Accelerating Training of Batch Normalization: A Manifold Perspective

Abstract

Batch normalization (BN) has become a critical component across diverse deep neural networks. The network with BN is invariant to positively linear re-scaling of weights, which makes there exist infinite functionally equivalent networks with different scales of weights. However, optimizing these equivalent networks with the first-order method such as stochastic gradient descent will converge to different local optima owing to different gradients across training. To alleviate this, we propose a quotient manifold *PSI manifold*, in which all the equivalent weights of the network with BN are regarded as the same one element. We then construct gradient descent and stochastic gradient descent on the proposed PSI manifold. The two algorithms guarantee that every group of equivalent weights (caused by positively re-scaling) converge to the equivalent optima. Besides that, we characterize the convergence rate of the proposed algorithms on PSI manifold and justify that they accelerate training compared with the algorithms on the Euclidean weight space. Empirical results show that our algorithms can consistently achieve better performances in the sense of both convergence rate and generalization ability over various experimental settings.

1 INTRODUCTION

Batch normalization (BN) [Ioffe and Szegedy, 2015] is one of the most critical innovations in deep learning, which appears to help optimization as well as generalization [Ioffe and Szegedy, 2015, Santurkar et al., 2018]. Despite the success of BN, there is one adverse effect in terms of optimization due to the positively scale-invariant (PSI) property brought by it. The PSI property is explained as the weights in each layer with BN are invariant to positively linear re-scaling. Due to this, there can be an infinite number of networks functionally equivalent to each other but with various scales of weights. These networks can converge to different local optima owing to different gradients [Cho and Lee, 2017, Huang et al., 2017]. Then the converged point may become sensitive to the scale of parameters. Hence, it is desired to obviate this ambiguity of training.

To this end, we leverage the technique of optimization on manifold [Absil et al., 2009, Cho and Lee, 2017, Huang et al., 2017, Badrinarayanan et al., 2015]. We propose to constrain the scale-invariant weights of deep neural networks with BN in a quotient manifold i.e. *PSI manifold* defined in Section 4, in which all the positively scale-equivalent weights are viewed as the same element. Constraining the scale-invariant weights on the PSI manifold maintains the representation ability of hypothesis space due to the PSI property of network with BN. More importantly, optimizing on the manifold avoids the training ambiguity caused by the PSI property.

By constructing the Riemannian metric and retraction function on the PSI manifold [Boumal et al., 2019], we propose the gradient descent (GD), stochastic gradient descent (SGD), and SGD with momentum on the manifold. We abbreviate the corresponding algorithms as PSI-GD, PSI-SGD, and PSI-SGDM, respectively. Compared with vanilla GD and SGD on the Euclidean weight space, the proposed algorithms guarantee the equivalent scale-invariant weights (caused by positively re-scaling) converge to the equivalent local optima.

In contrast to the literature of optimizing network with BN on manifold [Cho and Lee, 2017, Huang et al., 2017, Badrinarayanan et al., 2015], we give the exact convergence rates of PSI-GD and PSI-SGD. They are respectively converge in the order of $O(1/T)$ and $O(1/\sqrt{T})$ under the non-convex and smooth assumptions, where T is the number of iterations. The order of convergence rates match the optimal results with fine-tuned learning rates [Ghadimi and Lan,

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

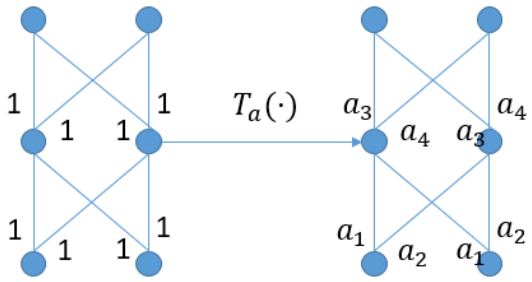

Figure 1: Batch normalized neural network with $\boldsymbol{W} = (1, \cdots, 1)^T$. For $\boldsymbol{a} = (a_1, a_2, a_3, a_4) > 0$, $f(\boldsymbol{x}, \boldsymbol{W}) = f(\boldsymbol{x}, T_{\boldsymbol{a}}(\boldsymbol{W}))$.

2013] in non-convex optimization. However, the proposed algorithms are actually training with adaptive learning rate, which is adjusted according to the smoothness of loss function. The adaptive learning rate allows us to show the proposed algorithms have better constant in the convergence rates compared with their vanilla versions on the Euclidean space.

To empirically study the proposed algorithms, we compare PSI-SGD and PSI-SGDM with the other three baseline algorithms. They are respectively vanilla SGD, adaptive learning rate algorithm Adam [Diederik P. Kingma, 2015], and another manifold based algorithm SGDG [Cho and Lee, 2017]. The experiments are conducted on image classification task on datasets `CIFAR` and `ImageNet` [Krizhevsky and Hinton, 2009, Deng et al., 2009]. We empirical observe the proposed methods consistently improve the baseline algorithms in the sense of both convergence rate and generalization over various experimental settings. The observed accelerated convergence rate and better generalization ability of PSI-GD and PSI-SGD verify our theoretical justification that the two method are able to find local minima with less iterations.

2 RELATED WORKS

Optimization on Manifold. Absil et al. [2009] presents an introduction and summarization to this topic. The foundations and some recent theoretical results in this topic refers to [Absil and Gallivan, 2006, Liu et al., 2017, Zhang and Sra, 2016, Boumal et al., 2019]. Roughly speaking, optimization on manifold converts the constrained problem into an unconstrained problem on a specific manifold. Huang et al. [2018], Lezcano-Casado and Martínez-Rubio [2019], Casado [2019], Li et al. [2019] leverage the technique to handle some explicit constraints to deep neural networks e.g., orthogonal constraints. [Cho and Lee, 2017, Huang et al., 2017] fix the ambiguity of optimizing network with BN by manifold based algorithms (Grassmann and Oblique manifold respectively). However, in contrast to this paper, the theoretical results of their method are not explored. For example, the convergence rate of their methods.

PSI property. Our method in this paper is build upon PSI property brought by BN. The PSI property properly defined in Section 4 are widely appears in modern deep neural networks. For example, the network with ReLU activation [Neyshabur et al., 2015] and the network with normalization layer, e.g., batch normalization [Ioffe and Szegedy, 2015], layer normalization [Ba et al., 2016] and group normalization [Wu and He, 2018]. Arora et al. [2018], Wu et al. [2018] use a similar theoretical framework as ours to show the network with PSI property allows a more flexible choice of the learning rate, and thus easier to be trained. However, we design algorithm to avoid the optimization brought by PSI property, and show the proposed methods can have improved convergence rate.

Due to the PSI property, Li and Arora [2019] uses exponentially increased learning rate for the network with BN to handle implicit effect brought by weight decay. In contrast to fixing the ambiguity of optimization by manifold based methods, Neyshabur et al. [2015], Meng et al. [2018], Zheng et al. [2019] re-parameterize the network by "path value" and optimizing on the path value space.

3 BACKGROUND

3.1 PROBLEM SET UP

In this paper, we consider the deep neural network with the following structure

$$f(\boldsymbol{x}, \boldsymbol{W}) = BN(\boldsymbol{W}_L \phi(BN(\boldsymbol{W}_{L-1}\phi(\cdots \phi(BN(\boldsymbol{W}_1 \boldsymbol{x})))))), \tag{1}$$

where $\phi(\cdot)$ is the non-linear activation, and $\boldsymbol{W}_l$ is the weight matrix of the l-th layer. $BN(\cdot)$ operator in the equation represents batch normalization [Ioffe and Szegedy, 2015] layer, which normalizes every hidden output $z = \boldsymbol{\theta}^T \boldsymbol{x}$

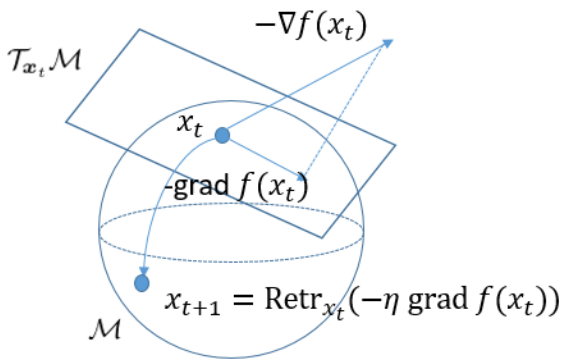

Figure 2: Gradient descent on Riemannian manifold.

across a batch of training samples as

$$BN(\boldsymbol{z}) = \gamma\frac{\boldsymbol{z}-\mathbb{E}(\boldsymbol{z})}{\sqrt{\mathrm{Var}(\boldsymbol{z})}} + \beta = \gamma\frac{\boldsymbol{\theta}^T(\boldsymbol{x}-\mathbb{E}(\boldsymbol{x}))}{\sqrt{\boldsymbol{\theta}^T\mathrm{Var}_{\boldsymbol{x}}\boldsymbol{\theta}}} + \beta, \quad (2)$$

where γ and β are learnable affine parameters. The loss is

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}_i, \boldsymbol{W}), y_i), \quad (3)$$

across this paper, where $\{(\boldsymbol{x}_i, y_i)\}$ is the training set, $\ell(\boldsymbol{x}, y)$ represents the loss function, and $\boldsymbol{g}$ are scale-variant parameters i.e. γ and β in BN layer. Let $\boldsymbol{W} = (\boldsymbol{w}^{(1)}\cdots, \boldsymbol{w}^{(m)})$, the $\boldsymbol{w}^{(i)}$ is the weight vector of the i-th neuron. Then $\mathcal{L}(\boldsymbol{W}, \boldsymbol{g})$ is positively scale-invariant w.r.t. $\boldsymbol{W}$. It means $\mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) = \mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}), \boldsymbol{g})$ for any $T_{\boldsymbol{a}}(\boldsymbol{W}) = (a_1\boldsymbol{w}^{(1)}, \cdots, a_m\boldsymbol{w}^{(m)})$ with $\{a_i > 0\}$. An illustration for the PSI property of the network with BN refers to Figure 1. Specially, let $\boldsymbol{V}$ be the normalized $\boldsymbol{W}$ which means $\boldsymbol{V} = (\boldsymbol{w}^{(1)}/\|\boldsymbol{w}^{(1)}\|, \cdots, \boldsymbol{w}^{(m)}/\|\boldsymbol{w}^{(m)}\|)$. Then we have $\mathcal{L}(\boldsymbol{V}, \boldsymbol{g}) = \mathcal{L}(\boldsymbol{W}, \boldsymbol{g})$.

It worth noting that our discussions below can be applied on the top of any functions with PSI property e.g., ReLU neural networks [Meng et al., 2018, Neyshabur et al., 2015]. However, we use NN with BN as a representation to illustrate our conclusions.

