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 $(\theta, \vartheta) \in \Theta^2$:*

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

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

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

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

We recall an important convergence result for the SGD with Momentum update from [Yan et al., 2018]:

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

Theorem 1. *[Yan et al., 2018] Under assumptions H 1, H 2, H 4 and the boundedness of the variance of the stochastic gradients, we have*

$$\begin{aligned}\min_{k=0,\ldots,n} \mathrm{E}\left[\left\|\nabla \ell\left(\theta_k\right)\right\|^2\right] \leq & \frac{2\left(\ell\left(\theta_0\right)-\ell_*\right)(1-\beta)}{n+1} \max \left\{\frac{2L}{1-\beta}, \frac{\sqrt{n+1}}{C}\right\} \\ & +\frac{C}{\sqrt{n+1}} \frac{L\beta^2((1-\beta)s-1)^2\left(G^2+\sigma^2\right)+L\sigma^2(1-\beta)^2}{(1-\beta)^3}\end{aligned} \tag{4}$$

We can easily check the following identity:

$$\nabla \ell\left(\theta_n, \xi_{n+1}\right)^{\top} \nabla \ell\left(\theta_{n-1}, \xi_n\right)=\frac{1}{\gamma} \nabla \ell\left(\theta_n, \xi_{n+1}\right)^{\top}\left(\theta_{n-1}-\theta_n\right)+\frac{\beta}{\gamma} \nabla \ell\left(\theta_n, \xi_{n+1}\right)^{\top}\left(\theta_{n-1}-\theta_{n-2}\right) \tag{5}$$

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

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

which yields:

$$\begin{aligned}\mathbb{E}\left[\nabla \ell\left(\theta_n, \xi_{n+1}\right)^{\top} \nabla \ell\left(\theta_{n-1}, \xi_n\right)\right] \leq & \frac{1}{\gamma}\left[f(\theta_{n-1})-f(\theta^*)-\frac{l}{2}\left\|\theta_{n-1}-\theta_n\right\|^2\right] \\ & +\frac{\beta}{\gamma}\left[f(\theta_n)-f(\theta^*)+\frac{L}{2}\left\|\theta_{n-1}-\theta_{n-2}\right\|^2\right]\end{aligned} \tag{7}$$

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

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

$$\left\|\Delta_n\right\|^2=\gamma^2\left\|\nabla \ell\left(\theta_{n-1}, \xi_n\right)\right\|^2+2\beta \Delta_n^{\top} \Delta_{n-1}-\beta^2\left\|\Delta_{n-1}\right\|^2 \tag{8}$$

Using assumptions H 3 and using Theorem 1 we obtain:

$$\mathbb{E}\left\|\Delta_n\right\|^2 \leq \gamma^2 G^2-(2\beta K+\beta^2) \mathbb{E}\left[\left\|\Delta_{n-1}\right\|^2\right] \tag{9}$$

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.

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.

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