The Adam and AMSGrad Algorithms for Online Convex Programming

Peter Prastakos

May 16, 2022

1 Introduction

In Kingma and Ba (2015), the authors propose a new method for stochastic optimization, which they denote the Adam algorithm. In this project, I will present this algorithm, and then discuss how Reddi et al. (2018) establish non-convergence of Adam in an online convex optimization example. I will finally present a variant of Adam, called AMSGrad, that Reddi et al. (2018) show has good convergence results.

1.1 Notation

We will adjust the notation of Reddi et al. (2018) slightly. We use $\mathcal{S}_d^+$ to denote the set of all positive semidefinite $d \times d$ real matrices. For a vector $a \in \mathbb{R}^d$ and a matrix $M \in \mathcal{S}_d^+$, we use $\|M_i\|_2$ to denote ℓ_2-norm of i^{th} row of M and $\sqrt{M}$ to represent $M^{1/2}$. Furthermore, for any vectors $a, b \in \mathbb{R}^d$, we use $\sqrt{a}$ for element-wise square root, a^2 for element-wise square, a/b to denote element-wise division and $\max(a, b)$ to denote element-wise maximum. For any vector $\theta_i \in \mathbb{R}^d$, $\theta_{i,j}$ denotes its j^{th} coordinate where $j \in [d]$. The projection operation $\Pi_{\mathcal{F},A}(y)$ for $A \in \mathcal{S}_+^d$ is defined as $\arg\min_{x \in \mathcal{F}} \|A^{1/2}(x - y)\|$ for $y \in \mathbb{R}^d$. Furthermore, for a sequence of vectors $g_1, \ldots, g_t \in \mathbb{R}^d$, we use $g_{1:t} = [g_1 \ldots g_t]$ to denote the $d \times t$ matrix obtained by concatenating the vector sequence. Finally, we say $\mathcal{F}$ has bounded diameter D_∞ if $\|x - y\|_\infty \leq D_\infty$ for all $x, y \in \mathcal{F}$.

1.2 Online Learning Preliminaries

Since the setup of the problem is in the field of online learning, we first begin with some preliminaries. *Online programming* refers to a class of problems (a game) where a decision-maker (DM) sequentially chooses points in a feasible set (i.e. the parameters of the model to be learned) in response to a time-varying sequence of loss functions chosen by the environment (E). The procedure can be described as follows:

Algorithm 1 Online Programming

for $t = 1, 2, \dots$ **do**
  DM chooses $x_t \in \mathcal{F}$, where $\mathcal{F} \subset \mathbb{R}^d$ is the feasible set
  E chooses loss function f_t
  DM incurs loss $f_t(x_t)$
end for

Since the nature of the sequence is unknown in advance of the choice of parameters at each time step t, the quality of the learning model is evaluated using the notion of the algorithm's *regret*, denoted R_T, which is the difference between the total loss until time step T and the smallest total loss that could have been achieved in hindsight by a single fixed point parameter $x^* \in \mathcal{F}$. In other words,

$$R_T = \sum_{i=1}^{T} f_t(x_t) - \min_{x \in \mathcal{F}} \sum_{i=1}^{T} f_t(x).$$

The goal is then to devise an algorithm that ensures $R_T = o(T)$, i.e. $R_T/T \to 0$ as $T \to \infty$. This implies that on average, the model's performance converges to the optimal one. An important note is that, though Kingma and Ba (2015) and Reddi et al. (2018) do not assume convexity of the functions f_t nor that $\mathcal{F}$ is a convex set in their presentation of AMSGrad, they do assume so in their proof of convergence. Thus, since this project explores the convergence of these algorithms, we will assume we are in the setting of online *convex* programming (OCP), that is, online programming in the case where the functions f_t are convex and the feasible set $\mathcal{F}$ is convex. We will furthermore assume, as in Reddi et al. (2018), that the feasible set $\mathcal{F}$ has bounded diameter, the functions f_t are differentiable, and $\|\nabla f_t(x)\|_\infty$ is bounded for all $t \in [T]$ and $x \in \mathcal{F}$.

2 Adam and related algorithms for OCP

The simplest algorithm for this online setting is the online gradient descent algorithm (see Zinkevich (2003)), which moves the point x_t in the opposite direction of the gradient $g_t = \nabla f_t(x_t)$ while maintaining the feasibility by projecting onto the set $\mathcal{F}$ via the update rule $x_{t+1} = \Pi_{\mathcal{F},\mathbb{I}}(x_t - \alpha_t g_t)$, where α_t is typically set to $\alpha/\sqrt{t}$ for some constant α. This differs from the basic stochastic gradient descent method (SGD) we discussed in lecture 14 only in the sense that, instead of randomly selecting one of the training samples to update the coefficients, we are selecting the most recent function's gradient at each time step t.