3.2 OPTIMIZATION ON MANIFOLD

We briefly introduce some definitions of optimization on the manifold, more details refers to Appendix A or [Absil et al., 2009]. A Matrix Manifold $\mathcal{M}$ is a subset of Euclidean space that for any $\boldsymbol{x} \in \mathcal{M}$, there exists a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$, such that $U_{\boldsymbol{x}}$ is homeomorphic to a Euclidean space. The PSI parameters $\boldsymbol{W}$ can be defined on a matrix manifold, e.g., Grassmann manifold or Oblique manifold without losing model capacity [Cho and Lee, 2017, Huang et al., 2017]. We propose a PSI manifold for them in this paper. For any $\boldsymbol{x} \in \mathcal{M}$, there is a tangent space $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$. For any $f(\boldsymbol{x})$ defined on $\mathcal{M}$, there is a Riemannian gradient $\mathrm{grad}f(\boldsymbol{x}) \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$. After defining a retraction function $R_{\boldsymbol{x}}(\cdot) : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \to \mathcal{M}$, it was shown in [Boumal et al., 2019] that gradient descent on manifold (refers to Figure 2)

$$\boldsymbol{x}_{t+1} = R_{\boldsymbol{x}_t}(-\eta\mathrm{grad}f(\boldsymbol{x}_t)) \quad (4)$$

has the same convergence rate with vanilla gradient descent in Euclidean space under various mild conditions.

4 PSI MANIFOLD

The optimization ambiguity of PSI parameters states as follows. By chain rule,

$$\nabla_{\boldsymbol{w}^{(i)}}\mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) = a_i\nabla_{a_i\boldsymbol{w}^{(i)}}\mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}), \boldsymbol{g}), \quad (5)$$

so the two positively scale-equivalent points $\boldsymbol{W}$ and $T_{\boldsymbol{a}}(\boldsymbol{W})$ with the same output converge to different local optima owing to the different gradients. In other words, gradient-based algorithms in Euclidean space can not make two iterates started from $\boldsymbol{W}$ and $T_{\boldsymbol{a}}(\boldsymbol{W})$ positively scale equivalent with each other across the training. This makes the converged point becomes sensitive the scale of weights. To alleviate this, we construct a quotient manifold i.e., PSI manifold. Constraining parameters in it avoids the training ambiguity brought by the positively scale-invariant property.

4.1 CONSTRUCTION OF PSI MANIFOLD

A quotient manifold is induced by an equivalent relationship. We formally formulate the positively scale-equivalence to derive the PSI manifold with PSI parameters defined in it.

Definition 1 (Positively Scale-Equivalent)**.** *For a pair of $\boldsymbol{W}, \boldsymbol{W}'$, they are positively scale-equivalent with each other, if there exists a transformation $T_{\boldsymbol{a}}(\cdot)$ such that $\boldsymbol{W}' = T_{\boldsymbol{a}}(\boldsymbol{W})$. $\boldsymbol{W} \sim \boldsymbol{W}'$ denotes the equivalence.*

We can verify that the positively scale-equivalent $\sim$ is an equivalent relationship. With the following proposition, we construct the PSI manifold.

Proposition 1. *For PSI parameters $\boldsymbol{W}$, the positively scale-equivalent relationship $\sim$ induces a quotient manifold $\mathcal{M} = \overline{\mathcal{M}}/\sim = \mathbb{R}^{d_1} \times \cdots \mathbb{R}^{d_m}/\sim$ named PSI manifold. Here d_i is the dimension of $\boldsymbol{w}^{(i)}$.*

Following the proof of Proposition 3.4.6 in [Absil et al., 2009], we get the conclusion.

Up till now, we formally defined the PSI manifold. Intuitively, all the parameters that are positively scale-equivalent with each other in the Euclidean space are viewed as the same element in the PSI manifold $\mathcal{M}$. Then, we can define the PSI parameters directly in the manifold without losing the representation ability of the model.

The PSI manifold is proposed to alleviate the ambiguity of optimization brought by the PSI property. To this end, we can optimize the PSI parameters directly in the PSI manifold, since all the positively scale-equivalent parameters are the same point in it. But (4) implies that optimization on PSI manifold requires the formulation of Riemannian gradient and retraction function on it. The following proposition defines a Riemannian metric in the PSI manifold, which formulates the Riemannian gradient (See this Definition in Appendix A).

Proposition 2. *The Riemannian metric $\langle \cdot, \cdot \rangle_{\boldsymbol{W}}$ in the PSI manifold can be defined as*

$$\langle \boldsymbol{\Xi}_1, \boldsymbol{\Xi}_2 \rangle_{\boldsymbol{W}} = \sum_{i=1}^{m} \frac{\langle \boldsymbol{\xi}_1^{(i)}, \boldsymbol{\xi}_2^{(i)} \rangle}{\|\boldsymbol{w}^{(i)}\|^2}, \tag{6}$$

for every $\boldsymbol{W} \in \mathcal{M}$ [1] and $\boldsymbol{\Xi}_1, \boldsymbol{\Xi}_2 \in \mathcal{T}_{\boldsymbol{W}}\mathcal{M}$, where $\boldsymbol{\Xi}_k = (\boldsymbol{\xi}_k^{(1)}, \cdots \boldsymbol{\xi}_k^{(m)})$.

The proof of this Proposition is in Appendix B.1. With the Riemannian metric, we can exactly compute the Riemannian gradient on the PSI manifold. According to Definition of Riemannian gradient in Appendix A, we have

$$\mathrm{grad}_{\boldsymbol{w}^{(i)}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) = \|\boldsymbol{w}^{(i)}\|^2 \nabla_{\boldsymbol{w}^{(i)}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) \tag{7}$$

for each $\boldsymbol{w}^{(i)}$. The retraction function $R_{\boldsymbol{W}}(\cdot)$ is also required to derive gradient-based algorithms on the PSI manifold.

Proposition 3. *A retraction function $R_{\boldsymbol{W}}(\boldsymbol{\Xi})$ on the PSI manifold can be defined as*

$$R_{\boldsymbol{W}}(\boldsymbol{\Xi}) = \boldsymbol{W} + \boldsymbol{\Xi} \tag{8}$$

for every $\boldsymbol{W} \in \mathcal{M}, \boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{W}}\mathcal{M}$.

The proof of this proposition appears in Appendix C.1. Now we are ready to optimize PSI parameters in the PSI manifold directly, which avoids the ambiguity of training brought by PSI property.

[1] Please note that for any $T_{\boldsymbol{a}}(\boldsymbol{W})$ and $\boldsymbol{W}$ are view as the same element in the PSI manifold $\mathcal{M}$. Here $\boldsymbol{W}$ is a representation from the elements $\{T_{\boldsymbol{a}}(\boldsymbol{W}) : a_i > 0\}$.

4.2 OPTIMIZATION ON THE PSI MANIFOLD

In this subsection, we give the update rules of GD, SGD, and SGD with momentum on the PSI manifold, abbreviated as PSI-GD, PSI-SGD, and PSI-SGDM respectively. We show that all of these update rules on the PSI manifold generate a unified optimization path for parameters that positively scale-equivalent with each other.

Combining equation (4) and Proposition 2, 3, we have the following update rule of GD

$$\begin{aligned} \boldsymbol{w}_{t+1}^{(i)} &= R_{\boldsymbol{w}_t^{(i)}} \left(-\eta_{\boldsymbol{w}_t^{(i)}} \mathrm{grad}_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right) \\ &= \boldsymbol{w}_t^{(i)} - \eta_{\boldsymbol{w}_t^{(i)}} \|\boldsymbol{w}_t^{(i)}\|^2 \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t), \end{aligned} \tag{9}$$

SGD

$$\begin{aligned} \boldsymbol{w}_{t+1}^{(i)} &= R_{\boldsymbol{w}_t^{(i)}} \left(-\eta_{\boldsymbol{w}_t^{(i)}} \mathrm{grad}_{\boldsymbol{w}_t^{(i)}} \hat{\mathcal{L}}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right) \\ &= \boldsymbol{w}_t^{(i)} - \eta_{\boldsymbol{w}_t^{(i)}} \|\boldsymbol{w}_t^{(i)}\|^2 \mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t), \end{aligned} \tag{10}$$

and SGD with momentum

$$\begin{aligned} \boldsymbol{u}_{t+1}^{(i)} &= R_{\rho \boldsymbol{u}_t^{(i)}} \left(-\eta_{\boldsymbol{w}_t^{(i)}} \mathrm{grad}_{\boldsymbol{w}_t^{(i)}} \hat{\mathcal{L}}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right) \\ &= \rho \boldsymbol{u}_t^{(i)} - \eta_{\boldsymbol{w}_t^{(i)}} \|\boldsymbol{w}_t^{(i)}\|^2 \mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t); \\ \boldsymbol{w}_{t+1}^{(i)} &= R_{\boldsymbol{w}_t^{(i)}} \left(\boldsymbol{u}_{t+1}^{(i)} \right) = \boldsymbol{w}_{t+1}^{(i)} + \boldsymbol{u}_{t+1}^{(i)}. \end{aligned} \tag{11}$$

for PSI parameters. Here

$$\begin{aligned} \mathrm{grad}_{\boldsymbol{w}_t^{(i)}} \hat{\mathcal{L}}(\boldsymbol{W}_t, \boldsymbol{g}_t) &= \|\boldsymbol{w}_t^{(i)}\|^2 \mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t) \\ &= \|\boldsymbol{w}_t^{(i)}\|^2 \nabla_{\boldsymbol{w}_t^{(i)}} \frac{1}{S} \sum_{k=1}^{S} \nabla_{\boldsymbol{w}_t^{(i)}} \ell\left(f(\boldsymbol{x}_k, \boldsymbol{W}_t), y_k\right) \end{aligned} \tag{12}$$

for a batch of $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_S, y_S)\}$ is an unbiased estimation to the Riemannian gradient. In addition, the update rules of non-scale-invariant parameters $\boldsymbol{g}$ follows the gradient-based algorithms in Euclidean space.

One can justify that the proposed algorithms on PSI manifold are essentially using adaptive learning rates decided by $\|\boldsymbol{w}_t^{(i)}\|^2$ to match the gradient scale brought by a variant of weight scale. The following proposition shows the well-posedness of the algorithms on PSI manifold.

Theorem 1. *For two positively scale-equivalent weights $\boldsymbol{W}_0$ and $T_{\boldsymbol{a}}(\boldsymbol{W}_0)$, let $\boldsymbol{W}_t$, $\hat{\boldsymbol{W}}_t$ be t-th iterate of PSI-SGDM respectively started from $\{\boldsymbol{W}_0, \boldsymbol{U}_0\}$ and $\{T_{\boldsymbol{a}}(\boldsymbol{W}_0), T_{\boldsymbol{a}}(\boldsymbol{U}_0)\}$. Then we have $\hat{\boldsymbol{W}}_t = T_{\boldsymbol{a}}(\boldsymbol{W}_t)$.*

This theorem can be easily generalized to PSI-GD and PSI-SGD; the proof of it is in Appendix C.1. The conclusion shows that the iterates generated by the update rules on the PSI manifold are equivalent with respect to applying $T_{\boldsymbol{a}}(\cdot)$. Hence, optimizing on the PSI manifold avoids the optimization ambiguity brought by the PSI property. The complete algorithm of PSI-SGDM refers to Algorithm 1. With $\rho = 0$, the PSI-SGDM degenerates to PSI-SGD.

Algorithm 1 SGD with momentum on the PSI manifold (PSI-SGDM).

