Benefits of staleness and asynchrony in machine learning algorithms

Hongyi Wang, Yuzhe Ma, Xin Jin

November 3rd, 2017

Abstract

The momentum method is widely used in optimization to accelerate the stochastic gradient descent (SGD). By inheriting part of the previous computed gradients, the momentum can improve the convergence rate while maintaining the optimality of the solution. Previous work have verified the usefulness of momentum and show that the momentum term is typically set to 0.9 or a similar value. We study the problem of how to exploit the momentum in the distributed machine learning setting, where stragglers are common and the computed gradient could be out of date. In particular, we show that by using a dynamic momentum term that is dependent on the date of the computed gradient, one can improve the convergence rate without waiting for the stragglers.

1 A Motivation Example

In this section, we motivate the usefulness of the momentum method. We first give the following general framework for optimization:

$$\min_x f(x) \tag{1}$$

$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, 2, ...m \tag{2}$$

$$h_j(x) = 0 \quad j = 1, 2, ...n, \tag{3}$$

where $f(x)$ is the objective to be minimized and $g(x)$ and $h(x)$ are the inequality and equality constraints. For ease of explanation, we first consider an unconstrained problem where $f(x) = x^2$. The gradient is $\nabla_x f(x) = 2x$ and thus the gradient descent is simply $x_{t+1} = x_t - \alpha \nabla_x f(x_t) = x_t - 2\alpha x_t = (1 - 2\alpha)x_t$, where α is the step size. Note that if $\alpha < \frac{1}{2}$, then the solution converges to 0, which is the global optimum. Now we introduce the momentum term: $M_t = \gamma \nabla_x f(x_{t-1})$ with $\gamma \in (0, 1)$. The M_t inherits a fraction of $\nabla_x f(x_{t-1})$ and is then combined with the gradient at time t: $\tilde{\nabla}_x f(x_t) = \alpha \nabla_x f(x_t) + M_t$, where we use $\tilde{\nabla}_x f(x_t)$ to distinguish it from the true gradient $\nabla_x f(x_t)$. Then the gradient descent is performed with $\tilde{\nabla}_x f(x_t)$ instead of $\nabla_x f(x_t)$: $x_{t+1} = x_t - \tilde{\nabla}_x f(x_t) = x_t - \alpha \nabla_x f(x_t) - \gamma \nabla_x f(x_{t-1}) = (1 - 2\alpha)x_t - 2\gamma x_{t-1}$. Intuitively, this would speed up the convergence rate if α is small since by adding another term $2\gamma x_{t-1}$, x_t is closer to 0. To make it clear, we assume

$x_0 = 8$, $\alpha = \frac{1}{4}$ and $\gamma = \frac{1}{32}$, this gradient descent with and without the momentum term are as follows respectively:

$$x_{t+1} = x_t - \alpha \nabla_x f(x_t) = \frac{1}{2} x_t \tag{4}$$

$$\begin{aligned} x_{t+1} &= x_t - \alpha \nabla_x f(x_t) - \gamma \nabla_x f(x_{t-1}) \\ &= \frac{1}{2} x_t - \frac{1}{16} x_{t-1} \end{aligned} \tag{5}$$

Solving (4) gives $x_t = (\frac{1}{2})^{t-3}$. Note that to achieve a solution within $[0, \epsilon]$, we need at least $\log_2 \frac{1}{8\epsilon}$ iterations. Now consider (5), assume that $x_1 = \frac{1}{2} x_0 = 4$, meaning from x_0 to x_1 we perform vanilla gradient descent since we do not have a previous gradient. Solving (5) gives $x_t = (2t + 2)(\frac{1}{4})^{t-1}$, therefore asymptotically, we only need $O(\log_4 \frac{1}{\epsilon})$ iterations to achieve a solution in $[0, \epsilon]$. Though the improvement is only a constant in this example, it can be significant in a distributed setting, where stragglers can be a huge drawback if the system waits for the completion of every working machine to compute their parts of gradients. However, by using the momentum, one can continuously running SGD with the momentums in hand, which could be some gradients out of date but are still useful.

To better exploit the momentum, we need to adjust the weight of the momentum in terms of its computation time. For example at iteration t, the most up-to-date momentum is M_{t-k}, then one should be convinced that the larger k is, the smaller the weight γ_k should be set since the momentum is less likely to be useful. Examples of γ_k are: $\gamma_k = \gamma e^{-k}$ (exponentially decreasing) or $\gamma_k = \gamma n^{-k}$ (polynomially decreasing), where γ is a global constant.

