Supplementary Material to “ Gradient Centralization: A New Optimization Technique for Deep Neural Networks”

Hongwei Yong[1,2], Jianqiang Huang[2], Xiansheng Hua[2], and Lei Zhang[1,2⋆]

[1] Department of Computing, The Hong Kong Polytechnic University
{cshyong,cslzhang}@comp.polyu.edu.hk
[2] DAMO Academy, Alibaba Group
{jianqiang.jqh,huaxiansheng}@gmail.com

In this supplementary file, we first provide proofs of the theoretical results in the main paper, including Corollary 4.1 and Proposition 4.2, and then present some additional experiments.

A1. Proof of Corollary 4.1

Corollary 4.1: *Suppose that SGD (or SGDM) with GC is used to update the weight vector* $\mathbf{w}$*, for any input feature vectors* $\mathbf{x}$ *and* $\mathbf{x}+\gamma\mathbf{1}$*, we have*

$$(\mathbf{w}^t)^T\mathbf{x} - (\mathbf{w}^t)^T(\mathbf{x}+\gamma\mathbf{1}) = \gamma\mathbf{1}^T\mathbf{w}^0 \tag{1}$$

where $\mathbf{w}^0$ *is the initial weight vector and* γ *is a scalar.*

Proof. First we show below a simple property of $\mathbf{P}$:

$$\mathbf{1}^T\mathbf{P} = \mathbf{1}^T(\mathbf{I}-\mathbf{e}\mathbf{e}^T) = \mathbf{1}^T - \frac{1}{M}\mathbf{1}^T\mathbf{1}\mathbf{1}^T = \mathbf{0}^T,$$

where M is the dimension of $\mathbf{e}$.

For each SGD step with GC, we have:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha^t\mathbf{P}\nabla_{\mathbf{w}^t}\mathcal{L}.$$

It can be easily derived that:

$$\mathbf{w}^t = \mathbf{w}^0 - \mathbf{P}\sum_{i=0}^{t-1}\alpha^{(i)}\nabla_{\mathbf{w}^{(i)}}\mathcal{L},$$

⋆ This research is supported by the Hong Kong RGC GRF grant (PolyU 152216/18E).

where t is the number of iterations. Then for the output activations of $\mathbf{x}$ and $\mathbf{x}+\gamma\mathbf{1}$, there is

$$\begin{aligned}
(\mathbf{w}^t)^T\mathbf{x}-(\mathbf{w}^t)^T(\mathbf{x}+\gamma\mathbf{1}) &= \gamma\mathbf{1}^T\mathbf{w}^t \\
&= \gamma\mathbf{1}^T(\mathbf{w}^0-\mathbf{P}\sum_{i=0}^{t-1}\alpha^{(i)}\nabla_{\mathbf{w}^{(i)}}\mathcal{L}) \\
&= \gamma\mathbf{1}^T\mathbf{w}^0-\gamma\mathbf{1}^T\mathbf{P}\sum_{i=0}^{t-1}\alpha^{(i)}\nabla_{\mathbf{w}^{(i)}}\mathcal{L} \\
&= \gamma\mathbf{1}^T\mathbf{w}^0.
\end{aligned} \tag{2}$$

Therefore,

$$(\mathbf{w}^t)^T\mathbf{x}-(\mathbf{w}^t)^T(\mathbf{x}+\gamma\mathbf{1}) = \gamma\mathbf{1}^T\mathbf{w}^0. \tag{3}$$

For SGD with momentum, the conclusion is the same, because we can obtain a term $\gamma\mathbf{1}^T\mathbf{P}\sum_{i=0}^{t-1}\alpha^{(i)}\mathbf{m}^i$ in the third row of Eq.(2), where $\mathbf{m}^i$ is the momentum in the ith iteration, and this term is also equal to zero.

The proof is completed. ■

A2. Proof of Proposition 4.2

Proposition 4.2: *Suppose $\nabla_{\mathbf{w}}\mathcal{L}$ is the gradient of loss function $\mathcal{L}$ w.r.t. weight vector $\mathbf{w}$. With the $\Phi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})$ defined in Eq.(2), we have the following conclusion for the loss function and its gradient, respectively:*

$$\begin{cases} ||\Phi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})||_2 \le ||\nabla_{\mathbf{w}}\mathcal{L}||_2, \\ ||\nabla_{\mathbf{w}}\Phi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})||_2 \le ||\nabla^2_{\mathbf{w}}\mathcal{L}||_2. \end{cases} \tag{4}$$

Proof. Because $\mathbf{e}$ is a unit vector, there is $\mathbf{e}^T\mathbf{e}=1$. We can easily prove that:

$$\begin{aligned}
\mathbf{P}^T\mathbf{P} &= (\mathbf{I}-\mathbf{e}\mathbf{e}^T)^T(\mathbf{I}-\mathbf{e}\mathbf{e}^T) \\
&= \mathbf{I}-2\mathbf{e}\mathbf{e}^T+\mathbf{e}\mathbf{e}^T\mathbf{e}\mathbf{e}^T \\
&= \mathbf{I}-\mathbf{e}\mathbf{e}^T \\
&= \mathbf{P}.
\end{aligned} \tag{5}$$

Table 1. Testing accuracies of different weight decay on CIFAR100 with ResNet50.

