
Optimistic Acceleration of AMSGrad for Nonconvex Optimization.

Anonymous Author(s)

Affiliation

Address

email

1 Nonconvex Analysis

We tackle the following classical optimization problem:

$$\min_{w \in \Theta} f(w) := \mathbb{E}[f(w, \xi)] \quad (1)$$

where ξ is some random noise and only noisy versions of the objective function are accessible in this work. The objective function $f(w)$ is (potentially) nonconvex and has Lipschitz gradients.

Optimistic Algorithm We present here the algorithm studied in this paper to tackle problem (1). Set the terminating iteration number, $K \in \{0, \dots, K_{\max} - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\eta_k}{\sum_{f=0}^{K_{\max}-1} \eta_f}. \quad (2)$$

where $K_{\max} \leftarrow$ is the maximum number of iteration. The random termination number (2) is inspired by [Ghadimi and Lan, 2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization. Consider constants $(\beta_1, \beta_2) \in [0, 1]$, a sequence of decreasing stepsizes $\{\eta_k\}_{k>0}$, Algorithm 1 introduces the new optimistic AMSGrad method.

Algorithm 1 OPTIMISTIC-AMSGRAD

```
1: Input: Parameters  $\beta_1, \beta_2, \epsilon, \eta_k$ 
2: Init.:  $w_1 = w_{-1/2} \in \mathcal{K} \subseteq \mathbb{R}^d$  and  $v_0 = \epsilon \mathbf{1} \in \mathbb{R}^d$ 
3: for  $k = 0, 1, 2, \dots, K$  do
4:   Get mini-batch stochastic gradient  $g_k$  at  $w_k$ 
5:    $\theta_k = \beta_1 \theta_{k-1} + (1 - \beta_1) g_k$ 
6:    $v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2$ 
7:    $\hat{v}_k = \max(\hat{v}_{k-1}, v_k)$ 
8:    $w_{k+\frac{1}{2}} = \Pi_{\mathcal{K}} \left[ w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}} \right]$ 
9:    $w_{k+1} = \Pi_{\mathcal{K}} \left[ w_{k+\frac{1}{2}} - \eta_k \frac{h_{k+1}}{\sqrt{\hat{v}_k}} \right]$ 
10:   where  $h_{k+1} := \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$ 
11:   and  $m_{k+1}$  is a guess of  $g_{k+1}$ 
12: end for
13: Return:  $w_{K+1}$ .
```

The final update at iteration k can be summarized as:

$$w_{k+1} = w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}} - \eta_k \frac{h_{k+1}}{\sqrt{\hat{v}_k}} \quad (3)$$

We make the following assumptions:

- 13 **H1.** The loss function $f(w)$ is nonconvex w.r.t. the parameter w .
 14 **H2.** The function $f(w)$ is L -smooth w.r.t. the parameter w . There exist some constant $L > 0$ such
 15 that for $(w, \vartheta) \in \Theta^2$:

$$f(w) - f(\vartheta) - \nabla f(\vartheta)^\top (w - \vartheta) \leq \frac{L}{2} \|w - \vartheta\|^2. \quad (4)$$

H3. There exists a constant $a > 0$ such that for any $k > 0$:

$$\|m_{k+1}\| \leq a \|g_{k+1}\|$$

- 16 Classically (see [Ghadimi and Lan, 2013]) in nonconvex optimization, we make an assumption on
 17 the magnitude of the gradient:

H4. There exists a constant $M > 0$ such that

$$\|\nabla f(w, \xi)\| < M \quad \text{for any } w \text{ and } \xi$$

- 18 We begin with some auxiliary Lemmas important for the analysis. The first one ensures bounded
 19 norms of various quantities of interests (boiling down from the classical stochastic gradient bound-
 20 edness assumption):

Lemma 1. Assume assumption H 4, then the quantities defined in Algorithm 1 satisfy for any $w \in \Theta$ and $k > 0$:

$$\|\nabla f(w)\| < M, \quad \|\theta_k\| < M^2, \quad \|\hat{v}_k\| < M.$$

Proof Assume assumption H 4 we have:

$$\|\nabla f(w)\| = \|\mathbb{E}[\nabla f(w, \xi)]\| \leq \mathbb{E}[\|\nabla f(w, \xi)\|] \leq M$$

