Accelerate Distributed Stochastic Descent for Nonconvex Optimization with Momentum

Guojing Cong [1], Tianyi Liu[2]

[1]IBM TJ Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY, 10598
[2]Georgia Institute of Technology, Atlanta, GA, 30332

MLHPC-SC2020

Outline

- Introduction – Background, existing approaches, and motivation
- Algorithm
- Theoretical analysis on convergence and scalability
- Experimental results
- Conclusions

Introduction

- Training deep neural networks is time consuming, and calls for distributed training.
- Of many algorithms that have been recently proposed, one main challenge remains that as the number of processors P increases, convergence suffers per constant number of samples processed
- How can we improve convergence speed?

Synchronous and asynchronous distributed training approaches

- Asynchronous SGD – Downpour, Hogwild!, Elastic averaging SGD, and Decentralized methods
- Synchronous SGD – Hardsync (most popular), K-AVG (with lots of nice properties)

K-step Averaging

Algorithm 1 KAVG

---

initialize $\widetilde{\mathbf{w}}_1, \mathbf{v} \leftarrow 0$

for $j = 1, \ldots, P$ in parallel **do**

  Learner P_j set $\mathbf{w}_1^j = \widetilde{\mathbf{w}}_1$

  for $n = 1, \ldots, N$ **do**

    for $k = 1, \ldots, K$ **do**

      randomly sample a mini-batch of size B_n and update:

$$\mathbf{w}_{n+k}^j \leftarrow \mathbf{w}_{n+k-1}^j - \frac{\eta_n}{B_n} \sum_{s=1}^{B_n} \nabla F(\mathbf{w}_{n+k-1}^j; \xi_{k,s}^j)$$

    end for

  end for

end for

Scaling issues evident from convergence bounds

Suppose KVAG is run for N steps, then the expected average squared gradient norms of F satisfy the following bounds for all $N \in \mathbb{N}$:

$$\frac{1}{N}\mathbb{E}\sum_{j=1}^{N} \|\nabla F(\widetilde{\mathbf{w}}_j)\|_2^2$$
$$\leq \left[\frac{2(F(\widetilde{\mathbf{w}}_1) - F^*)}{N(K-1+\delta)\bar{\eta}} + \frac{LK\bar{\eta}M}{\bar{B}(K-1+\delta)}\left(\frac{K}{P} + \frac{L(2K-1)(K-1)\bar{\eta}}{6}\right)\right],$$

With a constant number of S samples

$$\frac{1}{N}\mathbb{E}\sum_{j=1}^{N} \|\nabla F(\widetilde{\mathbf{w}}_j)\|_2^2$$
$$\leq \left[\frac{2(F(\widetilde{\mathbf{w}}_1) - F^*)PK}{S(K-1+\delta)\bar{\eta}} + \frac{LK\bar{\eta}M}{\bar{B}(K-1+\delta)}\left(\frac{K}{P} + \frac{L(2K-1)(K-1)\bar{\eta}}{6}\right)\right],$$

K-step Averaging with Momentum

Algorithm 2 MAVG

initialize $\widetilde{\mathbf{w}}_1$, $\mathbf{v} \leftarrow 0$
for $j = 1, \ldots, P$ in parallel **do**
  Learner P_j set $\mathbf{w}_1^j = \widetilde{\mathbf{w}}_1$
  for $n = 1, ..., N$ **do**
    for $k = 1, ..., K$ **do**
      randomly sample a mini-batch of size B_n and update:

$$\mathbf{w}_{n+k}^j \leftarrow \mathbf{w}_{n+k-1}^j - \frac{\eta_n}{B_n} \sum_{s=1}^{B_n} \nabla F(\mathbf{w}_{n+k-1}^j; \xi_{k,s}^j)$$

    end for
    $\mathbf{a} \leftarrow \frac{1}{P} \sum_{j=1}^{P} \mathbf{w}_{n+K}^j$;
    $\mathbf{d} \leftarrow \mathbf{a} - \widetilde{\mathbf{w}}_n$; $\mathbf{v} \leftarrow \mu \mathbf{v} + d$;
    $\widetilde{\mathbf{w}}_{n+1} = \widetilde{\mathbf{w}}_n + \mathbf{v}$;
  end for
end for

MAVG convergence bound

suppose MAVGis run with fixed step size $\eta > 0$, batch size $B > 0$ and momentum parameter $\mu \in [0, 1)$ such that the following condition holds.

$$1 \geq \frac{L^2\eta^2(K+1)(K-2)}{2(1-\mu)^2} + \frac{2\eta LK}{1-\mu}$$

and

$$1 - \delta \geq L^2\eta^2/(1-\mu)^2,$$

for some constant $\delta \in (0, 1)$. Then the expected average squared gradient norms of F satisfy the following bounds for all $N \in \mathbb{N}$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\nabla F(\widetilde{\mathbf{w}}_i)\|_2^2 \leq\ & \frac{2(1-\mu)(F(w_1)-F^*)}{N(K-1+\delta)\eta} \\ &+ \frac{L^2\eta^2\sigma^2(2K-1)K(K-1)}{6(K-1+\delta)B(1-\mu)^2} \\ &+ \frac{2LK^2\sigma^2\eta}{PB(K-1+\delta)(1-\mu)}\left(1+\frac{\mu^2}{2(1-\mu)^2}\right) \\ &+ \frac{L\eta\mu^2K^2M}{(K-1+\delta)(1-\mu)^3}. \end{aligned} \tag{1}$$

Notice how the first term is scaled by $(1-\mu)$.

