Optimistic Acceleration of AMSGrad for Nonconvex Optimization.

Anonymous Author(s)
Affiliation
Address
email

1 Nonconvex Analysis

We tackle the following classical optimization problem:

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

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, $K \in \{0, \dots, K_{\mathsf{max}} - 1\}$, as a discrete r.v. with:

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

where $K_{\mathsf{max}} \leftarrow$ 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

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

The final update at iteration k can be summarized as:

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

We make the following assumptions:

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

H1. *The loss function* $f(w)$ *is nonconvex* w.r.t. *the parameter* w.

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

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

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

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

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

H4. *There exists a constant* $\mathsf{M} > 0$ *such that*

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

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

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

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

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

$$\overline{w}_k = w_k + \frac{\beta_1}{1-\beta_1}(w_k - w_{k-1}) = \frac{1}{1-\beta_1} w_k - \frac{\beta_1}{1-\beta_1} w_{k-1} \ , \tag{5}$$

and derive an important Lemma:

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

$$\overline{w}_{k+1} - \overline{w}_k = \frac{\beta_1}{1-\beta_1} \tilde{\theta}_{k-1} \left[\eta_{k-1} v_{k-1}^{-1/2} - \eta_k v_k^{-1/2} \right] - \eta_k v_k^{-1/2} \tilde{g}_k \ , \tag{6}$$

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

Proof By definition (5) and using the Algorithm updates, we have:

$$\begin{aligned}
\overline{w}_{k+1} - \overline{w}_k &= \frac{1}{1-\beta_1}(w_{k+1} - w_k) - \frac{\beta_1}{1-\beta_1}(w_k - w_{k-1}) \\
&= -\frac{1}{1-\beta_1} \eta_k v_k^{-1/2} (\theta_k + h_{k+1}) + \frac{\beta_1}{1-\beta_1} \eta_{k-1} v_{k-1}^{-1/2} (\theta_{k-1} + h_k) \\
&= -\frac{1}{1-\beta_1} \eta_k v_k^{-1/2} (\theta_k + \beta_1 \theta_{k-1}) - \frac{1}{1-\beta_1} \eta_k v_k^{-1/2} (1-\beta_1) m_{k+1} \\
&+ \frac{\beta_1}{1-\beta_1} \eta_{k-1} v_{k-1}^{-1/2} (\theta_{k-1} + \beta_1 \theta_{k-2}) + \frac{\beta_1}{1-\beta_1} \eta_{k-1} v_{k-1}^{-1/2} (1-\beta_1) m_k
\end{aligned} \tag{7}$$

Denote $\tilde{\theta}_k = \theta_k + \beta_1 \theta_{k-1}$ we notice that $\tilde{\theta}_k = \beta_1 \tilde{\theta}_{k-1} + (1-\beta_1)(g_k + \beta_1 g_{k-1})$.

$$\overline{w}_{k+1} - \overline{w}_k \leq \tag{8}$$

$\square$

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

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

$$\mathbb{E}\left[\|\nabla f(w_K)\|^2\right] \leq tocomplete \tag{9}$$

Proof Using H 2 and the iterate $\overline{w}_k$ we have:

$$\begin{aligned} f(\overline{w}_{k+1}) &\leq f(\overline{w}_k) + \nabla f(\overline{w}_k)^\top (\overline{w}_{k+1} - \overline{w}_k) + \frac{L}{2}\|\overline{w}_{k+1} - \overline{w}_k\|^2 \\ &\leq f(\overline{w}_k) + \underbrace{\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k)}_{A} + \underbrace{\left(\nabla f(\overline{w}_k) - \nabla f(w_k)\right)^\top (\overline{w}_{k+1} - \overline{w}_k)}_{B} + \frac{L}{2}\|\overline{w}_{k+1} - \overline{w}_k\| \end{aligned} \tag{10}$$

Term A. Using Lemma 2, we have that:

$$\begin{aligned} \nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k) &= \nabla f(w_k)^\top \left[\frac{\beta_1}{1-\beta_1}\tilde{\theta}_{k-1}\left[\eta_{k-1}v_{k-1}^{-1/2} - \eta_k v_k^{-1/2}\right] - \eta_k v_k^{-1/2}\tilde{g}_k\right] \\ &\leq \frac{\beta_1}{1-\beta_1}\|\nabla f(w_k)\|\left\|\eta_{k-1}v_{k-1}^{-1/2} - \eta_k v_k^{-1/2}\right\|\left\|\tilde{\theta}_{k-1}\right\| - \nabla f(w_k)^\top \eta_k v_k^{-1/2}\tilde{g}_k \end{aligned} \tag{11}$$

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

$$\nabla f(w_k)^\top (\overline{w}_{k+1} - \overline{w}_k) \leq \frac{\beta_1(a+2)}{1-\beta_1}\mathsf{M}^2\left[\left\|\eta_{k-1}v_{k-1}^{-1/2}\right\| - \left\|\eta_k v_k^{-1/2}\right\|\right] - \nabla f(w_k)^\top \eta_k v_k^{-1/2}\tilde{g}_k \tag{12}$$

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

$$\begin{aligned} -\nabla f(w_k)^\top \eta_k v_k^{-1/2}\tilde{g}_k &= -\nabla f(w_k)^\top \eta_{k-1}v_{k-1}^{-1/2}\tilde{g}_k - -\nabla f(w_k)^\top \left[\eta_k v_k^{-1/2} - \eta_{k-1}v_{k-1}^{-1/2}\right]\tilde{g}_k \\ &\leq -\nabla f(w_k)^\top \eta_{k-1}v_{k-1}^{-1/2}\tilde{g}_k + (1-\beta_1)\mathsf{M}^2\left[\left\|\eta_{k-1}v_{k-1}^{-1/2}\right\| - \left\|\eta_k v_k^{-1/2}\right\|\right] \end{aligned} \tag{13}$$

using Lemma 1 on $\|g_k\|$ and recalling that $\tilde{g}_k = g_k - \beta_1 g_{k-1}$ □

2 Containment of the iterates for a DNN

References

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.