Input: Training steps T; batch size S; momentum parameters ρ; learning rate $\eta_{\boldsymbol{w}_t^{(i)}}$ and $\eta_{\boldsymbol{g}_t}$.
for $t = 0 \cdots T-1$ **do**
  Sampling a batch of data $\{(x_1, y_1), \cdots, (\boldsymbol{x}_S.y_S)\}$ from training set
  for $i = 1, \cdots, m$ **do**
    $\boldsymbol{u}_{t+1}^{(i)} = \rho \boldsymbol{u}_t^{(i)} - \eta_{\boldsymbol{w}_t^{(i)}} \|\boldsymbol{w}_t^{(i)}\|^2 \frac{1}{S} \sum_{k=1}^{S} \nabla_{\boldsymbol{w}_t^{(i)}} \ell\left(f(\boldsymbol{x}_k, \boldsymbol{W}_t), y_k\right)$
    $\boldsymbol{w}_{t+1}^{(i)} = \boldsymbol{w}_{t+1}^{(i)} + \boldsymbol{u}_{t+1}^{(i)}$
  end for
  $\boldsymbol{g}_{t+1} = \boldsymbol{g}_t - \eta_{\boldsymbol{g}_t} \sum_{k=1}^{S} \nabla_{\boldsymbol{g}_t} \ell\left(f(\boldsymbol{x}_k, \boldsymbol{W}_t), y_k\right)$
end for
return $(\boldsymbol{W}_T, \boldsymbol{g}_T)$

Remark 1. *The update rule of SGD with momentum on manifold [Cho and Lee, 2017, Liu et al., 2017] involves the parallel transformation to make sure the $\boldsymbol{U}_{t+1}$ locates in the tangent space $\mathcal{T}_{\boldsymbol{W}_t}\mathcal{M}$. However, it requires the proposed retraction in Proposition 3 to be an exponential retraction map [Absil et al., 2009], which may fails in-practical. Hence, we heuristically propose the PSI-SGDM without verifying the retraction function. Even though, our method is well-defined to positively scale-transformation.*

5 OPTIMIZATION ON THE PSI MANIFOLD ACCELERATES TRAINING

In this section, we give the convergence rates of PSI-GD and PSI-SGD. Furthermore, we show that they accelerate training compared with vanilla GD and SGD on the Euclidean space.

5.1 CONVERGENCE RATES OF OPTIMIZATION ON THE PSI MANIFOLD

We give the convergence rates of PSI-GD and PSI-SGD (Algorithm 1) in this subsection. For the PSI parameters, let $\boldsymbol{V}_t = (\boldsymbol{w}_t^{(1)}/\|\boldsymbol{w}_t^{(1)}\|, \cdots, \boldsymbol{w}_t^{(m)}/\|\boldsymbol{w}_t^{(m)}\|)$, we involve the following assumption to characterize the smoothness of normalized parameters

$$\begin{aligned} &\left\|\nabla^2_{\boldsymbol{v}^{(i)}\boldsymbol{v}^{(j)}}\mathcal{L}(\boldsymbol{V}, \boldsymbol{g})\right\|_2 \le L_{ij}^{\boldsymbol{vv}}, \\ &\left\|\nabla^2_{\boldsymbol{v}^{(i)}\boldsymbol{g}}\mathcal{L}(\boldsymbol{V}, \boldsymbol{g})\right\|_2 \le L_{i}^{\boldsymbol{vg}}, \\ &\left\|\nabla^2_{\boldsymbol{g}}\mathcal{L}(\boldsymbol{V}, \boldsymbol{g})\right\|_2 \le L^{\boldsymbol{gg}}. \end{aligned} \tag{13}$$

Here $\|\cdot\|_2$ is the spectral norm of matrix, it represents the Lipschitz constant of gradient. For PSI-SGD, we further assume the bounded variance:

$$\begin{aligned} \mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right] \le \sigma^2; \\ \mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right] \le \sigma^2. \end{aligned} \tag{14}$$

Remark 2. *One may figure out that due to* (5)*, the upper bound* (14) *can not hold as $\|\boldsymbol{w}_i\|$ goes to zero. However, due to Lemma 1 below, the weights $\boldsymbol{w}_i$ obtained by GD or SGD will never goes to zero. Thus the bounded variance* (14) *is reasonable in this regime.*

Let $\mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*) = \inf_{\boldsymbol{W}, \boldsymbol{g}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{g})$, we have the following two Theorems to give the convergence rates of PSI-GD and PSI-SGD.

Theorem 2. *Let $\{\boldsymbol{W}_t, \boldsymbol{g}_t\}$ be the iterates of PSI-GD* (9)*, then we have*

$$\min_{0\le t<T} \|\nabla_{\boldsymbol{V}_t, \boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t, \boldsymbol{g}_t)\|^2 \le \frac{2\tilde{L}(\mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*))}{T} \tag{15}$$

by choosing $\eta_{\boldsymbol{w}_t}^{(i)} = 1/\tilde{L}_{\boldsymbol{v}^{(i)}}$ and $\eta_{\boldsymbol{g}_t} = 1/\tilde{L}_{\boldsymbol{g}}$. Here $\tilde{L} = \max\{\tilde{L}_{\boldsymbol{v}^{(1)}}, \cdots, \tilde{L}_{\boldsymbol{v}^{(m)}}, \tilde{L}_{\boldsymbol{g}}\}$, where $\tilde{L}_{\boldsymbol{v}^{(i)}} = L_i^{\boldsymbol{vg}} + \sum_{j=1}^{m} L_{ij}^{\boldsymbol{vv}}$; $\tilde{L}_{\boldsymbol{g}} = mL^{\boldsymbol{gg}} + \sum_{i=1}^{m} L_i^{\boldsymbol{vg}}$.

Theorem 3. *Let $\{\boldsymbol{W}_t, \boldsymbol{g}_t\}$ be the iterates of PSI-SGD* (10)*, then we have*

$$\begin{aligned} \min_{0\le t<T} \mathbb{E}\left[\|\nabla_{\boldsymbol{V}_t, \boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t, \boldsymbol{g}_t)\|^2\right] \le \frac{2\tilde{L}(\mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*))}{\sqrt{T}} \\ + \frac{\sigma^2\tilde{L}}{2\sqrt{T}} \sum_{i=1}^{m} \left(\frac{1}{\tilde{L}_{\boldsymbol{g}}} + \frac{1}{\tilde{L}_{\boldsymbol{v}^{(i)}}}\right), \end{aligned} \tag{16}$$

by choosing $\eta_{\boldsymbol{w}_t^{(i)}} = \frac{1}{\tilde{L}_{\boldsymbol{v}^{(i)}}\sqrt{T}}$ and $\eta_{\boldsymbol{g}_t} = \frac{1}{\tilde{L}_{\boldsymbol{g}}\sqrt{T}}$, where $\tilde{L}, \tilde{L}_{\boldsymbol{v}^{(i)}}$ and $\tilde{L}_{\boldsymbol{g}}$ are defined in Theorem 2.

The proofs of the two theorems are respectively referred to Appendix D.3 and D.4. We see the convergence rates of PSI-GD and PSI-SGD are respectively $O(1/T)$ and $O(1/\sqrt{T})$.

One can note the convergence rate is about the gradient with respect to normalized parameters $\boldsymbol{V}_t$. But equation (5) implies

$$\begin{aligned} \|\nabla_{\boldsymbol{W}_t, \boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\|^2 &= \sum_{i=1}^{m} \frac{1}{\|\boldsymbol{w}_t^{(i)}\|^2} \left\|\nabla_{\boldsymbol{v}_t^{(i)}, \boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t, \boldsymbol{g}_t)\right\|^2 \\ &\quad + \|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t, \boldsymbol{g}_t)\|^2 \\ &\le \|\nabla_{\boldsymbol{V}_t, \boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t, \boldsymbol{g}_t)\|^2, \end{aligned} \tag{17}$$

if $\|\boldsymbol{w}_t^{(i)}\|^2 \ge 1$ for $1 \le i \le m$. Thus we can get the corresponded convergence rates of PSI-GD and PSI-SGD. They match the optimal results with well tuned learning rates in the Euclidean space [Ghadimi and Lan, 2013]. The following lemma shows that the $\|\boldsymbol{w}_t^{(i)}\|^2$ keeps increasing across training for gradient-based algorithms.

Lemma 1 (Lemma 2.4 in [Arora et al., 2018]). *For any PSI weight $\boldsymbol{w}^{(i)}$, $\boldsymbol{w}^{(i)}$ and $\nabla_{\boldsymbol{w}^{(i)}}\ell\left(f(\boldsymbol{x}_k,\boldsymbol{W}),y_k\right)$ are perpendicular for any $(\boldsymbol{x}_k,y_k)$. On the other hand*

$$\begin{aligned}&\left\|\boldsymbol{w}^{(i)}+\eta_{\boldsymbol{w}^{(i)}}\nabla_{\boldsymbol{w}^{(i)}}\ell\left(f(\boldsymbol{x}_k,\boldsymbol{W}),y_k\right)\right\|^2=\left\|\boldsymbol{w}^{(i)}\right\|^2\\&+\eta_{\boldsymbol{w}^{(i)}}^2\left\|\nabla_{\boldsymbol{w}^{(i)}}\ell\left(f(\boldsymbol{x}_k,\boldsymbol{W}),y_k\right)\right\|^2.\end{aligned}\tag{18}$$

Thus, $\|\boldsymbol{w}_t^{(i)}\|^2\geq 1$ holds, if $\|\boldsymbol{w}_0^{(i)}\|^2\geq 1$. Since the network is usually initialized by $\boldsymbol{w}^{(i)}\sim\mathcal{N}(0,2/\sqrt{d_i})$ [He et al., 2015], $\|\boldsymbol{w}_0^{(i)}\|^2\approx 2$ with high probability. Hence, we can conclude $\left\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\leq\left\|\nabla_{\boldsymbol{V}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|^2$. We will show the inequality and the increasing $\|\boldsymbol{w}_t^{(i)}\|^2$ is the reason for the acceleration of the algorithms on PSI manifold.

5.2 WHY OPTIMIZATION ON THE PSI MANIFOLD ACCELERATES TRAINING

In this subsection, we show PSI-GD and PSI-SGD accelerate training compared with their version on the Euclidean space.

Roughly speaking, PSI parameters move towards a smoother region (smaller gradient Lipschitz constant) across training, since the gradient Lipschitz constant is in inverse ratio to the weight scale which keeps increasing according to Lemma 1. Because a smoother region allows a larger learning rate, gradually increasing learning rate accelerates training in this regime. Fortunately, PSI-GD and PSI-SGD happen to be vanilla GD and SGD with adaptive increasing learning rates according to equation (9), (10) and (18).