- 21 By induction reasoning, since $\|\theta_0\| = 0 \leq M$ and suppose that for $\|\theta_k\| \leq M$ then we have

$$\|\theta_{k+1}\| = \|\beta_1 \theta_k + (1 - \beta_1) g_{k+1}\| \leq \beta_1 \|\theta_k\| + (1 - \beta_1) \|g_{k+1}\| \leq M \quad (5)$$

- 22 Using the same induction reasoning we prove that

$$\|\hat{v}_{k+1}\| = \|\beta_2 \hat{v}_k + (1 - \beta_2) g_{k+1}^2\| \leq \beta_2 \|\hat{v}_k\| + (1 - \beta_1) \|g_{k+1}^2\| \leq M^2 \quad (6)$$

23 \square

- 24 Then, following [Yan et al., 2018] and their study of the SGD with Momentum (not AMSGrad but
 25 simple momentum) we denote for any $k > 0$:

$$\bar{w}_k = w_k + \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) = \frac{1}{1 - \beta_1} w_k - \frac{\beta_1}{1 - \beta_1} w_{k-1}, \quad (7)$$

- 26 and derive an important Lemma:

- 27 **Lemma 2.** Assume a strictly positive and non increasing sequence of stepsizes $\{\eta_k\}_{k>0}$, $\beta \in [0, 1]$,
 28 then the following holds:

$$\bar{w}_{k+1} - \bar{w}_k = \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k, \quad (8)$$

- 29 where $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$ and $\tilde{g}_k = g_k - \beta_1 g_{k-1}$.

- 30 **Proof** By definition (7) and using the Algorithm updates, we have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{1}{1 - \beta_1} (w_{k+1} - w_k) - \frac{\beta_1}{1 - \beta_1} (w_k - w_{k-1}) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + h_{k+1}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + h_k) \\ &= -\frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (\theta_k + \beta_1 \theta_{k-1}) - \frac{1}{1 - \beta_1} \eta_k \hat{v}_k^{-1/2} (1 - \beta_1) m_{k+1} \\ &\quad + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (\theta_{k-1} + \beta_1 \theta_{k-2}) + \frac{\beta_1}{1 - \beta_1} \eta_{k-1} \hat{v}_{k-1}^{-1/2} (1 - \beta_1) m_k \end{aligned} \quad (9)$$

31 Denote $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$ and $\tilde{g}_k = g_k - \beta_1 g_{k-1}$. Notice that $\tilde{\theta}_k = \beta_1 \tilde{\theta}_{k-1} +$
 32 $(1 - \beta_1)(g_k + \beta_1 g_{k-1})$.

$$\bar{w}_{k+1} - \bar{w}_k = \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} \hat{v}_{k-1}^{-1/2} - \eta_k \hat{v}_k^{-1/2} \right] - \eta_k \hat{v}_k^{-1/2} \tilde{g}_k \quad (10)$$

33 □

34 **Lemma 3.** Assume H 4, a strictly positive and non increasing sequence of stepsizes $\{\eta_k\}_{k>0}$,
 35 $\beta \in [0, 1]$, then the following holds:

$$\sum_{k=1}^K \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] \leq \frac{\eta^2 d K (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \quad (11)$$

36 **Proof** We denote by index $p \in [1, d]$ the dimension of each component of vectors of interest. Noting
 37 that for any $k > 0$ and dimension p we have $\hat{v}_{k,p} \geq v_{k,p}$, then:

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &= \eta_k^2 \mathbb{E} \left[\sum_{p=1}^d \frac{\theta_{k,p}^2}{\hat{v}_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[\sum_{p=1}^d \frac{\theta_{k,p}^2}{v_{k,p}} \right] \\ &\leq \eta_k^2 \mathbb{E} \left[\sum_{p=1}^d \frac{(\sum_{t=1}^k (1 - \beta_1) \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k (1 - \beta_2) \beta_2^{k-t} g_{t,p}^2} \right] \end{aligned} \quad (12)$$