2 SGD with noised gradient

We do SGD with batch size of 1 (i.e. $B = 1$). Also, we focus our analysis on function f that are λ-strong convex, which means $f(\mathbf{w}) \geq f(\mathbf{w}') + \langle \nabla f(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{\lambda}{2} \|\mathbf{w} - \mathbf{w}'\|^2$: Also, we further assume $\mathbb{E}\|\nabla f(\mathbf{w})\|^2 \leq M^2$: Based on the update rule of traditional SGD, we have $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma \nabla f_{S_k}(\mathbf{w}_k)$: However, in our study, we only subset of gradients from workers gathered on the master node, we are now have gradient $\mathbf{g}_k$ for each iteration rather than $\nabla f_{S_k}(\mathbf{w}_k)$. Thus, currently our update role should be $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma \mathbf{g}_k$: Let $\mathbf{w}^*$ the weight vector, which gives us the global minimum value of the loss function.

$$\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}^*\|^2 = \mathbb{E}\|\mathbf{w}_k - \mathbf{w}^*\|^2 - 2\gamma \mathbb{E}\langle \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^* \rangle + \gamma^2 \mathbb{E}\|\mathbf{g}_k\|^2 \tag{6}$$

If we look at the most right item in the foregoing equation, and take $\mathbf{g}_k = (\mathbf{g}_k - \nabla f_{S_k}(\mathbf{w}_k)) + \nabla f_{S_k}(\mathbf{w}_k)$, using the fact that $\|\mathbf{g}_k - \nabla f_{S_k}(\mathbf{w}_k)\| \leq \epsilon \|\nabla f_{S_k}(\mathbf{w}_k)\|$, we

get:

$$\begin{aligned}\mathbb{E}\|\mathbf{g}_k\|^2 &= \mathbb{E}\|(\mathbf{g}_k - \nabla f_{S_k}(\mathbf{w}_k)) + \nabla f_{S_k}(\mathbf{w}_k)\|^2 \\ &= \mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k)\|^2 + 2\mathbb{E}\langle \nabla f_{S_k}(\mathbf{w}_k), \mathbf{g}_k - \nabla f_{S_k}(\mathbf{w}_k)\rangle + \mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k) - \mathbf{g}_k\|^2 \\ &\leq (\epsilon+1)^2\mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k)\|^2 \\ &\leq (\epsilon+1)^2 M^2\end{aligned}$$

Thus, equation (1) can be higher bounded by

$$\mathbb{E}[\Delta_k] - 2\gamma\mathbb{E}\langle \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^*\rangle + \gamma^2(\epsilon+1)^2M^2 \tag{7}$$

in which $\Delta_k = \|\mathbf{w}_k - \mathbf{w}^*\|^2$. Consider the second term in equation (2), and let $\mathbf{g}_k = \nabla f_{S_k}(\mathbf{w}_k) - (\nabla f_{S_k}(\mathbf{w}_k) - \mathbf{g}_k)$, then it becomes

$$\begin{aligned}\langle \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^*\rangle &= \langle \nabla f_{S_k}(\mathbf{w}_k) - (\nabla f_{S_k}(\mathbf{w}_k) - \mathbf{g}_k), \mathbf{w}_k - \mathbf{w}^*\rangle \\ &= \langle \nabla f_{S_k}(\mathbf{w}_k), \mathbf{w}_k - \mathbf{w}^*\rangle - \langle \nabla f_{S_k}(\mathbf{w}_k) - \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^*\rangle \\ &\geq \langle \nabla f_{S_k}(\mathbf{w}_k), \mathbf{w}_k - \mathbf{w}^*\rangle - \|\nabla f_{S_k}(\mathbf{w}_k) - \mathbf{g}_k\|\|\mathbf{w}_k - \mathbf{w}^*\| \\ &\geq \langle \nabla f_{S_k}(\mathbf{w}_k), \mathbf{w}_k - \mathbf{w}^*\rangle - \epsilon\|\nabla f_{S_k}(\mathbf{w}_k)\|\|\mathbf{w}_k - \mathbf{w}^*\| \\ &\geq \langle \nabla f_{S_k}(\mathbf{w}_k) - \nabla f_{S_k}(\mathbf{w}^*), \mathbf{w}_k - \mathbf{w}^*\rangle - \epsilon\|\nabla f_{S_k}(\mathbf{w}_k)\|\|\mathbf{w}_k - \mathbf{w}^*\|\end{aligned}$$

We add $f_{S_k}(\mathbf{w}^*)$ here because it equals to zero and adding it changes nothing. Plugging this into (2), we can upper bound (2) by:

$$\mathbb{E}[\Delta_k] - 2\gamma\mathbb{E}\langle \nabla f_{S_k}(\mathbf{w}_k) - \nabla f_{S_k}(\mathbf{w}^*), \mathbf{w}_k - \mathbf{w}^*\rangle + 2\gamma\epsilon\mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k)\|\|\mathbf{w}_k - \mathbf{w}^*\| + \gamma^2(\epsilon+1)^2M^2$$

using λ strong convex from our assumption we get:

$$\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}[\Delta_k] - 2\gamma\lambda\mathbb{E}[\Delta_k] + 2\gamma\epsilon\mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k)\|\|\mathbf{w}_k - \mathbf{w}^*\| + \gamma^2(\epsilon+1)^2M^2 \tag{8}$$

