

---

# Optimistic Acceleration of AMSGrad: Theory and Applications.

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Algorithm

Set the terminating iteration number,  $K \in \{0, \dots, K_{\max} - 1\}$ , as a discrete r.v. with:

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

where  $K_{\max} \leftarrow$  is the maximum number of iteration. The random termination number (1) is inspired by [Ghadimi and Lan, 2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization.

---

### 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:    $w_{k+1} = \Pi_{\mathcal{K}} \left[ w_{k+\frac{1}{2}} - \eta_{k+1} \frac{h_{k+1}}{\sqrt{v_k}} \right]$ 
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:

$$w_{k+1} = w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}} - \eta_{k+1} \frac{h_{k+1}}{\sqrt{v_k}} \quad (2)$$

## 2 Nonconvex Analysis

### 2.1 Containment of the iterates for a DNN

### 2.2 Non Asymptotic analysis

10 **References**

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