Stochastic Gradient Descent with Momentum Convergence Diagnostic for Nonconvex Optimization

Anonymous Author(s)
Affiliation
Address
`email`

1 Nonconvex case

We recall the SGD with Momentum update we are analyzing here:

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla \ell \left(\theta_n, \xi_{n+1}\right) + \beta \left(\theta_n - \theta_{n-1}\right) \tag{1}$$

where ℓ is the nonconvex loss function parametrized by $\theta \in \Theta \subset \mathbb{R}^p$ and ξ is some random noise. $\beta \in [0, 1)$ is the momentum parameter and γ_{n+1} the learning stepsize.

We also define $f(\theta) = \mathbb{E}[\ell(\theta, \xi]$ the expected loss. Following Pflug convergence diagnostic test, we construct the following test statistics:

$$\nabla \ell \left(\theta_n, \xi_{n+1}\right)^\top \nabla \ell \left(\theta_{n-1}, \xi_n\right) \tag{2}$$

and the goal will be to upperbound the expectation of this quantity in order to spot the two different phases through the iterates.

We make the following assumptions before analyzing (2).

H1. *The loss function $\ell(\theta, \xi)$ is nonconvex* w.r.t. *the parameter θ.*

We consider the very general setting where the loss function $\ell(;\xi)$ is (l, L)-smooth, see [Allen-Zhu, 2017, Zhou and Gu, 2019]

H2. *There exist some constant $l \in \mathbb{R}$ and $L > 0$ such that for $(\theta, \vartheta) \in \Theta^2$:*

$$\frac{l}{2} \left\|\theta - \vartheta\right\|^2 \leq \ell(\theta) - \ell(\vartheta) - \nabla \ell(\vartheta)^\top (\theta - \vartheta) \leq \frac{L}{2} \left\|\theta - \vartheta\right\|^2 \tag{3}$$

Note that if $l = -L$ we recover the conventional L-smoothness definition and if $l \geq 0$ (resp. $l > 0$) we have convexity (resp. strong convexity).

H3. *There exists $K > 1$ such that*

$$\mathbb{E}\left[\left(\theta_n - \theta_{n-1}\right)^\top \left(\theta_{n-1} - \theta_{n-2}\right)\right] \geq -K \mathbb{E}\left[\left\|\theta_n - \theta_{n-1}\right\|^2\right]$$

for large enough iteration index n.

Finally and classically in nonconvex optimization, we assume a bounded gradient term.

H4. *There exists a constant $G > 0$ such that*

$$\|\nabla \ell(\theta, \xi)\| < G \quad \textit{for any } \theta \textit{ and } \xi$$

We can easily check the following identity:

$$\nabla \ell \left(\theta_n, \xi_{n+1}\right)^\top \nabla \ell \left(\theta_{n-1}, \xi_n\right) = \frac{1}{\gamma} \nabla \ell \left(\theta_n, \xi_{n+1}\right)^\top \left(\theta_{n-1} - \theta_n\right) + \frac{\beta}{\gamma} \nabla \ell \left(\theta_n, \xi_{n+1}\right)^\top \left(\theta_{n-1} - \theta_{n-2}\right) \tag{4}$$

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

Taking expectations on both sides and using Assumption H 2, we have:

$$\begin{aligned}\mathbb{E}\left[\nabla\ell\left(\theta_n,\xi_{n+1}\right)^\top\nabla\ell\left(\theta_{n-1},\xi_n\right)\right]\leq&\frac{1}{\gamma}\left[f(\theta_{n-1})-f(\theta_n)-\frac{l}{2}\left\|\theta_{n-1}-\theta_n\right\|^2\right]\\&+\frac{\beta}{\gamma}\left[f(\theta_n)-f(\theta_n+\theta_{n-2}-\theta_{n-1})+\frac{L}{2}\left\|\theta_{n-1}-\theta_{n-2}\right\|^2\right]\end{aligned}\tag{5}$$

which yields:

$$\begin{aligned}\mathbb{E}\left[\nabla\ell\left(\theta_n,\xi_{n+1}\right)^\top\nabla\ell\left(\theta_{n-1},\xi_n\right)\right]\leq&\frac{1}{\gamma}\left[f(\theta_{n-1})-f(\theta^*)-\frac{l}{2}\left\|\theta_{n-1}-\theta_n\right\|^2\right]\\&+\frac{\beta}{\gamma}\left[f(\theta_n)-f(\theta^*)+\frac{L}{2}\left\|\theta_{n-1}-\theta_{n-2}\right\|^2\right]\end{aligned}\tag{6}$$

where θ^* is the global minimizer of the expected loss.

Denote $\Delta_n=\theta_n-\theta_{n-1}$ and observe that:

$$\left\|\Delta_n\right\|^2=\gamma^2\left\|\nabla\ell\left(\theta_{n-1},\xi_n\right)\right\|^2+2\beta\Delta_n^\top\Delta_{n-1}-\beta^2\left\|\Delta_{n-1}\right\|^2\tag{7}$$

Using assumptions H 4 and H 3 we obtain:

$$\mathbb{E}\left\|\Delta_n\right\|^2\leq\gamma^2G^2-(2\beta K+\beta^2)\mathbb{E}\left[\left\|\Delta_{n-1}\right\|^2\right]\tag{8}$$

References

Z. Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 89–97. JMLR. org, 2017.

D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:1901.11224*, 2019.