Before we tell something about ϵ, we need to firstly figure out the convergence rate of regular SGD (i.e. $\epsilon = 0$), which makes the equation (3) becomes:

$$\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}^*\|^2 \leq (1 - 2\gamma_k\lambda)\mathbb{E}[\Delta_k] + \gamma_k^2M^2$$

We write γ as γ_k here because we need to choose γ_k to get convergence rate of SGD, here we choose $\gamma_k = \frac{1}{\lambda k}$, We prove the convergence rate by induction. First it is easy to see that:

$$\|\mathbf{w}_1 - \mathbf{w}^*\| \leq \frac{\max\{\|\mathbf{w}_1 - \mathbf{w}^*\|^2, M^2/\lambda^2\}}{1}$$

Then, we assume that the convergence rate holds with k. Next we only need to show that it holds with $k+1$. Denote $L = \max\{\|x_1 - x^*\|^2, M^2/\lambda^2\}$, then based on (3), we

have:

$$\begin{aligned}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 &\leq (1-\frac{2}{k})\mathbb{E}\|\mathbf{w}_k-\mathbf{w}^*\|^2+\frac{M^2}{\lambda^2k^2}\\ &\leq (1-\frac{2}{k})\frac{L}{k}+\frac{M^2}{\lambda^2k^2}\\ &\leq (\frac{1}{k}-\frac{2}{k^2})L-\frac{L}{k^2}\\ &=\frac{L}{k+1}\end{aligned}$$

Thus, the convergence rate we get here is:

$$\mathbb{E}\|\mathbf{w}_k-\mathbf{w}^*\|\leq\frac{\max\{\|\mathbf{w}_1-\mathbf{w}^*\|^2,M^2/\lambda^2\}}{k}$$

Now, we focus on ϵ in equation (3), we bound the third term on the right hand side by:

$$\begin{aligned}2\gamma\epsilon\mathbb{E}\|\nabla f_{S_k}(\mathbf{w}_k)\|\|\mathbf{w}_k-\mathbf{w}^*\| &\leq \gamma\epsilon(\frac{\|\nabla f_{S_k}(\mathbf{w}_k)\|^2}{\frac{\epsilon}{\gamma}}+\frac{\epsilon}{\gamma}\|\mathbf{w}_k-\mathbf{w}^*\|^2)\\ &\leq \gamma^2M^2+\epsilon^2\|\mathbf{w}_k-\mathbf{w}^*\|^2\end{aligned}$$

Plugging this back into equation (3), then we get:

$$\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2\leq(1-2\gamma\lambda+\epsilon^2)\mathbb{E}[\Delta_k]+\gamma^2[(\epsilon+1)^2+1]M^2 \tag{9}$$

If we take ϵ proportional to $\gamma\lambda$, say $\alpha\gamma\lambda$, and take $\gamma=\frac{1}{\lambda k}$ then we get:

$$\gamma^2[(\epsilon+1)^2+1]M^2\leq(2\epsilon^2+3)\gamma^2M^2=2\alpha\gamma\lambda+3=\frac{2\alpha}{k}+3\leq C$$

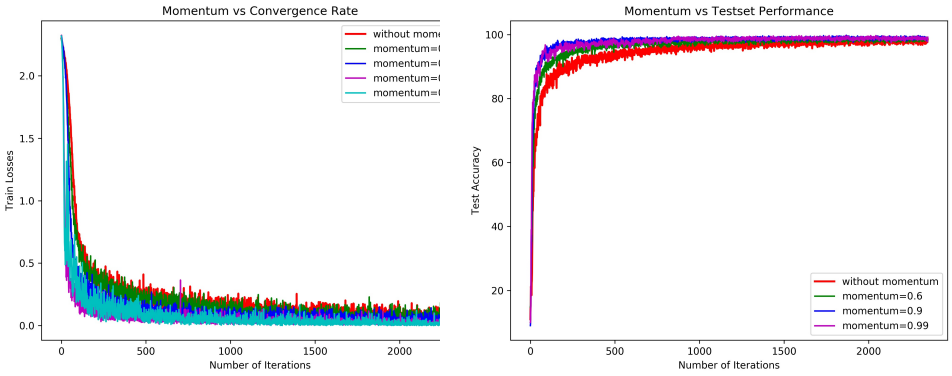
Then equation (4) become $\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2\leq(1-(2-\alpha)\gamma\lambda)\mathbb{E}[\Delta_k]+C$: Then the foregoing proof for convergence just follow.

3 Experiments

We firstly verify the relationship between momentum value and convergence rate of model. The result is given in Fig. 1, we can tell that with momentum around 0.9 we get the optimal (sub-optimal) convergence rate.

4 Our Next Steps

- Show the proof that how the optimal momentum will affect the convergence rate of model
- Set up experiment to show allowing appropriate amount of staleness gradient will lead to faster convergence rate



(a) Convergence Rate on Training Set

(b) Performance on Test Set

Figure 1: Relationship between convergence and Momentum value experiments running on MNIST and LeNet using mini-batched SGD (batch size at 256) with m4.2xlarge instance of AWS EC2