To begin with, we verify the convergence rates of vanilla GD and SGD. We assume a global smoothness for $\mathcal{L}(\boldsymbol{W},\boldsymbol{g})$,

$$\begin{aligned}&\left\|\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{w}^{(j)}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})\right\|_2\leq L_{ij}^{\boldsymbol{ww}},\\&\left\|\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{g}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})\right\|_2\leq L_{i}^{\boldsymbol{wg}},\\&\left\|\nabla^2_{\boldsymbol{g}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})\right\|_2\leq L^{\boldsymbol{gg}}.\end{aligned}\tag{19}$$

The assumption is stronger than (13) since (13) only involves the normalized parameters, thus $L_{ij}^{\boldsymbol{vv}}\leq L_{ij}^{\boldsymbol{ww}};L_i^{\boldsymbol{vg}}\leq L_i^{\boldsymbol{wg}}$. We consider the update rule of GD

$$\begin{aligned}\boldsymbol{w}_{t+1}^{(i)}&=\boldsymbol{w}_t^{(i)}-\eta_{\boldsymbol{w}_t^{(i)}}\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t);\\\boldsymbol{g}_{t+1}&=\boldsymbol{g}_t-\eta_{\boldsymbol{g}_t}\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t),\end{aligned}\tag{20}$$

and SGD

$$\begin{aligned}\boldsymbol{w}_{t+1}^{(i)}&=\boldsymbol{w}_t^{(i)}-\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t);\\\boldsymbol{g}_{t+1}&=\boldsymbol{g}_t-\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t),\end{aligned}\tag{21}$$

where $\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)$ and $\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t)$ are respectively unbiased estimations of $\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)$ and $\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)$ defined in equation (12). For SGD, we assume the two estimators satisfy the bounded variance assumption (14).

The following two Theorems characterize the convergence rate of GD and SGD.

Theorem 4. *Let $\{\boldsymbol{W}_t,\boldsymbol{g}_t\}$ updated by GD (20), by choosing $\eta_{\boldsymbol{w}_t}^{(i)}=1/\tilde{L}_{\boldsymbol{w}^{(i)}}$ and $\eta_{\boldsymbol{g}_t}=1/\tilde{L}_{\boldsymbol{g}}$,*

$$\min_{0\leq t<T}\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\|^2\leq\frac{2\tilde{L}(\mathcal{L}(\boldsymbol{W}_0,\boldsymbol{g}_0)-\mathcal{L}(\boldsymbol{W}^*,\boldsymbol{g}^*))}{T}.\tag{22}$$

Here $\tilde{L}=\max\{\tilde{L}_{\boldsymbol{w}^{(1)}},\cdots,\tilde{L}_{\boldsymbol{w}^{(m)}},\tilde{L}_{\boldsymbol{g}}\}$, where $\tilde{L}_{\boldsymbol{w}^{(i)}}=L_i^{\boldsymbol{wg}}+\sum_{j=1}^m L_{ij}^{\boldsymbol{ww}};\tilde{L}_{\boldsymbol{g}}=mL^{\boldsymbol{gg}}+\sum_{i=1}^m L_i^{\boldsymbol{wg}}$.

Theorem 5. *Let $\{\boldsymbol{W}_t,\boldsymbol{g}_t\}$ updated by SGD (21), then*

$$\begin{aligned}\min_{0\leq t<T}\mathbb{E}\left[\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\|^2\right]\leq&\frac{2\tilde{L}(\mathcal{L}(\boldsymbol{W}_0,\boldsymbol{g}_0)-\mathcal{L}(\boldsymbol{W}^*,\boldsymbol{g}^*))}{\sqrt{T}}\\&+\frac{\sigma^2\tilde{L}}{2\sqrt{T}}\sum_{i=1}^m\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}}+\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right),\end{aligned}\tag{23}$$

by choosing $\eta_{\boldsymbol{w}_t^{(i)}}=\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}\sqrt{T}}$ and $\eta_{\boldsymbol{g}_t}=\frac{1}{\tilde{L}_{\boldsymbol{g}}\sqrt{T}}$, where $\tilde{L},\tilde{L}_{\boldsymbol{w}^{(i)}}$ and $\tilde{L}_{\boldsymbol{g}}$ are defined in Theorem 4.

The two convergence rates of $\|\nabla_{\boldsymbol{W},\boldsymbol{g}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})\|$ match the optimal results with well tuned learning rates [Ghadimi and Lan, 2013]. Since $L_{ij}^{\boldsymbol{vv}}\leq L_{ij}^{\boldsymbol{ww}};L_i^{\boldsymbol{vg}}\leq L_i^{\boldsymbol{wg}}$, the right hand side in (22) is smaller than the one in (15). To see the results for SGD, under (19), the upper bound in (16) can be replaced by the smaller one of the upper bounds in (16) and (23). Thus the convergence rate of PSI-SGD is also sharper than the one of SGD.

On the other hand, equation (17) and the discussion in the above section implies $\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\|^2\leq\|\nabla_{\boldsymbol{V}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|^2$. Thus, from equation (17), one can verify the proposed algorithms accelerate training in a factor $\|\boldsymbol{w}_t^{(i)}\|^2\geq 1$ compared with the versions on Euclidean space. In addition, $\|\boldsymbol{w}_t^{(i)}\|^2$ keeps increasing to a bounded constant across training, thus the acceleration is more significant after a number of iterations. A detailed discussion to $\|\boldsymbol{w}_t^{(i)}\|^2$ is in Appendix E.

The proofs of the two theorems are delegated to Appendix D.1 and D.2. In the proof, we see that scaling the learning rate with $1/\tilde{L}_{\boldsymbol{w}^{(i)}}$ and $1/\tilde{L}_{\boldsymbol{g}}$ are respectively the optimal schedule of $\eta_{\boldsymbol{w}_t^{(i)}}$ and $\eta_{\boldsymbol{g}_t}$. Since smaller $\tilde{L}_{\boldsymbol{w}^{(i)}}$ and $\tilde{L}_{\boldsymbol{g}}$ corresponds with a smoother loss landscapes, it explains that $\mathcal{L}(\boldsymbol{W},\boldsymbol{g})$ allows a larger learning rate in a smoother region to get the optimal convergence rate.

Now we are ready to illustrate the source of acceleration of PSI-GD and PSI-SGD. Due to $\mathcal{L}(\boldsymbol{W},\boldsymbol{g})=\mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}),\boldsymbol{g})$, we have

$$\begin{aligned}\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{w}^{(j)}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})&=a_ia_j\nabla^2_{a_i\boldsymbol{w}^{(i)}a_j\boldsymbol{w}^{(j)}}\mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}),\boldsymbol{g});\\\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{g}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})&=a_i\nabla^2_{a_i\boldsymbol{w}^{(i)}\boldsymbol{g}}\mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}),\boldsymbol{g}).\end{aligned}\tag{24}$$

The fact shows that the smoothness of PSI parameters increasing with its scale. Because a larger

Table 1: Performance of ResNet on various datasets. The ResNet for CIFAR and ImageNet are respectively the specific structure for the corresponded dataset, more details about the structure refers to [He et al., 2016].

Dataset	CIFAR-10				
Algorithm	SGD	Adam	SGDG	PSI-SGD	PSI-SGDM
ResNet20	91.14	89.90	91.38	92.25	**92.41**
ResNet32	91.32	90.40	92.48	**93.56**	93.30
ResNet44	91.95	91.02	92.99	**93.90**	93.39
ResNet56	92.19	91.49	93.11	**94.08**	93.90
Dataset	CIFAR-100				
ResNet20	65.53	63.35	68.19	**69.11**	68.41
ResNet32	67.41	64.43	70.13	**70.81**	70.14
ResNet44	67.72	65.03	71.32	71.94	**72.03**
ResNet56	68.02	65.77	71.46	**72.88**	72.40
Dataset	ImageNet				
ResNet18	67.72	68.16	68.97	**70.38**	69.42
ResNet34	71.30	70.68	72.40	**73.31**	72.88
ResNet50	73.46	74.00	74.67	74.67	**75.11**

$\boldsymbol{a}$ in the right hand side of the above results in a smaller $\|\nabla^2_{a_i \boldsymbol{w}^{(i)} a_j \boldsymbol{w}^{(j)}} \mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}), \boldsymbol{g})\|_2$ and $\|\nabla^2_{a_i \boldsymbol{w}^{(i)} \boldsymbol{g}} \mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}), \boldsymbol{g})\|_2$. Due to Lemma 1, the increasing $\|\boldsymbol{w}_t^{(i)}\|$ implies the loss landscape of PSI parameters becomes smoother across training. Thus, gradually increasing the learning rate to update PSI parameters can accelerate training. The proposed PSI-GD and PSI-SGD are proven to be vanilla GD and SGD with the increasing learning rate that is proportional to $\|\boldsymbol{w}_t^{(i)}\|^2$, which interprets the acceleration.

Finally, equation (24) shows that our algorithms are optimal in the manner of leveraging the smoothness to accelerate training. The intuition is that the proposed methods have a more accurate estimation of local smoothness across training. We point out that $L_{ij}^{\boldsymbol{vv}} \leq L_{ij}^{\boldsymbol{ww}}; L_i^{\boldsymbol{vg}} \leq L_i^{\boldsymbol{wg}}$ in equation (19) also gives a sharper convergence rate of PSI-GD and PSI-SGD.

6 EXPERIMENTS

6.1 IMPROVED CONVERGENCE RATE

We use a toy example to verify that our proposed algorithms indeed have improved convergence rate compared with other baselines methods.

Data. We sample 1000 training samples $\{\boldsymbol{x}_i\}$ from 10-dimensional normal distribution with its label $y_i = \mu^\top \boldsymbol{x}_i + \epsilon_i$ for $\epsilon_i \sim \mathcal{N}(0, 1)$.

Setup. We use a toy neural network $f(\boldsymbol{x}) = \boldsymbol{w}_1^\top \phi(W_2 \boldsymbol{x})$ with $\boldsymbol{w}_1 \in \mathbb{R}^{100}$, $W_2 \in \mathbb{R}^{100 \times 10}$, and the activation function $\phi(\cdot)$ is ELU function. The training loss is MES loss. We compare the convergence rates in terms of gradient norm of the proposed algorithms PSI-SGD and PSI-SGDM with three baselines algorithms. The benchmark optimization algorithm SGD with momentum (abbrev. SGD); adaptive learning rate algorithm Adam [Diederik P. Kingma, 2015]; manifold based algorithm SGDG [Cho and Lee, 2017] which is applied to the network with BN.

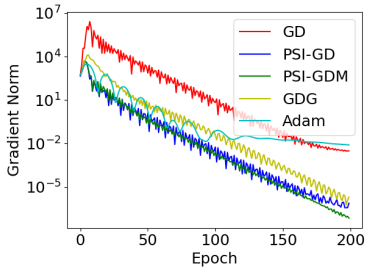

Figure 3: Convergence rates of gradient norm on a toy example.

Main Results. The convergence rates of gradient norm are in Figure **??**. As can be seen, the three manifold based methods PSI-SGD, PSI-SGDM, SGDG exhibit improved convergence rates compared other methods. However, we still observe that PSI-SGD and PSI-SGDM are slightly better than SGDG. More important is that in contrast to SGDG, our algorithms have proved convergence rate. The toy example verifies that our algorithms have improved convergence rates compared with other methods.

6.2 EXPERIMENTS ON REAL-WORLD DATASET

In this section, we empirically study the proposed algorithms PSI-SGD and PSI-SGDM on real-world dataset.