MAVG – optimal $\mu > 0$

Suppose MAVG is run with fixed step size $\eta > 0$, batch size $B > 0$, number of learners $P > 0$. For N meta iterations, such that

$$1 > \frac{L^2\eta^2(K+1)(K-2)}{2} + 2\eta LK$$

and

$$1 - \delta > L^2\eta^2,$$

for some constant $\delta \in (0, 1)$. When the following conditions hold,

$$\eta^2 < \frac{B(F(w_1) - F^*)}{5LN\sigma^2(5/P + 6L)} \text{ and } K \leq 5$$

or

$$1 > \frac{N\sigma^2}{2B(F(w_1) - F^*)}\left(\frac{1}{2LP} + \frac{1}{L}\right) \text{ and } K > 5,$$

we have

$$\mu_{\text{optimal}} > 0.$$

MAVG with regard to scaling P

Let $S = N * P * B * K$, be a constant. Suppose the Algorithm 1 is run with a fixed step size η, a fixed batch size B, and a fixed frequency K. Suppose for $P = P_0$, the optimal momentum parameter is μ_0^*. If the number of processors is increased from P_0 to λP_0, where $\lambda > 1$, the momentum parameter μ_λ^* satisfies

$$\mu_\lambda^* > \mu_0^*.$$

MAVG with regard to optimal communication frequency

Suppose

$$\frac{1-\delta}{\delta}\frac{(F(w_1)-F^*)}{S\eta} > \frac{1}{(1-\mu)^3}\frac{L^2\eta^2\sigma^2}{2B}$$
$$+\frac{1}{(1-\mu)^2}\frac{3\delta-1}{2\delta}\left(\frac{\mu^2}{(1-\mu)^2}(\frac{L\sigma^2\eta}{PB}+L\eta M)+\frac{2L\sigma^2\eta}{PB}\right),$$

we have

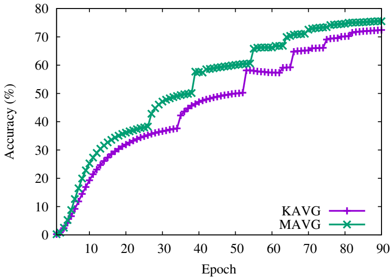
$$K_{opt}(\mu) \leq K_{opt}(0).$$

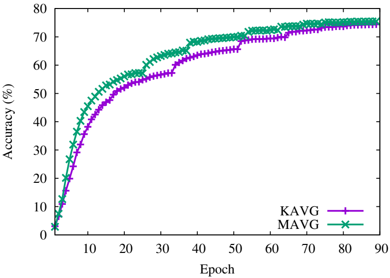
Experimental Results with 7 networks after 200 epochs, P=6

Model	KAVG	MAVG
ResNet-18	94.81	95.31
DenseNet	95.2	95.5
SENet	94.73	94.91
GoogLeNet	94.36	95.00
MoibleNet	91.77	92.16
PreActResNet-18	94.54	95.03
DPN	95.69	95.75

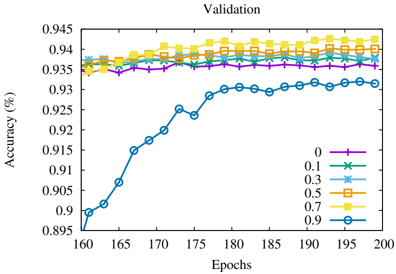
Table: Validation accuracy

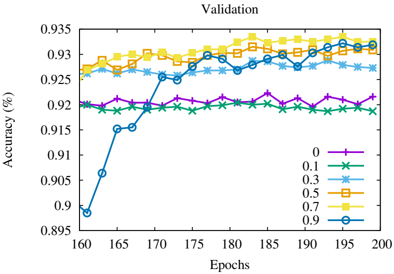
ResNet50 – Training and Validation, P=48, μ=0.6



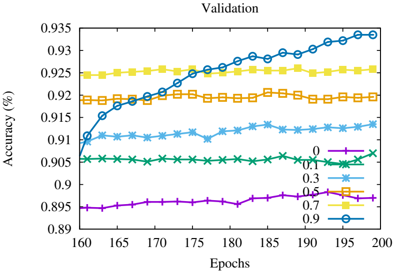


Optimal μ, P=6, 12, ResNet18





Optimal μ, P=24, ResNet18



Conclusions

We show that momentum in MAVG accelerates convergence. MAVG keeps the desirable property of KAVG that the optimal K is larger than 1 that implies low communication cost. In terms of scaling, when P increases, a larger momentum term should be used. When we switch from KAVG to MAVG, we need to use a smaller K.