The general setup of algorithms in this online framework can be primarily encapsulated by the following algorithm:

Algorithm 2 Generic Method for OCP

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t > 0\}_{t=1}^T$, sequence of functions $\{\phi_t, \psi_t\}_{t=1}^T$
for $t = 1$ **to** T **do**
  $g_t = \nabla f_t(x_t)$
  $m_t = \phi_t(g_1, \ldots, g_t)$ and $v_t = \psi_t(g_1, \ldots, g_t)$
  $\hat{x}_{t+1} = x_t - \alpha_t m_t / \sqrt{v_t}$
  $V_t = \mathrm{diag}(v_t)$
  $x_{t+1} = \Pi_{\mathcal{F}, \sqrt{V_t}}(\hat{x}_{t+1})$
end for

Note the algorithm is still quite general since the "averaging" functions ϕ_t and ψ_t have not been specified. Here $\phi_t : \mathcal{F}^t \to \mathbb{R}^d$ and $\psi_t : \mathcal{F}^t \to \mathbb{R}^d_{\geq 0}$. For ease of exposition, we refer to α_t as step size and $\frac{\alpha_t}{\sqrt{v_t}}$ as learning rate of the algorithm. We first observe that online gradient descent falls in this framework by setting

$$\phi_t(g_1, \ldots, g_t) = g_t, \quad \psi_t(g_1, \ldots, g_t) = \mathbf{1}, \text{ and } \alpha_t = \alpha/\sqrt{t} \quad \forall t \in [T].$$

According to Reddi et al. (2018), while the decreasing step size is necessary for convergence, such an aggressive decay of learning rate typically translates into poor empirical performance. Thus, this motivated the creation of adaptive methods, which choose averaging functions appropriately so as to entail good convergence. The first adaptive method, called ADAGRAD (Duchi et al., 2011), uses the following averaging

functions:

$$\phi_t(g_1, \ldots, g_t) = g_t \text{ and } \psi_t(g_1, \ldots, g_t) = \frac{1}{t}\sum_{i=1}^{t} g_i^2, \qquad \text{(ADAGRAD)}$$

and step size $\alpha_t = \alpha/\sqrt{t}$ for all $t \in [T]$. In contrast to a learning rate of $\alpha/\sqrt{t}$ in online gradient descent, such a setting effectively implies a modest learning rate decay of $\alpha/\sqrt{\sum_i g_{i,j}^2}$ for $j \in [d]$. When the gradients are sparse, this can potentially lead to huge gains in terms of convergence (see Duchi et al. (2011)).

Adam. Although ADAGRAD seems to works well in practice for sparse settings, its performance has been observed to deteriorate in settings where the loss functions are nonconvex and gradients are dense due to rapid decay of the learning rate in these settings since it uses all the past gradients in the update. This has motivated the development of adaptive methods based on exponential moving averages of squared gradients which essentially limit the reliance of the $\hat{x}_{t+1}$ update to only the past few gradients (since $\lim_{n\to\infty} a^n = 0$ for all $a \in [0, 1)$). ADAM, a particularly popular algorithm in this category, uses the following functions:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \text{ and } v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad \forall t \in [T] \qquad (1)$$

for some $\beta_1, \beta_2 \in [0, 1)$, where m_0 and v_0 are initialized to be 0. Solving the recursion, we can reduce this to:

$$\phi_t(g_1, \ldots, g_t) = (1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g_i \text{ and } \psi_t(g_1, \ldots, g_t) = (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2. \qquad (2)$$

Note that the presentation of ADAM in Kingma and Ba (2015) includes a debiasing step wherein

$$\hat{m}_t = m_t/(1-\beta_1^t) \text{ and } \hat{v}_t = v_t/(1-\beta_1^t)$$

and then the $\hat{x}_{t+1}$ update is done using $\hat{m}_t$ and $\hat{v}_t$ instead of m_t and v_t, but the non-convergence argument of Reddi et al. (2018) works for the debiased version as well, so we remove it for simplicity (as did the authors). As with ADAGRAD, we use step size $\alpha_t = \alpha/\sqrt{t}$. A value of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is typically recommended in practice.