Data. We consider the image classification task on three benchmark datasets.CIFAR10 and CIFAR100 [Krizhevsky and Hinton, 2009] are respectively colorful images with 50K training samples and 10K validation samples from 10 and 100 categories. ImageNet [Deng et al., 2009] are colorful images with 1M+ training samples from

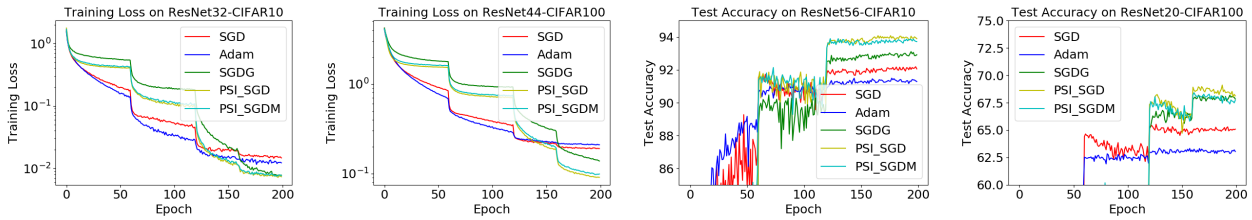

Figure 4: Results of ResNet trained over various Algorithms on CIFAR10 and CIFAR100.

1K object classes.

Setup. The model is a unified structure ResNet with various structures. As in the above section, we compare our methods with SGD, Adam, and SGDG. For PSI-SGD or PSI-SGDM, the PSI parameters are updated by the two algorithms, while the other parameters are updated by SGD.

We do not use the regularizer to the PSI weights e.g., l_2-regularizer in the loss function since it breaks the PSI property of PSI weights. More experiments with regularized refer to Appendix F.

For CIFAR we conduct 200 epochs of training for each algorithm. The learning rate starts from 0.1 and decays by a factor 0.2 at epochs 60, 120, and 160. For ImageNet, the training is conducted for 100 epochs, and the learning rate starts from 0.1 and decays by a factor 0.2 at epochs 30, 60, and 90. For the hyperparameters of baseline methods, we grid search the learning rates in the range of {0.01, 0.1, 1.0} and {0.0001, 0.001, 0.01} respectively for SGD and Adam, and the hyperparameters of SGDG follow the one of [Cho and Lee, 2017]. The other hyperparameters of all these methods are summarized in Appendix F.2.

It worth noting that the PSI weights updated by PSI-SGD and PSI-SGDM may overflow after quite a number of iterations. Hence, we normalize the PSI weights $\boldsymbol{w}^{(i)}$ if $\|\boldsymbol{w}^{(i)}\|$ is larger than 10000 during training. This operation will not change the output of the model due to the PSI property.

Main Results. We report the test accuracy of each method. The results refers to Table 1 and Figure 4. We have the following observations and conclusions from the experimental results.

1. In Table 1, the model trained by manifold based methods i.e., SGDG, PSI-SGD and, PSI-SGDM generalize better in most cases. this is due to the manifold based algorithms can avoid quite a lot of local minima with poor generalization, since they have a unified optimization path for the parameters equivalent to each other.

   Besides that, the proposed PSI-SGD and PSI-SGDM are significantly better than SGDG, and the performances of PSI-SGD and PSI-SGDM are comparable. We speculate this is due to the sharper convergence rate of PSI-SGD and PSI-SGDM allow them to find local minima in less number of iterations.

2. Figure 4 show that the PSI-SGD and PSI-SGDM converge faster than the three baselines after a certain number of iterations i.e., after 120 epochs of update. This justifies our theoretical results in Section 5, since the acceleration is linearly with $\|\boldsymbol{w}_t^{(i)}\|^2$ which is large after a while of training (Lemma 1).

   We present the results of training error instead of gradient norm here because evaluating the gradient norm requires implementing back propagations on all training data which brings a great extra computational effort, especially for large scale dataset e.g., ImageNet.

We have one remark about the efficiency of our methods. They are simple and efficient compared with SGDG, since SGDG involves the operators like trigonometric functions in the update rule. For example, the elapsed time of ResNet56 with respect to one-epoch training of CIFAR10 under SGD, Adam, SGDG, PSI-SGD, and PSI-SGDM are respectively 16.98, 18.1, 26.6, 18.2, and 18.5 seconds. All the experiments are conducted on a server with single NVIDIA V100 GPU.

7 CONCLUSION

In this paper, we fix the optimization ambiguity brought by the PSI property of the network with BN. Our scenario is built upon optimization on manifold by constructing a specific manifold and optimizing the PSI weights of the network with BN in it. The developed gradient-based algorithms on PSI manifold are shown to have a well-defined optimization path with respect to positively equivalent rescaling.

We also give the convergence rates of the proposed methods. Besides that, we theoretically justify that PSI-GD and PSI-SGD accelerate training by a clever schedule of adaptive learning rate.

Finally, we conduct various experiments to show that the proposed methods have better performance in the generalization and efficiency compared with the other three baselines.

REFERENCES

P-A Absil and Kyle A Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Vijay Badrinarayanan, Bamdev Mishra, and Roberto Cipolla. Symmetry-invariant optimization in deep networks. *arXiv preprint arXiv:1511.01754*, 2015.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1): 1–33, 2019.

Mario Lezcano Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, pages 9157–9168, 2019.

Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. In *Advances in Neural Information Processing Systems*, pages 5225–5235, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Lei Huang, Xianglong Liu, Bo Lang, and Bo Li. Projection based weight normalization for deep neural networks. *arXiv preprint arXiv:1710.02338*, 2017.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. 2016.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *CIFAR*, 2009.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. *arXiv preprint arXiv:1901.08428*, 2019.

Jun Li, Fuxin Li, and Sinisa Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. In *International Conference on Learning Representations*, 2019.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2019.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4868–4877, 2017.

Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, Nenghai Yu, and Tie-Yan Liu. G-sgd: Optimizing relu neural networks in its positively scale-invariant space. In *International Conference on Learning Representations*, 2018.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in neural information processing systems*, pages 2422–2430, 2015.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.

Xiaoxia Wu, Rachel Ward, and Léon Bottou. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

Shuxin Zheng, Qi Meng, Huishuai Zhang, Wei Chen, Nenghai Yu, and Tie-Yan Liu. Capacity control of relu neural networks by basis-path norm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5925–5932, 2019.

A RELATED DEFINITIONS OF OPTIMIZATION ON MANIFOLD

We give a brief introduction to optimization on manifold in this section. A summary of this topic can be referred to [Absil et al., 2009]. This paper focuses on the matrix manifold which is a subspace of Euclidean space $\mathbb{R}^n$. We start with the definition of matrix manifold.

Definition 2 (Matrix manifold)**.** *A Matrix Manifold $\mathcal{M}$ is a subset of Euclidean space $\mathbb{R}^n = \mathbb{R}^{m\times p}$ with any $\boldsymbol{x} \in \mathcal{M}$ has a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$ such that $U_{\boldsymbol{x}}$ is homeomorphic to a Euclidean space. In addition, for a given equivalent relation $\sim$ of elements in the manifold $\overline{\mathcal{M}}$, we use $[\boldsymbol{x}]$ to denote set $\{\boldsymbol{y} \in \overline{\mathcal{M}} : \boldsymbol{y} \sim \boldsymbol{x}\}$. Then $\mathcal{M} = \overline{\mathcal{M}}/\sim = \{[\boldsymbol{x}] : \boldsymbol{x} \in \overline{\mathcal{M}}\}$ is a quotient manifold.*[2]

Various spaces can be categorized into matrix manifold, e.g. Euclidean space, unit ball. For a manifold $\mathcal{M}$ with $\boldsymbol{x} \in \mathcal{M}$, there is a tangent space $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ which gives the tangential direction of $\boldsymbol{x} \in \mathcal{M}$. For matrix manifold, tangent space $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ must be a linear space with finite dimension [Absil et al., 2009], it specifies the moving direction of a point in the manifold. Since manifold can be non-linear space, the movement of a point $\boldsymbol{x} \in \mathcal{M}$ along a specific direction $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ is decided by the retraction function $R_{\boldsymbol{x}}(\cdot) : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} :\to \mathcal{M}$, which is defined as follows.

Definition 3 (Retraction)**.** *A retraction on a manifold $\mathcal{M}$ is a smooth mapping $R_{\boldsymbol{x}}(\cdot) : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} :\to \mathcal{M}$ satisfies:*

(1) $R_{\boldsymbol{x}}(0_{\boldsymbol{x}}) = \boldsymbol{x}$, where $0_{\boldsymbol{x}}$ denotes the zero element of $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$.

(2) $R_{\boldsymbol{x}}(\cdot)$ satisfies

$$\lim_{t\to\infty} \frac{R_{\boldsymbol{x}}(0_{\boldsymbol{x}} + t\boldsymbol{\Xi}) - R_{\boldsymbol{x}}(0_{\boldsymbol{x}})}{t} = \boldsymbol{\Xi}$$

for any $\boldsymbol{\Xi} \in \mathcal{T}_{\mathcal{M}}$.

For instance, the retraction function of $\mathbb{R}^n$ can be defined as $R_{\boldsymbol{x}}(\boldsymbol{\Xi}) = \boldsymbol{x} + \boldsymbol{\Xi}$; and for unit ball it can be written as $R_{\boldsymbol{x}}(\boldsymbol{\Xi}) = \frac{\boldsymbol{x}+\boldsymbol{\Xi}}{\|\boldsymbol{x}+\boldsymbol{\Xi}\|}$ where $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$.

To obtain the gradient based optimization algorithms on manifold, we should give the definition of Riemannian gradient in manifold. The following Riemannian metric derives the definition of corresponded gradient.

Definition 4 (Riemannian metric)**.** *For matrix manifold, Reimannian metric is an inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{x}}$ of tangent space $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ for any $\boldsymbol{x} \in \mathcal{M}$. In addition, given a set of coordinate basis vector $\{E_i\}$ of $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$, then for any $\boldsymbol{\Xi}_{\boldsymbol{x}}$ and $\boldsymbol{\zeta}_{\boldsymbol{x}}$, we have*

$$\langle \boldsymbol{\Xi}_{\boldsymbol{x}}, \boldsymbol{\zeta}_{\boldsymbol{x}} \rangle_{\boldsymbol{x}} = \hat{\boldsymbol{\Xi}}_{\boldsymbol{x}}^T G_{\boldsymbol{x}} \hat{\boldsymbol{\zeta}}_{\boldsymbol{x}}. \tag{25}$$

Here $\hat{\boldsymbol{\Xi}}_{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{x}}$ are respective coordinates of $\boldsymbol{\Xi}_{\boldsymbol{x}}$ and $\boldsymbol{\zeta}_{\boldsymbol{x}}$ under the coordinate basis vector $\{E_i\}$, and $G_{\boldsymbol{x}} = (g)_{ij}$ where $g_{ij} = \langle E_i, E_j \rangle_{\boldsymbol{x}}$.

Based on the Riemannian metric, the Riemannian gradient is defined as follows.

Definition 5 (Riemannian gradient)**.** *The Riemannian gradient of function $f(\cdot)$ defined in an open set containing manifold $\mathcal{M}$ at $\boldsymbol{x} \in \mathcal{M}$ is the tangent vector $\mathrm{grad} f(\boldsymbol{x})$ belongs to $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ satisfies that*

$$\lim_{t\to\infty} \frac{f(\boldsymbol{x} + t\boldsymbol{\Xi}) - f(\boldsymbol{x})}{t} = \langle \boldsymbol{\Xi}_{\boldsymbol{x}}, \mathrm{grad} f(\boldsymbol{x}) \rangle_{\boldsymbol{x}}$$

for any $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$.