38 where the last inequality is due to initializations. Denote $\gamma = \frac{\beta_1}{\beta_2}$. Then,

$$\begin{aligned} \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 (1 - \beta_1)^2}{1 - \beta_2} \mathbb{E} \left[\sum_{p=1}^d \frac{(\sum_{t=1}^k \beta_1^{k-t} g_{t,p})^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\stackrel{(a)}{\leq} \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{p=1}^d \frac{\sum_{t=1}^k \beta_1^{k-t} g_{t,p}^2}{\sum_{t=1}^k \beta_2^{k-t} g_{t,p}^2} \right] \\ &\leq \frac{\eta_k^2 (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{p=1}^d \sum_{t=1}^k \gamma^{k-t} \right] = \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=1}^k \gamma^{k-t} \right] \end{aligned} \quad (13)$$

39 where (a) is due to $\sum_{t=1}^k \beta_1^{k-t} \leq \frac{1}{1 - \beta_1}$. Summing from $k = 1$ to $k = K$ on both sides yields:

$$\begin{aligned} \sum_{k=1}^K \eta_k^2 \mathbb{E} \left[\left\| \hat{v}_k^{-1/2} \theta_k \right\|_2^2 \right] &\leq \frac{\eta_k^2 d (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{k=1}^K \sum_{t=1}^k \gamma^{k-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{1 - \beta_2} \mathbb{E} \left[\sum_{t=1}^K \gamma^{K-t} \right] \\ &\leq \frac{\eta^2 d K (1 - \beta_1)}{(1 - \beta_2)(1 - \gamma)} \end{aligned} \quad (14)$$

40 where the last inequality is due to $\sum_{t=1}^k \gamma^{k-t} \leq \frac{1}{1 - \gamma}$ as a consequence of the definition of γ . □

41 We now formulate the main result of our paper giving a finite-time upper bound of the quantity
 42 $\mathbb{E} [\|\nabla f(w_K)\|^2]$ where K is a random termination number distributed according to 2, see [Ghadimi
 43 and Lan, 2013].

44 **Theorem 1.** Assume H 2-H 4, $(\beta_1, \beta_2) \in [0, 1]$ and a sequence of decreasing stepsizes $\{\eta_k\}_{k>0}$,
 45 then the following result holds:

$$\mathbb{E} [\|\nabla f(w_K)\|^2] \leq \text{tocomplete} \quad (15)$$

46 **Proof** Using H 2 and the iterate \bar{w}_k we have:

$$\begin{aligned} f(\bar{w}_{k+1}) &\leq f(\bar{w}_k) + \nabla f(\bar{w}_k)^\top (\bar{w}_{k+1} - \bar{w}_k) + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 \\ &\leq f(\bar{w}_k) + \underbrace{\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k)}_A + \underbrace{(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k)}_B + \frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\| \end{aligned} \quad (16)$$

47 **Term A.** Using Lemma 2, we have that:

$$\begin{aligned} \nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) &= \nabla f(w_k)^\top \left[\frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \right] \\ &\leq \frac{\beta_1}{1 - \beta_1} \|\nabla f(w_k)\| \left\| \eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right\| \left\| \tilde{\theta}_{k-1} \right\| - \nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k \end{aligned} \quad (17)$$

48 where the inequality is due to trivial inequality for positive diagonal matrix. Using Lemma 1 and
49 assumption H3 we obtain:

$$\nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \frac{\beta_1(1 + \beta_1 + a)}{1 - \beta_1} M^2 \left[\left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] - \nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k \quad (18)$$

50 where we have used the fact that $\eta_k v_k^{-1/2}$ is a diagonal matrix such that $\eta_{k-1} v_{k-1}^{-1/2} \succcurlyeq \eta_k v_k^{-1/2} \succcurlyeq 0$
51 (decreasing stepsize and max operator). Also note that:

$$\begin{aligned} -\nabla f(w_k)^\top \eta_k v_k^{-1/2} \tilde{g}_k &= -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k - \nabla f(w_k)^\top \left[\eta_k v_k^{-1/2} - \eta_{k-1} v_{k-1}^{-1/2} \right] \tilde{g}_k \\ &\leq -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k + (1 - \beta_1) M^2 \left[\left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] \end{aligned} \quad (19)$$