3 The Non-Convergence of Adam

Reddi et al. (2018) show that ADAM can fail to converge to an optimal solution even in simple one-dimensional convex settings, thereby contradicting the claim of convergence in Theorem 4.1 of Kingma and Ba (2015). The main issue lies in investigating the following quantity:

$$\delta_{t+1} = \frac{\sqrt{v_{t+1}}}{\alpha_{t+1}} - \frac{\sqrt{v_t}}{\alpha_t}. \tag{3}$$

This quantity essentially measures the change in the inverse of the learning rate with respect to time. One key observation is that for online gradient descent and ADAGRAD, all entries of δ_t are nonnegative for all $t \in [T]$. This simply follows from the update rules of the algorithms in Section 2, if one observes that the learning rates, $\alpha/\sqrt{t}$ and $\alpha/\sqrt{\sum_i g_i^2}$ are nonincreasing functions of t. However, this is not necessarily the case for exponential moving average variants like ADAM, which can lead to undesirable convergence behavior as in the counterexample below.

Theorem 1. *There is an online convex optimization problem where* ADAM *has nonzero average regret i.e.,* $R_T/T \nrightarrow 0$ *as* $T \to \infty$.

Proof. We consider the setting where f_t are linear functions and $\mathcal{F} = [-1, 1]$. In particular, we define the following function sequence:

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

where $C \geq 2$. For this function sequence, we have that

$$\sum_{i=1}^{T} f_t(x) = \begin{cases} \frac{T}{3}(C-2)x, & \text{for } T \bmod 3 = 0 \\ \frac{T-1}{3}(C-2)x + Cx, & \text{for } T \bmod 3 = 1 \\ \frac{T-2}{3}(C-2)x + (C-1)x, & \text{for } T \bmod 3 = 2. \end{cases}$$

In all cases, since the functions are linear and $C \geq 2$ (thus slope is nonnegative), the minimization is solved by the leftmost endpoint in the feasible set $\mathcal{F}$, i.e. $x^* = -1$. Thus, we would ideally like our algorithm to converge to -1, as that would provide the minimum regret. Without loss of generality, assume that the initial point is $x_1 = 1$.

This can be assumed without any loss of generality because for any choice of initial point, we can always translate the coordinate system such that the initial point is $x_1 = 1$ in the new coordinate system and then choose the sequence of functions as above in the new coordinate system. Consider the execution of ADAM algorithm for this sequence of functions with

$$\beta_1 = 0, \beta_2 = \frac{1}{1+C^2} \text{ and } \alpha_t = \frac{\alpha}{\sqrt{t}}$$

where $\alpha < \sqrt{1-\beta_2}$.

Now note that, in their convergence analysis, Kingma and Ba (2015) assume that the sequence of functions has bounded gradients, the feasible set $\mathcal{F}$ has bounded diameter, and that $\beta_1^2/\sqrt{\beta_2} < 1$. In this specific sequence of functions, those conditions are satisfied since $\|\nabla f_t(x)\|_\infty \leq C$ for all $t \in [T]$ and $x \in \mathcal{F}$, $\|x-y\|_\infty \leq 2$ for all $x, y \in \mathcal{F}$, and trivially $\beta_1^2 = 0 < \frac{1}{\sqrt{1+C^2}} = \sqrt{\beta_2}$.

Our main claim is that for iterates $\{x_t\}_{t=1}^\infty$ arising from the updates of ADAM, we do not observe convergence to -1. In fact, we have $x_t > 0$ for all $t \in \mathbb{N}$ and furthermore, $x_{3t+1} = 1$ for all $t \in \mathbb{N} \cup \{0\}$. We prove this by induction on t. Since $x_1 = 1$, both the aforementioned conditions hold for the base case. Suppose for some $t \in \mathbb{N} \cup \{0\}$, we have $x_i > 0$ for all $i \in [3t+1]$ and $x_{3t+1} = 1$. Our aim is to prove that x_{3t+2} and x_{3t+3} are positive and $x_{3t+4} = 1$. We first observe that the gradients have the following form:

$$\nabla f_i(x) = \begin{cases} C, & \text{for } i \bmod 3 = 1 \\ -1, & \text{otherwise} \end{cases}$$

From $(3t+1)^{\text{th}}$ update of ADAM in Equation (1), we obtain

$$\hat{x}_{3t+2} = x_{3t+1} - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}} = 1 - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}}.$$

The equality follows from the induction hypothesis that $x_{3t+1} = 1$. We now observe

that

$$\frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}} \leq \frac{\alpha C}{\sqrt{(3t+1)(1-\beta_2)C^2}} = \frac{\alpha}{\sqrt{(3t+1)(1-\beta_2))}}, \tag{4}$$

and since $\alpha < \sqrt{1-\beta_2}$, we have that

$$\frac{\alpha}{\sqrt{(3t+1)(1-\beta_2))}} < 1.$$

Therefore, we have $0 < \hat{x}_{3t+2} < 1$ and hence $x_{3t+2} = \hat{x}_{3t+2} > 0$. Furthermore, after the $(3t+2)^{\text{th}}$ and $(3t+3)^{\text{th}}$ updates of ADAM in Equation (1), we have the following:

$$\hat{x}_{3t+3} = x_{3t+2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}},$$
$$\hat{x}_{3t+4} = x_{3t+3} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}.$$

Since $x_{3t+2} > 0$, we clearly also have that $x_{3t+3} > 0$.