In fact, if function $f(\cdot)$ has gradient at point $\boldsymbol{x}$, then we see

$$\begin{aligned} \langle \boldsymbol{\Xi}_{\boldsymbol{x}}, \nabla f(\boldsymbol{x}) \rangle &= \lim_{t\to\infty} \frac{f(\boldsymbol{x} + t\boldsymbol{\Xi}_{\boldsymbol{x}}) - f(\boldsymbol{x})}{t} = \langle \boldsymbol{\Xi}_{\boldsymbol{x}}, \mathrm{grad} f(\boldsymbol{x}) \rangle_{\boldsymbol{x}} \\ &= \boldsymbol{\Xi}_{\boldsymbol{x}}^T G_{\boldsymbol{x}} \mathrm{grad} f(\boldsymbol{x}). \end{aligned} \tag{26}$$

Hence, we have $\mathrm{grad} f(\boldsymbol{x}) = G_{\boldsymbol{x}}^{-1} \nabla f(\boldsymbol{x})$ by the arbitrariness of $\boldsymbol{\Xi}_{\boldsymbol{x}}$.

Boumal et al. [2019] proves that for function $f(\cdot)$ defined on manifold $\mathcal{M}$ with Lipschitz Riemannnian gradient[3], gradient descent in manifold

$$\boldsymbol{x}_{t+1} = R_{\boldsymbol{x}_t}(-\mathrm{grad} f(\boldsymbol{x}_t)) \tag{27}$$

can convergence to a local minimum of $f(\cdot)$.

[2]Please notice we use $\mathcal{M}$ to represent quotient manifold while $\overline{\mathcal{M}}$ is the original manifold equipped with equivalent relationship.

[3]Which means there exist a positive number L_g satisfy $\|\mathrm{grad} f(\boldsymbol{x}) - \mathrm{grad} f(\boldsymbol{y})\| \le L_g \|\boldsymbol{x} - \boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$

B PROOFS IN SECTION 4

B.1 PROOF OF PROPOSITION 2

Proof. It is easily to verify that the function $\langle \cdot, \cdot \rangle_{\boldsymbol{W}}$ is an inner product for every $\boldsymbol{W} \in \mathcal{M}$. We notice that $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{W}}\mathcal{M}$ is a representation $\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}} \in \mathcal{T}_{\overline{\boldsymbol{W}}}\overline{\mathcal{M}}$ for some $\overline{\boldsymbol{W}} \in \pi^{-1}(\boldsymbol{W})$. Similar to Example 3.5.4 in [Absil et al., 2009], we see different representation of $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{W}}\mathcal{M}$ satisfies that

$$\bar{\boldsymbol{\Xi}}_{T_{\boldsymbol{a}}(\overline{\boldsymbol{W}})} = T_{\boldsymbol{a}}(\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}}). \tag{28}$$

Then, combining equation (6), we can conclude that the Riemannian metric is well-defined to the choice of horizontal lift $\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}}$. Hence, we get the conclusion. □

B.2 PROOF OF PROPOSITION 3

We need a Lemma to give the proof.

Lemma 2 (Proposition 4.1.3 in [Absil et al., 2009])**.** *Let $\mathcal{M} = \overline{\mathcal{M}}/\sim$ be a quotient manifold and $\bar{R}$ be a retraction on $\overline{\mathcal{M}}$ such that for all $x \in \mathcal{M}$ any two $\bar{\boldsymbol{x}}_a, \bar{\boldsymbol{x}}_b \in \pi^{-1}(\boldsymbol{x})$ and $\boldsymbol{\xi} \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$,*

$$\pi(\bar{R}_{\bar{\boldsymbol{x}}_a}(\bar{\boldsymbol{\xi}}_{\bar{\boldsymbol{x}}_a})) = \pi(\bar{R}_{\bar{\boldsymbol{x}}_b}(\bar{\boldsymbol{\xi}}_{\bar{\boldsymbol{x}}_b})). \tag{29}$$

Here $\pi(\cdot)$ is a canonical projection from $\overline{\mathcal{M}}$ to $\mathcal{M}$ satisfies that $\pi(\boldsymbol{x}) = \pi(\boldsymbol{y})$ if and only if $\boldsymbol{x} \sim \boldsymbol{y}$. $\bar{\boldsymbol{\xi}}_{\bar{\boldsymbol{x}}_a}$, $\bar{\boldsymbol{\xi}}_{\bar{\boldsymbol{x}}_b}$ are respectively tangent vector in $\mathcal{T}_{\bar{\boldsymbol{x}}_a}\overline{\mathcal{M}}$ and $\mathcal{T}_{\bar{\boldsymbol{x}}_b}\overline{\mathcal{M}}$. $\bar{R}_{\cdot}(\cdot)$ is a retraction in $\overline{\mathcal{M}}$. Then

$$R_{\boldsymbol{x}}(\boldsymbol{\xi}) = \pi(\bar{R}_{\bar{\boldsymbol{x}}}(\bar{\boldsymbol{\xi}}_{\bar{\boldsymbol{x}}})) \tag{30}$$

defines a retraction on quotient manifold $\mathcal{M}$.

With this Lemma, we now provide the proof of Proposition 3.

Proof of Proposition 3. According to the lemma 2, we only need to verify the conditions. First we notice that $\boldsymbol{\Xi} \in \mathcal{T}_W\mathcal{M}$ is a representation $\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}} \in \mathcal{T}_{\overline{\boldsymbol{W}}}\overline{\mathcal{M}}$ for some $\overline{\boldsymbol{W}} \in \pi^{-1}(\boldsymbol{W})$. Similar to the Example 3.5.4 in [Absil et al., 2009], different representation of $\boldsymbol{\Xi} \in \mathcal{T}_{\boldsymbol{W}}\mathcal{M}$ satisfies

$$\bar{\boldsymbol{\Xi}}_{T_{\boldsymbol{a}}(\overline{\boldsymbol{W}})} = T_{\boldsymbol{a}}(\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}}). \tag{31}$$

Hence we can choose

$$\bar{R}_{\overline{\boldsymbol{W}}}(\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}}) = \overline{\boldsymbol{W}} + \bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}}. \tag{32}$$

Thus $\bar{R}_{\overline{\boldsymbol{W}}}(\cdot)$ is a retraction in $\overline{\mathcal{M}}$. Then, for any canonical projection $\pi(\cdot)$ there exist $\pi(\bar{R}_{\overline{\boldsymbol{W}}}(\bar{\boldsymbol{\Xi}}_{\overline{\boldsymbol{W}}})) = \pi(\bar{R}_{\mathcal{T}_{\boldsymbol{a}}(\overline{\boldsymbol{W}})}(\bar{\boldsymbol{\Xi}}_{\mathcal{T}_{\boldsymbol{a}}(\overline{\boldsymbol{W}})}))$. Hence, we get the conclusion. □

C PROOFS IN SECTION 4.2

C.1 PROOF OF THEOREM 1

Proof. Since

$$\nabla_{\boldsymbol{w}^{(i)}}\mathcal{L}(\boldsymbol{W}, \boldsymbol{g}) = a_i \nabla_{a_i\boldsymbol{w}^{(i)}}\mathcal{L}(T_{\boldsymbol{a}}(\boldsymbol{W}), \boldsymbol{g}) \tag{33}$$

due to the PSI property. Then, combining the update rule of PSI-SGDM (11), we have

$$\begin{aligned} \boldsymbol{U}_0 &\xrightarrow{\text{update}} \boldsymbol{U}_1; \qquad T_{\boldsymbol{a}}(\boldsymbol{U}_0) \xrightarrow{\text{update}} T_{\boldsymbol{a}}(\boldsymbol{U}_1) \\ \boldsymbol{W}_0 &\xrightarrow{\text{update}} \boldsymbol{W}_1; \qquad T_{\boldsymbol{a}}(\boldsymbol{W}_0) = \hat{\boldsymbol{W}}_0 \xrightarrow{\text{update}} \hat{\boldsymbol{W}}_1 = T_{\boldsymbol{a}}(\boldsymbol{W}_1). \end{aligned} \tag{34}$$

By induction, we get the conclusion. □

D PROOFS IN SECTION 5

D.1 PROOF OF THEOREM 4

Proof. By Taylor's expansion, and condition of Lipschitz gradient

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}_{t+1}, \boldsymbol{g}_{t+1}) - \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) &= \int_0^1 \langle \nabla_{\boldsymbol{W}_t, \boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t + s\Delta \boldsymbol{W}_t, \boldsymbol{g}_t + s\Delta \boldsymbol{g}_t), (\Delta \boldsymbol{W}_t, \Delta \boldsymbol{g}_t) \rangle \, ds \\
&= \langle \nabla_{\boldsymbol{W}, \boldsymbol{g}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t), (\boldsymbol{W}_t, \boldsymbol{g}_t) \rangle \\
&+ \int_0^1 \langle \nabla_{\boldsymbol{W}_t, \boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t + s\Delta \boldsymbol{W}_t, \boldsymbol{g}_t + s\Delta \boldsymbol{g}_t) - \nabla_{\boldsymbol{W}, \boldsymbol{g}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t), (\Delta \boldsymbol{W}_t, \Delta \boldsymbol{g}_t) \rangle \, ds \\
&\leq \langle \nabla_{\boldsymbol{W}, \boldsymbol{g}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t), (\boldsymbol{W}_t, \boldsymbol{g}_t) \rangle + \frac{1}{2} \sum_{i,j=1}^m L_{ij}^{\boldsymbol{ww}} \left\| \Delta \boldsymbol{w}_t^{(i)} \right\| \left\| \Delta \boldsymbol{w}_t^{(j)} \right\| \\
&+ \sum_{i=1}^m L_i^{\boldsymbol{wg}} \left\| \Delta \boldsymbol{w}_t^{(i)} \right\| \| \Delta \boldsymbol{g}_t \| + \frac{m L^{\boldsymbol{gg}}}{2} \| \| \Delta \boldsymbol{g}_t \|
\end{aligned} \tag{35}$$

Thus,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}_{t+1}, \boldsymbol{g}_{t+1}) - \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \leq & - \sum_{i=1}^m \left(\eta_{\boldsymbol{w}_t^{(i)}} \left\| \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 + \eta_{\boldsymbol{g}_t} \left\| \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 \right) \\
& + \frac{1}{2} \sum_{i,j=1}^m L_{ij}^{\boldsymbol{ww}} \left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \left\| \eta_{\boldsymbol{w}_t^{(j)}} \nabla_{\boldsymbol{w}_t^{(j)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \\
& + \sum_{i=1}^m L_i^{\boldsymbol{wg}} \left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \left\| \eta_{\boldsymbol{g}_t} \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \\
& + \frac{m L^{\boldsymbol{gg}}}{2} \left\| \eta_{\boldsymbol{g}_t} \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 .
\end{aligned} \tag{36}$$