52 using Lemma 1 on $\|g_k\|$ and recalling that $\tilde{g}_k = g_k - \beta_1 g_{k-1}$. Plugging (19) into (18) yields:

$$\begin{aligned} \nabla f(w_k)^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq -\nabla f(w_k)^\top \eta_{k-1} v_{k-1}^{-1/2} \tilde{g}_k + \frac{1}{1 - \beta_1} (\beta_1^2 + a\beta_1 + 1) M^2 \left[\left\| \eta_{k-1} v_{k-1}^{-1/2} \right\| - \left\| \eta_k v_k^{-1/2} \right\| \right] \end{aligned} \quad (20)$$

53 **Term B.** By Cauchy-Schwarz (CS) inequality we have:

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| \|\bar{w}_{k+1} - \bar{w}_k\| \quad (21)$$

54 Using smoothness assumption H 2:

$$\begin{aligned} \|\nabla f(\bar{w}_k) - \nabla f(w_k)\| &\leq L \|\bar{w}_k - w_k\| \\ &\leq L \frac{\beta_1}{1 - \beta_1} \|w_k - w_{k-1}\| \end{aligned} \quad (22)$$

55 By Lemma 2 we also have:

$$\begin{aligned} \bar{w}_{k+1} - \bar{w}_k &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \tilde{\theta}_{k-1} \eta_{k-1} v_{k-1}^{-1/2} \left[I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \\ &= \frac{\beta_1}{1 - \beta_1} \left[I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right] (w_{k-1} - w_k) - \eta_k v_k^{-1/2} \tilde{g}_k \end{aligned} \quad (23)$$

56 where the last equality is due to $\tilde{\theta}_{k-1} \eta_{k-1} v_{k-1}^{-1/2} = w_{k-1} - w_k$ by construction of $\tilde{\theta}_k$. Taking the
57 norms on both sides, observing $\left\| I - (\eta_k v_k^{-1/2})(\eta_{k-1} v_{k-1}^{-1/2})^{-1} \right\| \leq 1$ due to the decreasing stepsize
58 and the construction of \hat{v}_k and using CS inequality yield:

$$\|\bar{w}_{k+1} - \bar{w}_k\| \leq \frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \quad (24)$$

59 Plugging (22) and (24) into (21) returns:

$$\begin{aligned}
(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) &\leq L \frac{\beta_1}{1 - \beta_1} \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \|w_k - w_{k-1}\| \\
&\quad + L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2
\end{aligned} \tag{25}$$

We recall Young's inequality with a constant $\delta \in (0, 1)$ as follows:

$$\langle X | Y \rangle \leq \frac{1}{\delta} \|X\|^2 + \delta \|Y\|^2$$

60 Applying Young's inequality with $\delta \rightarrow \frac{\beta_1}{1 - \beta_1}$ on the product $\left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \|w_k - w_{k-1}\|$ yields:

$$(\nabla f(\bar{w}_k) - \nabla f(w_k))^\top (\bar{w}_{k+1} - \bar{w}_k) \leq L \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\|^2 + 2L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2 \tag{26}$$

61 The last term $\frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|$ can be upper bounded using (24):

$$\begin{aligned}
\frac{L}{2} \|\bar{w}_{k+1} - \bar{w}_k\|^2 &\leq \frac{L}{2} \left[\frac{\beta_1}{1 - \beta_1} \|w_{k-1} - w_k\| + \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\| \right] \\
&\leq L \left\| \eta_k v_k^{-1/2} \tilde{g}_k \right\|^2 + 2L \left(\frac{\beta_1}{1 - \beta_1} \right)^2 \|w_{k-1} - w_k\|^2
\end{aligned} \tag{27}$$

62

□

63 Plugging (20), (26) and (27) into (16) and taking the expectations on both sides give:

$$\mathbb{E}[f(\bar{w}_{k+1})] \leq \tag{28}$$

64 2 Containment of the iterates for a Deep Neural Network

65 **References**

- 66 S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic pro-
67 gramming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- 68 Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods
69 for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.