It now remains to show that $x_{3t+4} = 1$ to complete the inductive proof. In order to show this, we show that $\hat{x}_{3t+4} \geq 1$, which readily translates to $x_{3t+4} = 1$ because $x_{3t+4} = \Pi_{\mathcal{F}, \sqrt{V_{3t+3}}}(\hat{x}_{3t+4})$ and $\mathcal{F} = [-1, 1]$. Note here that since we are in the one dimensional case, $\Pi_{\mathcal{F}, \sqrt{V_t}} = \Pi_{\mathcal{F}, \mathbb{I}}$ so this is the simple Euclidean projection. We now observe the following:

$$\hat{x}_{3t+4} = \min(\hat{x}_{3t+3}, 1) + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}. \tag{5}$$

The above equality is due to the fact that $\hat{x}_{3t+3} > 0$ and the property of the projection operation onto the set $\mathcal{F} = [-1, 1]$. We consider the following two cases:

1. Suppose $\hat{x}_{3t+3} \geq 1$, then it is clear from Equation (5) that $\hat{x}_{3t+4} > 1$.

2. Suppose $\hat{x}_{3t+3} < 1$, then we have the following:

$$\begin{aligned}
\hat{x}_{3t+4} &= \hat{x}_{3t+3} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \\
&= x_{3t+2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \\
&= 1 - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} \\
&\qquad + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}.
\end{aligned}$$

The third equality is due to the fact that $x_{3t+2} = \hat{x}_{3t+2}$. Thus, to prove $\hat{x}_{3t+4} > 1$, it is enough to prove:

$$\underbrace{\frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}}}_{T_1} \leq \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} \underbrace{+ \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}}_{T_2}$$

We have the following bound on term T_1 from Equation (4):

$$T_1 \leq \frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}}. \tag{6}$$

Now, we lower bound T_2. Note that since the norm of the gradient at any feasible point is upper bounded by C, we have that $v_t \leq C^2$ for all $t \in \mathbb{N}$. Thus, we have

$$\begin{aligned}
T_2 &= \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \\
&\geq \frac{\alpha}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{3t+2}} + \frac{1}{\sqrt{3t+3}} \right) \\
&\geq \frac{\alpha}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{2(3t+1)}} + \frac{1}{\sqrt{2(3t+1)}} \right) \\
&= \frac{\sqrt{2}\alpha}{\sqrt{(3t+1)(\beta_2 C^2 + (1-\beta_2))}}.
\end{aligned} \tag{7}$$

Now note that since $\beta_2 = 1/(1+C^2)$, we have that

$$\sqrt{\frac{\beta_2 C^2 + (1-\beta_2)}{2}} = \sqrt{1-\beta_2}$$

and hence we get that

$$\frac{\sqrt{2}\alpha}{\sqrt{(3t+1)(\beta_2 C^2 + (1-\beta_2))}} = \frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}} \geq T_1.$$

where the inequality follows from inequality in Equation (6). Therefore, we have $T_2 \geq T_1$ and hence, $\hat{x}_{3t+4} \geq 1$.

Therefore, from both cases, we have that $x_{3t+4} = 1$. Therefore, by induction we have that $x_t > 0$ for all $t \in \mathbb{N}$ and $x_{3t+1} = 1$ for all $t \in \mathbb{N} \cup \{0\}$. Thus, noting that $f_{3t+1}(x_{3t+1}) = C$, $f_{3t+2}(x_{3t+2}) \in [-1, 0)$, $f_{3t+3}(x_{3t+3}) \in [-1, 0)$, $f_{3t+1}(-1) = -C$, $f_{3t+2}(-1) = 1$, and $f_{3t+3}(-1) = 1$ we have

$$\begin{aligned}
&f_{3t+1}(x_{3t+1}) + f_{3t+2}(x_{3t+2}) + f_{3t+3}(x_{3t+3}) - f_{3t+1}(-1) - f_{3t+2}(-1) - f_{3t+3}(-1) \\
&\geq 2C - 2 + f_{3t+2}(x_{3t+2}) + f_{3t+3}(x_{3t+3}) \geq 2C - 4.
\end{aligned}$$

Therefore, for every 3 steps, ADAM suffers a regret of at least $2C - 4$. This then implies that $R_T \geq (2C-4)T/3$. Since $C \geq 2$, this regret can be very large and furthermore, $R_T/T \nrightarrow 0$ as $T \to \infty$, which completes the proof. ■

