lambda^{t-1} \leq \beta_1 \leq 1$, we obtain

$$\begin{aligned}(8) &= \sum_{i=1}^{d}\sum_{t=2}^{T}\frac{\beta_{1,t}\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_{1,t})}(x_{t,i}-x_i^*)^2\\ &\leq \frac{D_\infty^2}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sum_{t=2}^{T}\beta_{1,t}\sqrt{(t-1)\hat{v}_{t-1,i}}.\end{aligned}$$

Therefore, from Lemma 5.3, we obtain

$$(8) \leq \frac{dD_\infty^2G_\infty}{2\alpha(1-\beta_1)^2}\sum_{t=2}^{T}\beta_{1,t}\sqrt{(t-1)}.$$

Finally, we will bound (6). By (9) and replacing $\alpha_t = \frac{\alpha}{\sqrt{t}}(1 \leq t \leq T)$, we obtain

$$\begin{aligned}(6) \leq &\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha(1-\beta_1)}(x_{1,i}-x_i^*)^2\\ &+\frac{1}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{T}(x_{t,i}-x_i^*)^2\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}}-\frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right)\end{aligned}$$

Moreover, by the update rule of Algorithm 2, we have

$$\hat{v}_{t,i} = \max\left\{\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1,i}, v_{t,i}\right\}.$$

Therefore, $\hat{v}_{t,i} \geq \frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1,i}$, and hence

$$\begin{aligned}
&\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}} \\
&\geq \frac{\sqrt{t\frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2}\hat{v}_{t-1,i}}}{1-\beta_{1,t}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}} \\
&= \frac{\sqrt{t\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}} > 0.
\end{aligned}$$

Now by the positivity of the essential formula

$$\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}},$$

we obtain

$$\begin{aligned}
(6) &\leq \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{1-\beta_1} \\
&+ \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\sum_{t=2}^{T}\left(\frac{\sqrt{t\hat{v}_{t,i}}}{1-\beta_{1,t}} - \frac{\sqrt{(t-1)\hat{v}_{t-1,i}}}{1-\beta_{1,t-1}}\right) \\
&= \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\frac{\sqrt{T\hat{v}_{T,i}}}{1-\beta_{1,T}} \\
&\leq \frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)^2}\sqrt{T},
\end{aligned}$$

where the last inequality is by Lemma 5.3. Hence we obtain the desired upper bound for $R(T)$. □

Corollary 5.5: With the same assumption as in Theorem 5.1, and for all $0 \leq \beta_{1,t} < 1$ satisfying

$$\lim_{T\to\infty}\frac{\sum_{t=2}^{T}\beta_{1,t}\sqrt{t-1}}{T} = 0,$$

AdamX achieves the following guarantee:

$$\lim_{T\to\infty}\frac{R(T)}{T} = 0.$$

Proof: By Theorem 5.1, it is sufficient to consider the term

$$\frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)^2}\sum_{t=2}^{T}\beta_{1,t}\sqrt{t-1}$$

on the right hand side of the upper bound for $R(T)$ in Theorem 5.1. Because $\frac{dD_\infty^2 G_\infty}{2\alpha(1-\beta_1)^2}$ is bounded and does not depend on T, the statement follows. □

When either $\beta_{1,t} = \beta_1\lambda^{t-1}$ for some $\lambda \in (0,1)$, or $\beta_{1,t} = \frac{1}{t}$ in Theorem 5.1, we obtain the following guarantee that the average regret of AdamX converges.

Corollary 5.6: With the same assumption as in Theorem 5.1, and either $\beta_{1,t} = \beta_1\lambda^{t-1}$ for some $\lambda \in (0,1)$, or $\beta_{1,t} = \frac{1}{t}$, AdamX achieves the following guarantee:

$$\lim_{T\to\infty}\frac{R(T)}{T} = 0.$$

Proof: By Corollary 5.5, it is sufficient to consider the term

$$\sum_{t=2}^{T}\beta_{1,t}\sqrt{t-1}.$$

When $\beta_{1,t} = \beta_1\lambda^{t-1}$ for some $\lambda \in (0,1)$, we have

$$\begin{aligned}
\sum_{t=2}^{T}\beta_{1,t}\sqrt{t-1} &= \sum_{t=2}^{T}\beta_1\lambda^{t-1}\sqrt{t-1} \\
&\leq \sum_{t=2}^{T}\sqrt{(t-1)}\lambda^{t-1} \\
&\leq \sum_{t=2}^{T}t\lambda^{t-1} \\
&\leq \frac{1}{(1-\lambda)^2} \qquad (21)
\end{aligned}$$

where the first inequality is from the property that $\beta_1 \leq 1$, and the last inequality is from Lemma 2.4. When $\beta_{1,t} = \frac{1}{t}$, we obtain

$$\begin{aligned}
\sum_{t=2}^{T}\beta_{1,t}\sqrt{t-1} &= \sum_{t=2}^{T}\frac{\sqrt{t-1}}{t} \\
&\leq \sum_{t=2}^{T}\frac{1}{\sqrt{t}} \\
&\leq 2\sqrt{T}, \qquad (22)
\end{aligned}$$

where the last inequality is from Lemma 2.6. Now, by combining (21) and (22) with Corollary 5.5, we obtain the desired result. □

VI. EXPERIMENTS

While we consider our main contributions as the theoretical analyses on AMSGrad and AdamX in the previous sections, we provide experimental results in this section for AMSGrad and AdamX. Concretely, we use the PyTorch code for AMSGrad[2] via setting the Boolean flag `amsgrad = True`. The code for AdamX is based on that of AMSGrad, with corresponding modifications as in Algorithm 2. The parameters for AMSGrad and AdamX are identical in our experiments, namely $(\beta_1, \beta_2) = (0.9, 0.999)$, the term added to the denominator to improve numerical stability is $\epsilon = 10^{-8}$, and additionally we set $\beta_{1,t} = \beta_1\lambda^{t-1}$ with $\lambda = 0.001$ to make use of Corollary 5.6 on the convergence of AdamX.