Then, by Young's inequality,

$$\begin{aligned}
\left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \left\| \eta_{\boldsymbol{w}_t^{(j)}} \nabla_{\boldsymbol{w}_t^{(j)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| &\leq \frac{1}{2} \left(\left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 + \left\| \eta_{\boldsymbol{w}_t^{(j)}} \nabla_{\boldsymbol{w}_t^{(j)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 \right) \\
\left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| \left\| \eta_{\boldsymbol{g}_t} \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\| &\leq \frac{1}{2} \left(\left\| \eta_{\boldsymbol{w}_t^{(i)}} \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 + \left\| \eta_{\boldsymbol{g}_t} \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 \right).
\end{aligned} \tag{37}$$

Then, we see

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}_{t+1}, \boldsymbol{g}_{t+1}) - \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) &\leq \sum_{i=1}^m \left(-\eta_{\boldsymbol{w}_t^{(i)}} + \frac{\tilde{L}_{\boldsymbol{w}^{(i)}}}{2} \eta_{\boldsymbol{w}_t^{(i)}}^2 \right) \left\| \nabla_{\boldsymbol{w}_t^{(i)}} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \right\|^2 \\
&+ \left(-\eta_{\boldsymbol{g}_t} + \frac{\tilde{L}_{\boldsymbol{g}}}{2} \eta_{\boldsymbol{g}_t}^2 \right) \| \nabla_{\boldsymbol{g}_t} \mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) \|^2 \\
&\leq \| \nabla \mathcal{L}_{\boldsymbol{W}_t, \boldsymbol{g}_t}(\boldsymbol{W}_t, \boldsymbol{g}_t) \|^2 \max_{1 \leq i \leq m} \left\{ \left(-\eta_{\boldsymbol{w}_t^{(i)}} + \frac{\tilde{L}_{\boldsymbol{w}^{(i)}}}{2} \eta_{\boldsymbol{w}_t^{(i)}}^2 \right) \right\} \wedge \left(-\eta_{\boldsymbol{g}_t} + \frac{\tilde{L}_{\boldsymbol{g}}}{2} \eta_{\boldsymbol{g}_t}^2 \right)
\end{aligned} \tag{38}$$

where

$$\tilde{L}_{\boldsymbol{w}^{(i)}} = L_i^{\boldsymbol{wg}} + \sum_{j=1}^m L_{ij}^{\boldsymbol{ww}}; \qquad \tilde{L}_{\boldsymbol{g}} = m L^{\boldsymbol{gg}} + \sum_{i=1}^m L_i^{\boldsymbol{wg}}. \tag{39}$$

Summing over equation (38) from 0 to $T-1$, let $0 \leq \eta_{\boldsymbol{w}_t^{(i)}} \leq 2/\tilde{L}_{\boldsymbol{w}^{(i)}}$, we get

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_T, \boldsymbol{g}_T) - \mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) \\
&\qquad \leq \min_{0 \leq t < T} \| \nabla \mathcal{L}_{\boldsymbol{W}_t, \boldsymbol{g}_t}(\boldsymbol{W}_t, \boldsymbol{g}_t) \|^2 \sum_{t=0}^{T-1} \max_{1 \leq i \leq m} \left\{ \left(-\eta_{\boldsymbol{w}_t^{(i)}} + \frac{\tilde{L}_{\boldsymbol{w}^{(i)}}}{2} \eta_{\boldsymbol{w}_t^{(i)}}^2 \right) \right\} \wedge \left(-\eta_{\boldsymbol{g}_t} + \frac{\tilde{L}_{\boldsymbol{g}}}{2} \eta_{\boldsymbol{g}_t}^2 \right).
\end{aligned} \tag{40}$$

We see the best selection of $\eta_{\boldsymbol{w}_t^{(i)}}$ and $\eta_{\boldsymbol{g}_t}$ are respectively $1/\tilde{L}_{\boldsymbol{w}^{(i)}}$ and $1/\tilde{L}_{\boldsymbol{g}}$. By the value of $\eta_{\boldsymbol{w}_t^{(i)}}$ and $\eta_{\boldsymbol{g}_t}$, we get

$$\min_{0 \leq t < T} \| \nabla \mathcal{L}_{\boldsymbol{W}_t, \boldsymbol{g}_t}(\boldsymbol{W}_t, \boldsymbol{g}_t) \|^2 \max \left\{ \frac{1}{2\tilde{L}_{\boldsymbol{w}^{(i)}}}, \cdots, \frac{1}{2\tilde{L}_{\boldsymbol{w}^{(m)}}}, \frac{1}{2\tilde{L}_{\boldsymbol{g}}} \right\} \leq \frac{\mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*)}{T}. \tag{41}$$

Then we conclude the proof. □

D.2 PROOF OF THEOREM 5

Proof. By conditions of Lipschitz gradient, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_{t+1},\boldsymbol{g}_{t+1})-\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\leq&-\sum_{i=1}^{m}\left(\eta_{\boldsymbol{w}_t^{(i)}}\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2+\eta_{\boldsymbol{g}_t}\left\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right)\\&+\frac{1}{2}\sum_{i,j=1}^{m}L_{ij}^{\boldsymbol{ww}}\left\|\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{w}_t^{(j)}}\mathcal{G}_{\boldsymbol{w}_t^{(j)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\\&+\sum_{i=1}^{m}L_i^{\boldsymbol{wg}}\left\|\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\\&+\frac{mL^{\boldsymbol{gg}}}{2}\left\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2.\end{aligned}\tag{42}$$

By Young's inequality, taking expectation conditional in $(\boldsymbol{W}_t,\boldsymbol{g}_t)$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{w}_t^{(j)}}\mathcal{G}_{\boldsymbol{w}_t^{(j)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\right]\\&\leq\frac{1}{2}\left(\mathbb{E}\left[\left\|\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]+\mathbb{E}\left[\left\|\eta_{\boldsymbol{w}_t^{(j)}}\mathcal{G}_{\boldsymbol{w}_t^{(j)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right)\\&=\frac{1}{2}\eta_{\boldsymbol{w}_t^{(i)}}^2\mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)-\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)+\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&+\frac{1}{2}\eta_{\boldsymbol{w}_t^{(j)}}^2\mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(j)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)-\nabla_{\boldsymbol{w}_t^{(j)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)+\nabla_{\boldsymbol{w}_t^{(j)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&\overset{a}{=}\frac{1}{2}\eta_{\boldsymbol{w}_t^{(i)}}^2\mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)-\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2+\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&+\frac{1}{2}\eta_{\boldsymbol{w}_t^{(j)}}^2\mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(j)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)-\nabla_{\boldsymbol{w}_t^{(j)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2+\left\|\nabla_{\boldsymbol{w}_t^{(j)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&\leq\frac{1}{2}\eta_{\boldsymbol{w}_t^{(i)}}^2\left(\sigma^2+\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right)+\frac{1}{2}\eta_{\boldsymbol{w}_t^{(j)}}^2\left(\sigma^2+\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(j)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right),\end{aligned}\tag{43}$$

where a is due to equation (14). Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\eta_{\boldsymbol{w}_t^{(i)}}\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|\right]\leq&\frac{1}{2}\eta_{\boldsymbol{w}_t^{(i)}}^2\left(\sigma^2+\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right)\\&+\frac{1}{2}\eta_{\boldsymbol{g}_t}^2\left(\sigma^2+\mathbb{E}\left[\left\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right),\end{aligned}\tag{44}$$

and

$$\mathbb{E}\left[\left\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\leq\frac{1}{2}\eta_{\boldsymbol{g}_t}^2\left(\sigma^2+\mathbb{E}\left[\left\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\right).\tag{45}$$

Fixing the randomness of $\boldsymbol{W}_t,\boldsymbol{g}_t$, plugging equation (43), (44) and (45) into (42), then taking expectation, we get

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\boldsymbol{W}_{t+1},\boldsymbol{g}_{t+1})\right]-\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\leq&\sum_{i=1}^{m}\left(-\eta_{\boldsymbol{w}_t^{(i)}}+\frac{\tilde{L}_{\boldsymbol{w}^{(i)}}}{2}\eta_{\boldsymbol{w}_t^{(i)}}^2\right)\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&+\left(-\eta_{\boldsymbol{g}_t}+\frac{\tilde{L}_{\boldsymbol{g}}}{2}\eta_{\boldsymbol{g}_t}^2\right)\mathbb{E}\left[\left\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]+\frac{\sigma^2}{2}\left(\eta_{\boldsymbol{g}_t}^2\tilde{L}_{\boldsymbol{g}}+\sum_{i=1}^{m}\eta_{\boldsymbol{w}_t^{(i)}}^2\tilde{L}_{\boldsymbol{w}^{(i)}}\right)\\\overset{a}{\leq}&-\frac{1}{\sqrt{T}}\sum_{i=1}^{m}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]-\frac{1}{\sqrt{T}\tilde{L}_{\boldsymbol{g}}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\\&+\frac{\sigma^2}{2T}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}}+\sum_{i=1}^{m}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right)\\\leq&-\frac{1}{\sqrt{T}\tilde{L}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]+\frac{\sigma^2}{2T}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}}+\sum_{i=1}^{m}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right),\end{aligned}\tag{46}$$

where a is due to the value of $\eta_{\boldsymbol{w}_t^{(i)}}$ and $\eta_{\boldsymbol{g}_t}$, $\tilde{L}$ is $\max\{\tilde{L}_{\boldsymbol{w}^{(1)}},\cdots,\tilde{L}_{\boldsymbol{w}^{(m)}},\tilde{L}_{\boldsymbol{g}}\}$. Summing over from 0 to $T-1$ of equation (46), we get

$$\sum_{t=0}^{T-1}\frac{1}{\sqrt{T}\tilde{L}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\right\|^2\right]\leq\mathcal{L}(\boldsymbol{W}_0,\boldsymbol{g}_0)-\mathbb{E}\left[\mathcal{L}(\boldsymbol{W}_T,\boldsymbol{g}_T)\right]+\frac{\sigma^2}{2}\sum_{i=1}^{m}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}}+\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right)\sqrt{T}\tag{47}$$

Thus, we have

$$\left(\frac{\sqrt{T}}{\tilde{L}}\right)\min_{0\le t<T}\mathbb{E}\left[\|\nabla_{\boldsymbol{W}_t,\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\|^2\right]\le\mathcal{L}(\boldsymbol{W}_0,\boldsymbol{g}_0)-\mathcal{L}(\boldsymbol{W}^*,\boldsymbol{g}^*)+\frac{\sigma^2}{2}\sum_{i=1}^{m}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}}+\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right)\tag{48}$$

Then we get the conclusion. □

D.3 PROOF OF THEOREM 2

Proof. By assumption (13) and PSI property, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_{t+1},\boldsymbol{g}_{t+1})-\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\le&-\sum_{i=1}^{m}\left(\eta_{\boldsymbol{v}_t^{(i)}}\left\|\nabla_{\boldsymbol{v}_t^{(i)}}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|^2+\eta_{\boldsymbol{g}_t}\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|^2\right)\\&+\frac{1}{2}\sum_{i,j=1}^{m}L_{ij}^{\boldsymbol{vv}}\left\|\eta_{\boldsymbol{v}_t^{(i)}}\nabla_{\boldsymbol{v}_t^{(i)}}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{v}_t^{(j)}}\nabla_{\boldsymbol{v}_t^{(j)}}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\\&+\sum_{i=1}^{m}L_{i}^{\boldsymbol{vg}}\left\|\eta_{\boldsymbol{v}_t^{(i)}}\nabla_{\boldsymbol{v}_t^{(i)}}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\|\eta_{\boldsymbol{g}_t}\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|\\&+\frac{mL^{\boldsymbol{gg}}}{2}\|\eta_{\boldsymbol{g}_t}\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|^2.\end{aligned}\tag{49}$$