Weight decay	0	$1e^{-4}$	$2e^{-4}$	$5e^{-4}$	$1e^{-3}$
w/o GC	71.62±0.31	73.91±0.35	75.57±0.33	78.23±0.42	77.43±0.30
w/ GC	**72.83±0.29**	**76.56±0.31**	**77.62±0.37**	**79.14±0.33**	**78.10±0.36**

Table 2. Testing accuracies of different learning rates on CIFAR100 with ResNet50 for SGDM and Adam.

Algorithm	SGDM	SGDM	SGDM	Adam	Adam	Adam
Learning rate	0.05	0.1	0.2	0.0005	0.001	0.0015
w/o GC	76.81±0.27	78.23±0.42	76.53±0.32	73.88±0.46	71.64±0.56	**70.63±0.44**
w/ GC	**78.12±0.33**	**79.14±0.33**	**77.71±0.35**	**74.32±0.55**	**72.80±0.62**	**71.22±0.49**

Then for $\varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})$, we have:

$$\begin{aligned}
||\varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})||_2^2 &= \varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})^T\varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L}) \\
&= (\mathbf{P}\nabla_{\mathbf{w}}\mathcal{L})^T(\mathbf{P}\nabla_{\mathbf{w}}\mathcal{L}) \\
&= \nabla_{\mathbf{w}}\mathcal{L}^T\mathbf{P}^T\mathbf{P}\nabla_{\mathbf{w}}\mathcal{L} \\
&= \nabla_{\mathbf{w}}\mathcal{L}^T\mathbf{P}\nabla_{\mathbf{w}}\mathcal{L} \\
&= \nabla_{\mathbf{w}}\mathcal{L}^T(\mathbf{I}-\mathbf{e}\mathbf{e}^T)\nabla_{\mathbf{w}}\mathcal{L} \\
&= \nabla_{\mathbf{w}}\mathcal{L}^T\nabla_{\mathbf{w}}\mathcal{L} - \nabla_{\mathbf{w}}\mathcal{L}^T\mathbf{e}\mathbf{e}^T\nabla_{\mathbf{w}}\mathcal{L} \\
&= ||\nabla_{\mathbf{w}}\mathcal{L}||_2^2 - ||\mathbf{e}^T\nabla_{\mathbf{w}}\mathcal{L}||_2^2 \\
&\leq ||\nabla_{\mathbf{w}}\mathcal{L}||_2^2.
\end{aligned} \tag{6}$$

For $\nabla_{\mathbf{w}}\varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})$, we also have

$$\begin{aligned}
||\nabla_{\mathbf{w}}\varPhi_{GC}(\nabla_{\mathbf{w}}\mathcal{L})||_2^2 &= ||\mathbf{P}\nabla_{\mathbf{w}}^2\mathcal{L}||_2^2 \\
&= \nabla_{\mathbf{w}}^2\mathcal{L}^T\mathbf{P}^T\mathbf{P}\nabla_{\mathbf{w}}^2\mathcal{L} \\
&= \nabla_{\mathbf{w}}^2\mathcal{L}^T\mathbf{P}\nabla_{\mathbf{w}}^2\mathcal{L} \\
&= ||\nabla_{\mathbf{w}}^2\mathcal{L}||_2^2 - ||\mathbf{e}^T\nabla_{\mathbf{w}}^2\mathcal{L}||_2^2 \\
&\leq ||\nabla_{\mathbf{w}}^2\mathcal{L}||_2^2.
\end{aligned} \tag{7}$$

The proof is completed. ■

A3. More Experiments

Different hyper-parameter settings: In order to illustrate that GC can achieve consistent improvement with different hyper-parameters, we present the results of GC with different settings of weight decay and learning rates on the CIFAR100 dataset. ResNet50 is used as the backbone.

Table 1 shows the testing accuracies with different settings of weight decay, including 0, $1e^{-4}$, $2e^{-4}$, $5e^{-4}$ and $1e^{-3}$. The optimizer is SGDM with learning rate 0.1. Other settings are the same as those in the manuscript. It can be seen that the performance of weight decay is consistently improved by GC.

Table 2 shows the testing accuracies with different learning rates for SGDM and Adam. For SGDM, the learning rates are 0.05, 0.1 and 0.2, and for Adam, the learning rates are 0.0005, 0.001 and 0.0015. The weight decay is set to $5e^{-4}$. Other settings are the same as those in the manuscript. We can see that GC consistently improves the performance.