Additional Results for Convergence Diagnostic with SGDM

1 No Early Restarts with Pflug on SGDM

1.1 SGD for convex

In [Pesme et al., 2020], the authors claim that Pflug's statistic fails to detect convergence of a simple SGD procedure for convex objective functions:

$$\theta_{n+1} = \theta_n - \gamma \nabla \ell(\theta_n, \xi_{n+1}) \tag{1}$$

Also denote the noise term $\epsilon_n(\theta) = \nabla \ell(\theta_n, \xi_{n+1}) - \nabla \ell(\theta_n)$ as the gap between the stochastic gradient and the full one. On the simple example of a quadratic function and under the following assumption:

H1. *(Quadratic semi-stochastic setting). There exists a symmetric positive semi-definite matrix H such that $\ell(\theta) = \frac{1}{2}\theta^\top H \theta$ and the noise $\epsilon_n(\theta) = \epsilon_n$ is independent of θ with:*

$$(\epsilon_n)_{n \geq 0} \text{ are i.i.d.}, \mathbb{E}[\epsilon_n] = 0, \mathbb{E}\left[\epsilon_n^T \epsilon_n\right] = C. \tag{2}$$

And also:

H2. *Noise symmetry and continuity:*

$$\mathbb{P}\left(\epsilon_1^T \epsilon_2 \geq x\right) = \mathbb{P}\left(\epsilon_1^T \epsilon_2 \leq -x\right) \text{ for all } x \geq 0$$

Then under H 1 and H 2, they prove:

Proposition 1. *Assume an initial point $\theta_0 \sim \pi_{old}$ sampled from the stationary distribution π_{old} for a SGD trajectory ran with a constant stepsize γ_{old} and run SGD with the new decayed stepsize $\gamma = r \times \gamma_{old}$. Then for any $0 < \alpha < 2$ and iteration number $n_\gamma = O(\gamma^{-\alpha})$ we have:*

$$\lim_{\gamma \to 0} \mathbb{P}_{\theta_0 \sim \pi_{\gamma_{old}}}\left(S_{n_\gamma} \leq 0\right) = \frac{1}{2}$$

where S_{n_γ} is the Pflug statistic.

The signal during the transient phase is positive and if order $O(\gamma)$. However the variance of Sn is $O(1/n)$. Hence $\Omega(1/\gamma^2)$ iterations are typically needed in order to have a clean signal. Then, the main claim of this Proposition is that before this threshold, S_n resembles a random walk and its sign gives no information on whether saturation is reached or not, this leads to early on restart

1.2 SGD with Momentum for convex

$$\theta_{n+1} = \theta_n - \gamma \nabla \ell(\theta_n, \xi_{n+1}) + \beta(\theta_n - \theta_{n-1}) \tag{3}$$

We ought to show that with our modified Pflug statistics, and the consideration of a momentum based SGD algorithm, the convergence diagnostic does not fail, i.e. there are no early restarts.

Consider the same set of assumptions H 1 and H 2 with the update (3). We show the following:

Proposition 2. *Assume an initial point* $\theta_0 \sim \pi_{old}$ *sampled from the stationary distribution* π_{old} *run SGDM with the decayed stepsize* $\gamma = r \times \gamma_{old}$*. Then for any* $0 < \alpha < 2$ *and iteration number* $n_\gamma = O(\gamma^{-\alpha})$ *we have:*

$$\lim_{\gamma \to 0} \mathbb{P}_{\theta_0 \sim \pi_{\gamma_{old}}} \left(S_{n_\gamma} \leq 0 \right) = 0$$

where $\bar{S}_{n_\gamma} = \nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)$ *is our modified Pflug statistic.*

The proof goes as follows:

Proof. □

References

[Pesme et al., 2020] Pesme, S., Dieuleveut, A., and Flammarion, N. (2020). On convergence-diagnostic based step sizes for stochastic gradient descent. *arXiv preprint arXiv:2007.00534*.