The learning rate is scheduled for both optimizers AMSGrad and AdamX as follows: 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, $10^{-6}/2$ if the epoch is correspondingly in the ranges

[2]https://pytorch.org/docs/stable/_modules/torch/optim/adam.html

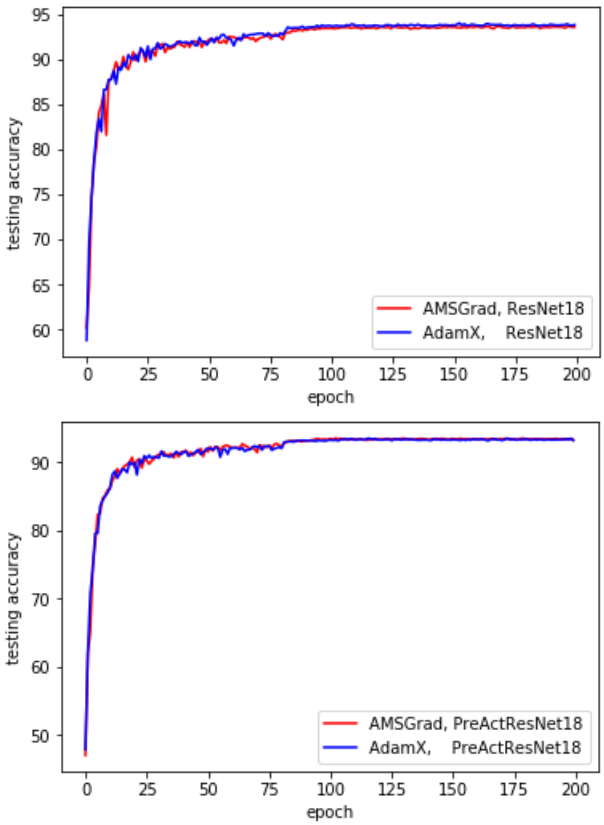


FIGURE 1. Testing accuracies over CIFAR-10 using AMSGrad and AdamX, with different neural network models.

[0, 80], [81, 120], [121, 160], [161, 180], [181, 200]. We use CIFAR[3]-10 (containing 50000 training images and 10000 test images of size 32 × 32) as the dataset and the residual networks ResNet18 [8] and PreActResNet18 [9] for training with batch size is 128. The testing result is given in Figure 1 where one can see that AMSGrad and AdamX behaves similarly, which supports our theoretical results on the convergence of both AMSGrad (Section IV) and AdamX (Section V).

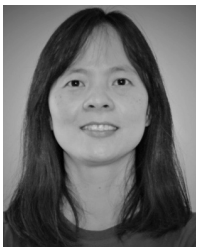
VII. CONCLUSION

We have shown that the convergence proof of AMSGrad [3] is problematic, and presented various fixes for it, which include a new and slightly modified version called AdamX. Our work helps ensure the theoretical foundation of those optimizers.

[3]https://www.cs.toronto.edu/ kriz/cifar.html

REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.
[2] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15. [Online]. Available: https://arxiv.org/pdf/1412.6980.pdf
[3] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–23. [Online]. Available: https://openreview.net/pdf?id=ryQu7f-RZ
[4] H. B. McMahan and M. Streeter, "Adaptive bound optimization for online convex optimization," in *Proc. 23rd Annu. Conf. Learn. Theory (COLT)*, 2010, pp. 244–256. [Online]. Available: https://arxiv.org/abs/1002.4908
[5] L. Luo, Y. Xiong, and Y. Liu, "Adaptive gradient methods with dynamic bound of learning rate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–19. [Online]. Available: https://openreview.net/pdf?id=Bkg3g2R9FX
[6] S. Bock, J. Goppold, and M. Weiß. (2018). "An improvement of the convergence proof of the ADAM-Optimizer." [Online]. Available: https://arxiv.org/abs/1804.10587
[7] J. Chen and Q. Gu. (2018). "Closing the generalization gap of adaptive gradient methods in training deep neural networks." [Online]. Available: https://arxiv.org/abs/1806.06763
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. [Online]. Available: https://arxiv.org/abs/1512.03385
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 630–645. [Online]. Available: https://arxiv.org/abs/1603.05027



TRAN THI PHUONG received the B.S. degree from the University of Natural Sciences, Ho Chi Minh City, Vietnam, in 2002, and the Ph.D. degree from Meiji University, Japan, in 2012. She is currently with Ton Duc Thang University, Vietnam, with Meiji University, and with the National Institute of Information and Communications Technology (NICT), Japan. Her current research interests include mathematics and deep learning.

LE TRIEU PHONG received the B.S. degree from the University of Natural Sciences, Ho Chi Minh City, Vietnam, in 2002, and the Ph.D. degree from the Tokyo Institute of Technology, in 2009. He is currently a Senior Researcher with the Cybersecurity Research Institute, National Institute of Information and Communications Technology (NICT), Japan. His current research interests include applied cryptography and privacy-preserving data mining.