Optimistic Acceleration of AMSGrad: Theory and Applications.

Anonymous Author(s)
Affiliation
Address
email

1 Algorithm

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

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

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

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 $k = 0, 1, 2, \ldots, K$ **do**
  Get mini-batch stochastic gradient g_k at w_k
  $\theta_k = \beta_1 \theta_{k-1} + (1 - \beta_1)\, g_k$
  $v_k = \beta_2 v_{k-1} + (1 - \beta_2)\, g_k^2$
  $\hat{v}_k = \max\left(\hat{v}_{k-1}, v_k\right)$
  $w_{k+\frac{1}{2}} = \Pi_{\mathcal{K}}\left[w_k - \eta_k \frac{\theta_k}{\sqrt{\hat{v}_k}}\right]$
  $w_{k+1} = \Pi_{\mathcal{K}}\left[w_{k+\frac{1}{2}} - \eta_{k+1} \frac{h_{k+1}}{\sqrt{v}_k}\right]$
    where$h_{k+1} := \beta_1 \theta_{k-1} + (1 - \beta_1) m_{k+1}$
      andm_{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:

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

2 Nonconvex Analysis

2.1 Containment of the iterates for a DNN

2.2 Non Asymptotic analysis

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

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.