Reddi et al. (2018) generalize this result by establishing non-convergence in the case where β_1 and β_2 are any constants in $[0, 1)$ such that $\beta_1 < \sqrt{\beta_2}$. Though I do not provide the proof here, the function sequence they use as a counterexample is quite similar to the one above, as follows:

$$f_t(x) = \begin{cases} Cx, & \text{for } t \bmod C = 1 \\ -x, & \text{otherwise.} \end{cases}$$

Here C is a positive even number satisfying the following constraints:

$$\begin{aligned}(1-\beta_1)\beta_1^{C-1}C &\leq 1-\beta_1^{C-1},\\ \beta_2^{(C-2)/2}C^2 &\leq 1,\\ \frac{3(1-\beta_1)}{2\sqrt{1-\beta_2}}\left(1+\frac{\gamma(1-\gamma^{C-1})}{1-\gamma}\right)+\frac{\beta_1^{C/2-1}}{1-\beta_1} &< \frac{C}{3},\end{aligned} \tag{8}$$

where $\gamma = \beta_1/\sqrt{\beta_2} < 1$ (see Theorem 2 in Reddi et al. (2018) for details). Reddi et al. (2018) also establish non-convergence in the case where the update rule of $\hat{x}_{t+1}$ is modified as follows

$$\hat{x}_{t+1} = x_t - \alpha_t m_t/\sqrt{v_t + \epsilon\mathbf{1}}, \tag{9}$$

where $\epsilon > 0$. In practice, selection of the ϵ parameter can critically improve the performance of the algorithm according to the authors. However, using the function sequence

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

and the same feasible set $\mathcal{F} = [-1, 1]$, the authors show that ADAM again fails to converge to non-zero average regret asymptotically (see Theorem 6 in Reddi et al. (2018) for details).

Note that while the results stated above use constant β_1 and β_2, the analysis of ADAM in Kingma and Ba (2015) actually relies on decreasing β_1 over time. However, extending the examples to the case where β_1 is decreased over time is certainly feasible since the critical parameter for the arguments is β_2 rather than β_1, and as long as β_2 is strictly less than 1, the analysis goes through.

4 The AMSGrad algorithm and its convergence

The primary constraint that AMSGRAD tries to satisfy in order to guarantee non-zero average regret asymptotically is that δ_t in Equation (3) has nonnegative entries for all $t \in [T]$. The algorithm is as follows:

Algorithm 3 AMSGrad

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t\}_{t=1}^T$, $\{\beta_{1t}\}_{t=1}^T$, β_2

Set $m_0 = 0$, $v_0 = 0$ and $\hat{v}_0 = 0$

for $t = 1$ **to** T **do**

$g_t = \nabla f_t(x_t)$

$m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$

$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ and $\hat{V}_t = \mathrm{diag}(\hat{v}_t)$

$x_{t+1} = \Pi_{\mathcal{F}, \sqrt{\hat{V}_t}}(x_t - \alpha_t m_t / \sqrt{\hat{v}_t})$

end for

The key difference of AMSGrad with Adam is that it maintains the maximum of all v_t until the present time step and uses this maximum value for normalizing the running average of the gradient instead of v_t in Adam. By doing this, AMSGrad results in a nonincreasing learning rate and avoids the pitfalls of Adam i.e., $\delta_t \geq 0$ for all $t \in [T]$.

The following result establishes a regret of $O(\sqrt{T})$ for AMSGrad.

Theorem 2. *Let $\{x_t\}$ and $\{v_t\}$ be the sequences obtained from Algorithm 3, $\alpha_t = \alpha/\sqrt{t}$, $\beta_1 = \beta_{11}$, $\beta_{1t} \leq \beta_1$ for all $t \in [T]$ and $\gamma = \beta_1/\sqrt{\beta_2} < 1$. Assume that $\mathcal{F}$ has bounded diameter D_∞ and $\|\nabla f_t(x)\|_\infty \leq G_\infty$ for all $t \in [T]$ and $x \in \mathcal{F}$. For x_t generated using the* AMSGrad *(Algorithm 3), we have the following bound on the regret*

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

Proof. First we will need the following lemma:

Lemma 1. *For any $Q \in \mathcal{S}_+^d$ and convex feasible set $\mathcal{F} \subset \mathbb{R}^d$, suppose $u_1 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_1)\|$ and $u_2 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_2)\|$ then we have $\|Q^{1/2}(u_1 - u_2)\| \leq \|Q^{1/2}(z_1 - z_2)\|$.*

Proof. Since $u_1 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_1)\|$ and $u_2 = \min_{x \in \mathcal{F}} \|Q^{1/2}(x - z_2)\|$, we have from the property of the projection operator that

$$\langle z_1 - u_1, Q(z_2 - z_1) \rangle \geq 0 \text{ and } \langle z_2 - u_2, Q(z_1 - z_2) \rangle \geq 0.$$

Combining the above inequalities, we have

$$\langle u_2 - u_1, Q(z_2 - z_1)\rangle \geq \langle z_2 - z_1, Q(z_2 - z_1)\rangle. \tag{10}$$

Observe that since $Q \in \mathcal{S}_+^d$, we have that

$$\langle (u_2 - u_1) - (z_2 - z_1), Q((u_2 - u_1) - (z_2 - z_1))\rangle \geq 0,$$

so rearranging, we get

$$\langle u_2 - u_1, Q(z_2 - z_1)\rangle \leq \frac{1}{2}[\langle u_2 - u_1, Q(u_2 - u_1)\rangle + \langle z_2 - z_1, Q(z_2 - z_1)\rangle].$$

Combining the above inequality with Equation (10), we have the required result. ■

We now begin the proof of the theorem. Note that, by definition of the projection operator, we have

$$x_{t+1} = \Pi_{\mathcal{F}, \sqrt{\hat{V}_t}}(x_t - \alpha_t m_t / \sqrt{\hat{v}_t}) = \min_{x \in \mathcal{F}} \|\hat{V}_t^{1/4}(x - (x_t - \alpha_t m_t / \sqrt{\hat{v}_t}))\|.$$

Furthermore, $\Pi_{\mathcal{F}, \sqrt{\hat{V}_t}}(x^*) = x^*$ for all $x^* \in \mathcal{F}$. In this proof, we will use x_i^* to denote the i^{th} coordinate of x^*. Using Lemma 1 with $u_1 = x_{t+1}$ and $u_2 = x^*$, we have the following:

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

Rearranging the above inequality, we have

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

$$\leq \frac{1}{2\alpha_t(1-\beta_{1t})}\left[\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2\right] + \frac{\alpha_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_t\|^2$$
$$+ \frac{\beta_{1t}}{2(1-\beta_{1t})}\alpha_t\|\hat{V}_t^{-1/4}m_{t-1}\|^2 + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2. \tag{11}$$

The second inequality follows from a direct application of Cauchy-Schwarz and Young's inequality. Now note that, using convexity of the functions f_t, we have that

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \sum_{t=1}^{T}\langle g_t, x_t - x^*\rangle.$$

Furthermore, using the bound from Equation (11) and the fact that $\hat{v}_{t-1,i} \leq \hat{v}_{t,i}$, we have that

$$\sum_{t=1}^{T}\langle g_t, x_t - x^*\rangle \leq \sum_{t=1}^{T}\left[\frac{1}{2\alpha_t(1-\beta_{1t})}\left[\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2\right]\right.$$
$$\left. + \frac{\alpha_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_t\|^2 + \frac{\beta_{1t}}{2(1-\beta_{1t})}\alpha_{t-1}\|\hat{V}_{t-1}^{-1/4}m_{t-1}\|^2 + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2\right]$$
$$\leq \sum_{t=1}^{T}\left[\frac{1}{2\alpha_t(1-\beta_{1t})}\left[\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2\right] + \frac{\alpha_t}{(1-\beta_1)}\|\hat{V}_t^{-1/4}m_t\|^2\right.$$
$$\left. + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2\right]. \tag{12}$$

For further bounding this inequality, we need the following intermediate result.

Lemma 2. *For the parameter settings and conditions assumed in Theorem 2, we have*

$$\sum_{t=1}^{T}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 \leq \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2$$

Proof. We start with the following:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 &= \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha_T\sum_{i=1}^{d}\frac{m_{T,i}^2}{\sqrt{\hat{v}_{T,i}}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha_T\sum_{i=1}^{d}\frac{m_{T,i}^2}{\sqrt{v_{T,i}}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}(1-\beta_{1j})\Pi_{k=1}^{T-j}\beta_{1(T-k+1)}g_{j,i})^2}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}g_{j,i}^2)}}\end{aligned}$$

The first inequality follows from the definition of $\hat{v}_{T,i}$, as the maximum of all $v_{T,i}$ until the current time step. The second inequality follows from the update rule of AMSGrad (this is similar to Equation (2) except now β_1 is not assumed to be constant for all t). We further bound the above inequality in the following manner:

$$\begin{aligned}&\sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}(1-\beta_{1j})\Pi_{k=1}^{T-j}\beta_{1(T-k+1)}g_{j,i})^2}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}g_{j,i}^2)}} \qquad (13)\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{1(T-k+1)})(\sum_{j=1}^{T}\Pi_{k=1}^{T-j}\beta_{1(T-k+1)}g_{j,i}^2)}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}g_{j,i}^2)}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \alpha\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}\beta_1^{T-j})(\sum_{j=1}^{T}\beta_1^{T-j}g_{j,i}^2)}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}g_{j,i}^2)}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \frac{\alpha}{1-\beta_1}\sum_{i=1}^{d}\frac{\sum_{j=1}^{T}\beta_1^{T-j}g_{j,i}^2}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}g_{j,i}^2)}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \frac{\alpha}{(1-\beta_1)\sqrt{T(1-\beta_2)}}\sum_{i=1}^{d}\sum_{j=1}^{T}\frac{\beta_1^{T-j}g_{j,i}^2}{\sqrt{\beta_2^{T-j}g_{j,i}^2}}\\ &\le \sum_{t=1}^{T-1}\alpha_t\|\hat{V}_t^{-1/4}m_t\|^2 + \frac{\alpha}{(1-\beta_1)\sqrt{T(1-\beta_2)}}\sum_{i=1}^{d}\sum_{j=1}^{T}\gamma^{T-j}|g_{j,i}| \qquad (14)\end{aligned}$$

The first inequality follows from an application of Cauchy-Schwarz. The second inequality is due to the fact that $\beta_{1k} \le \beta_1$ for all $k \in [T]$. The third inequality follows from the inequality $\sum_{j=1}^{T}\beta_1^{T-j} \le \frac{1}{1-\beta_1}$. By using similar upper bounds for all time

steps, the quantity in Equation (14) can further be bounded as follows:

$$\sum_{t=1}^{T} \alpha_t \|\hat{V}_t^{-1/4} m_t\|^2 \leq \sum_{t=1}^{T} \frac{\alpha}{(1-\beta_1)\sqrt{t(1-\beta_2)}} \sum_{i=1}^{d} \sum_{j=1}^{t} \gamma^{t-j} |g_{j,i}|$$

$$= \frac{\alpha}{(1-\beta_1)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{j=1}^{t} \gamma^{t-j} |g_{j,i}| = \frac{\alpha}{(1-\beta_1)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \sum_{t=1}^{T} |g_{t,i}| \sum_{j=t}^{T} \frac{\gamma^{j-t}}{\sqrt{j}}$$

$$\leq \frac{\alpha}{(1-\beta_1)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \sum_{t=1}^{T} |g_{t,i}| \sum_{j=t}^{T} \frac{\gamma^{j-t}}{\sqrt{t}} \leq \frac{\alpha}{(1-\beta_1)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \sum_{t=1}^{T} |g_{t,i}| \frac{1}{(1-\gamma)\sqrt{t}}$$

$$\leq \frac{\alpha}{(1-\beta_1)(1-\gamma)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2 \sqrt{\sum_{t=1}^{T} \frac{1}{t}} \leq \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)(1-\gamma)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

The third inequality follows from the fact that $\sum_{j=t}^{T} \gamma^{j-t} \leq \frac{1}{1-\gamma}$. The fourth inequality is due to simple application of Cauchy-Schwarz. The final inequality is due to the following bound on harmonic sum: $\sum_{t=1}^{T} 1/t \leq (1 + \log T)$. This completes the proof of the lemma. ■

We now return to the proof of Theorem 2. Using the above lemma in Equation (12), we have:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \sum_{t=1}^{T} \left[\frac{1}{2\alpha_t(1-\beta_{1t})} \left[\|\hat{V}_t^{1/4}(x_t - x^*)\|^2 - \|\hat{V}_t^{1/4}(x_{t+1} - x^*)\|^2 \right] \right.$$
$$\left. + \frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right] + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

$$= \frac{1}{2\alpha_1(1-\beta_1)} \|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \frac{1}{2} \sum_{t=2}^{T} \left[\frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1t})} - \frac{\|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2}{\alpha_{t-1}(1-\beta_{1(t-1)})} \right]$$
$$+ \sum_{t=1}^{T} \left[\frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right] + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

$$= \frac{1}{2\alpha_1(1-\beta_1)} \|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \frac{1}{2} \sum_{t=2}^{T} \left[\frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1(t-1)})} - \frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1(t-1)})} + \frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1t})} \right.$$
$$\left. - \frac{\|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2}{\alpha_{t-1}(1-\beta_{1(t-1)})} \right] + \sum_{t=1}^{T} \left[\frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})} \|\hat{V}_t^{1/4}(x_t - x^*)\|^2 \right] + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|g_{1:T,i}\|_2.$$

Now, noting that $\beta_{1t} \leq \beta_1$, and using the observations that δ_t as defined in Equa-

tion (3) has nonnegative entries for all t (i.e. $\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t} \geq \frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}$) and that

$$\frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1t})} - \frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1(t-1)})} \leq \frac{\beta_{1t}}{\alpha_t(1-\beta_1)^2}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2,$$

we have that

$$\begin{aligned}
&\frac{1}{2\alpha_1(1-\beta_1)}\|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \frac{1}{2}\sum_{t=2}^{T}\left[\frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1(t-1)})} - \frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1(t-1)})} + \frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t(1-\beta_{1t})}\right. \\
&\left. - \frac{\|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2}{\alpha_{t-1}(1-\beta_{1(t-1)})}\right] + \sum_{t=1}^{T}\left[\frac{\beta_{1t}}{2\alpha_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2\right] + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2 \\
&\leq \frac{1}{2\alpha_1(1-\beta_1)}\|\hat{V}_1^{1/4}(x_1 - x^*)\|^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\left[\frac{\|\hat{V}_t^{1/4}(x_t - x^*)\|^2}{\alpha_t} - \frac{\|\hat{V}_{t-1}^{1/4}(x_t - x^*)\|^2}{\alpha_{t-1}}\right] \\
&\qquad + \sum_{t=1}^{T}\left[\frac{\beta_{1t}}{\alpha_t(1-\beta_1)^2}\|\hat{V}_t^{1/4}(x_t - x^*)\|^2\right] + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2 \\
&= \frac{1}{2\alpha_1(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{1,i}^{1/2}(x_{1,i} - x_i^*)^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}(x_{t,i} - x_i^*)^2\left[\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t} - \frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right] \\
&\qquad + \frac{1}{(1-\beta_1)^2}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}(x_{t,i} - x_i^*)^2\hat{v}_{t,i}^{1/2}}{\alpha_t} + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2.
\end{aligned} \tag{15}$$

Now, using the D_∞ bound on the diameter of the feasible region $\mathcal{F}$ to further simplify the bound in Equation (15), we have:

$$\begin{aligned}
\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) &\leq \frac{1}{2\alpha_1(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{1,i}^{1/2}D_\infty^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}D_\infty^2\left[\frac{\hat{v}_{t,i}^{1/2}}{\alpha_t} - \frac{\hat{v}_{t-1,i}^{1/2}}{\alpha_{t-1}}\right] \\
&\qquad + \frac{1}{(1-\beta_1)^2}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{D_\infty^2\beta_{1t}\hat{v}_{t,i}^{1/2}}{\alpha_t} + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2 \\
&= \frac{D_\infty^2}{2\alpha_T(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{D_\infty^2}{(1-\beta_1)^2}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}\hat{v}_{t,i}^{1/2}}{\alpha_t} + \frac{\alpha\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|g_{1:T,i}\|_2.
\end{aligned}$$

The equality follows from using a simple telescopic sum, yielding the desired result. We thus conclude from this bound that the worst case dependence of regret on T is $O(\sqrt{T})$. ■

5 Conclusion

In this project, we presented an introduction to online programming as well as the general structure of adaptive methods used to achieve minimum regret. Though online gradient descent and ADAGRAD exhibit good convergence properties, the rate of decay of their learning rate can lead to poor empirical performance, particularly when the loss functions are nonconvex or gradients are dense. ADAM, an algorithm introduced by Kingma and Ba (2015), is a popular adaptive method that uses exponential moving averages to update the parameters, thus mitigating the rapid decay of the learning rate by essentially limiting the reliance of the update to only the past few gradients.

However, Reddi et al. (2018) showed that ADAM can fail to converge (i.e. achieve zero average regret asymptotically) in certain settings due to δ_t as defined in Equation (3) potentially containing negative entries for certain time steps. In the proof of Theorem 1, we provided an example for non-convergence in a simple one-dimensional online convex programming framework.

To mitigate the issue of non-convergence of ADAM, Reddi et al. (2018) proposed the AMSGRAD algorithm, which uses the maximum of all v_t until the present time step and uses this maximum value for normalizing the running average of the gradient instead of v_t in ADAM. This algorithm has $\delta_t \geq 0$ for all time steps t, while also resolving the issue of rapidly deteriorating learning rate that online gradient descent and ADAGRAD exhibit. In Theorem 2, we presented the convergence analysis for AMSGRAD, showing that it can achieve worse case regret of $O(\sqrt{T})$ in a convex setting, and thus $R_T/T \to 0$ as $T \to \infty$.

References

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12 (7), 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations*, 2015.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and

beyond. In *Proceedings of 6th International Conference on Learning Representations*, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.