Here we use one fact that

$$\begin{aligned}\|\boldsymbol{w}^{(i)}\|\|\boldsymbol{w}^{(j)}\|\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{w}^{(j)}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})=\nabla^2_{\boldsymbol{v}^{(i)}\boldsymbol{v}^{(j)}}\mathcal{L}(\boldsymbol{V},\boldsymbol{g});\\\|\boldsymbol{w}^{(i)}\|\nabla^2_{\boldsymbol{w}^{(i)}\boldsymbol{g}}\mathcal{L}(\boldsymbol{W},\boldsymbol{g})=\nabla^2_{\boldsymbol{v}^{(i)}\boldsymbol{g}}\mathcal{L}(\boldsymbol{V},\boldsymbol{g}).\end{aligned}\tag{50}$$

Then follow the proof of Theorem 4, we get the conclusion. □

D.4 PROOF OF THEOREM 3

Proof. By noticing that

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_{t+1},\boldsymbol{g}_{t+1})-\mathcal{L}(\boldsymbol{W}_t,\boldsymbol{g}_t)\le&-\sum_{i=1}^{m}\left(\eta_{\boldsymbol{v}_t^{(i)}}\left\|\nabla_{\boldsymbol{v}_t^{(i)}}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|^2+\eta_{\boldsymbol{g}_t}\|\nabla_{\boldsymbol{g}_t}\mathcal{L}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|^2\right)\\&+\frac{1}{2}\sum_{i,j=1}^{m}L_{ij}^{\boldsymbol{vv}}\left\|\eta_{\boldsymbol{v}_t^{(i)}}\mathcal{G}_{\boldsymbol{v}_t^{(i)}}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\left\|\eta_{\boldsymbol{v}_t^{(j)}}\mathcal{G}_{\boldsymbol{v}_t^{(j)}}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\\&+\sum_{i=1}^{m}L_{i}^{\boldsymbol{vg}}\left\|\eta_{\boldsymbol{v}_t^{(i)}}\mathcal{G}_{\boldsymbol{v}_t^{(i)}}(\boldsymbol{V}_t,\boldsymbol{g}_t)\right\|\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|\\&+\frac{mL^{\boldsymbol{gg}}}{2}\|\eta_{\boldsymbol{g}_t}\mathcal{G}_{\boldsymbol{g}_t}(\boldsymbol{V}_t,\boldsymbol{g}_t)\|^2.\end{aligned}\tag{51}$$

Following the proof of Theorem 5, we get the desired conclusion. □

E UPPER BOUND TO $\|\boldsymbol{w}_t^{(i)}\|^2$

In this section we respectively give an upper bound to $\|\boldsymbol{w}_t^{(i)}\|^2$ for vanilla GD and SGD. The conclusions are similar to the results in [Arora et al., 2018].

Proposition 4. *For iterates in Theorem 4, we have*

$$\|\boldsymbol{w}_t^{(i)}\|^2\le\|\boldsymbol{w}_0^{(i)}\|^2+\frac{2(\mathcal{L}(\boldsymbol{W}_0,\boldsymbol{g}_0)-\mathcal{L}(\boldsymbol{W}^*,\boldsymbol{g}^*))}{L_{\boldsymbol{w}^{(i)}}},\tag{52}$$

for any $0\le t\le T$ *and* $1\le i\le m$.

Proof. From equation (38), we have

$$\frac{\|\boldsymbol{w}_{t+1}^{(i)}\|^2 - \|\boldsymbol{w}_t^{(i)}\|^2}{\eta_{\boldsymbol{w}_t^{(i)}}^2} = \left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2 \leq 2L_{\boldsymbol{w}^{(i)}}(\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \mathcal{L}(\boldsymbol{W}_{t+1}, \boldsymbol{g}_{t+1})) \tag{53}$$

for $1 \leq i \leq m$. By the value of $\eta_{\boldsymbol{w}^{(i)}}$, we then have

$$L_{\boldsymbol{w}^{(i)}}^2\left(\|\boldsymbol{w}_{t+1}^{(i)}\|^2 - \|\boldsymbol{w}_t^{(i)}\|^2\right) \leq 2L_{\boldsymbol{w}^{(i)}}(\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \mathcal{L}(\boldsymbol{W}_{t+1}, \boldsymbol{g}_{t+1})). \tag{54}$$

Summing over t from 0 to $T-1$, we have

$$\|\boldsymbol{w}_T^{(i)}\|^2 - \|\boldsymbol{w}_0^{(i)}\|^2 \leq \frac{2(\mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*))}{L_{\boldsymbol{w}^{(i)}}}. \tag{55}$$

Then we get the conclusion due to the increasing property of $\|\boldsymbol{w}_t^{(i)}\|^2$ with respect to t. □

Proposition 5. *For iterates in Theorem 5, we have*

$$\mathbb{E}\left[\|\boldsymbol{w}_t^{(i)}\|^2\right] \leq \|\boldsymbol{w}_0^{(i)}\|^2 + O\left(\frac{\sigma^2}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right), \tag{56}$$

for any $0 \leq t \leq T$, *and* $1 \leq i \leq m$.

Proof. From equation (46), we have

$$\begin{aligned}
\tilde{L}_{\boldsymbol{w}^{(i)}} T\mathbb{E}\left[\|\boldsymbol{w}_{t+1}^{(i)}\|^2 - \|\boldsymbol{w}_t^{(i)}\|^2\right] &= \mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t) + \nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right] \\
&\overset{a}{=} \mathbb{E}\left[\left\|\mathcal{G}_{\boldsymbol{w}_t^{(i)}}(\boldsymbol{W}_t, \boldsymbol{g}_t) - \nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2 + \left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right] \\
&\overset{b}{\leq} \sigma^2 + \mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_t^{(i)}}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\right\|^2\right]
\end{aligned} \tag{57}$$

where a and b are due to equation (46). We then have

$$\sum_{s=0}^{t}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}\sqrt{T}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_s^{(i)}}\mathcal{L}(\boldsymbol{W}_s, \boldsymbol{g}_s)\right\|^2\right] \leq \mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*) + \sum_{s=0}^{t}\frac{\sigma^2}{2T}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}} + \sum_{i=1}^{m}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right). \tag{58}$$

for any $0 \leq t \leq T$ due to equation (46), the monotonic decreasing of $\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)$ and the value of $\eta_{\boldsymbol{w}_t^{(i)}}$. Let

$$S_t^{(i)} = \sum_{s=1}^{t}\frac{1}{\sqrt{T}}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{w}_s^{(i)}}\mathcal{L}(\boldsymbol{W}_s, \boldsymbol{g}_s)\right\|^2\right] \leq \tilde{L}_{\boldsymbol{w}^{(i)}}(\mathcal{L}(\boldsymbol{W}_0, \boldsymbol{g}_0) - \mathcal{L}(\boldsymbol{W}^*, \boldsymbol{g}^*)) + \frac{\sigma^2\tilde{L}_{\boldsymbol{w}^{(i)}}}{2}\left(\frac{1}{\tilde{L}_{\boldsymbol{g}}} + \sum_{i=1}^{m}\frac{1}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right) \tag{59}$$

Summing over equation (57) with respect to t from 0 to $T-1$, and combining the above equation, we conclude that

$$\mathbb{E}\left[\|\boldsymbol{w}_T^{(i)}\|^2 - \|\boldsymbol{w}_0^{(i)}\|^2\right] \leq \frac{\sigma^2}{\tilde{L}_{\boldsymbol{w}^{(i)}}} + \frac{S_T^{(i)}}{\sqrt{T}} = O\left(\frac{\sigma^2}{\tilde{L}_{\boldsymbol{w}^{(i)}}}\right). \tag{60}$$

Thus, we get the conclusion due to the increasing property of $\|\boldsymbol{w}_t^{(i)}\|^2$ with respect to t. □

We respectively bound the $\|\boldsymbol{w}_t^{(i)}\|^2$ for GD and SGD. Due to there is a scalar $\|\boldsymbol{w}_t^{(i)}\|^2$ is in $\|\nabla_{\boldsymbol{w}_t}\mathcal{L}(\boldsymbol{W}_t, \boldsymbol{g}_t)\|$, we conclude that both PSI-GD and PSI-SGD accelerate training within a constant scalar.

Table 2: Performance of ResNet with regularizer on `CIFAR10` and `CIFAR100`. The results in parenthesis are reported in [He et al., 2016] and [Huang et al., 2017].

Dataset	CIFAR-10					CIFAR-100				
Algorithm	SGD	Adam	SGDG	PSI-SGD	PSI-SGDM	SGD	Adam	SGDG	PSI-SGD	PSI-SGDM
ResNet20	92.35(91.25)	90.89	91.99	92.49	**92.70**	68.62(67.72)	65.66	68.67	69.62	**69.64**
ResNet32	93.64(92.49)	91.83	92.53	93.75	**93.80**	70.67(69.38)	67.72	70.49	**71.21**	71.07
ResNet44	93.74(92.83)	91.85	93.30	**94.15**	93.92	71.31(70.05)	68.24	71.49	**72.03**	72.00
ResNet56	**94.19**(93.03)	91.78	93.37	93.98	93.99	72.44(70.93)	68.78	71.56	**73.05**	72.62

F MORE EXPERIMENTAL RESULTS

F.1 EXPERIMENTS WITH REGULARIZER

The experimental results in Section 6 are conducted without regularizer e.g. l_2-regularizer. This is due to the regularizer breaks the PSI property of PSI parameters. Even though, it is desired to verify the proposed method within loss with regularizer.

Thus we conduct the experiments on `CIFAR10` and `CIFAR100` to see the performance of the proposed methods on the loss with l_2-regularizer. The settings follow Section 6 in the main part of the paper, expected for the loss has l_2-regularizer [4]. The results refers to Table 2. We can observe from the result that adding regularizer significantly improve the performance of SGD and Adam, while the improvement on the manifold based methods is incremental. Even though, optimization on PSI manifold always finds minimum generalize better. We suggest this is because optimizing on the PSI manifold corresponds with a larger learning rate, then the iterates are away from the initialized point. Hence they are able to find flatter minima which is empirically observed to generalize better [Keskar et al., 2016].

Finally, all the experimental results refers to Figure 5 and 6.

F.2 HYPERPARAMETERS

The PSI parameters are updated by the various methods while non-PSI parameters are all updated by SGD. The hyperparameters refers to Table 3.

Table 3: The hyperparameters of the experiments

Hyperparameters	SGD	Adam	SGDG	PSI-SGD	PSI-SGDM
Learning Rate	0.1	0.001	0.2	1.0	0.1
Regularizer	0.0005	0.0005	0.1	0.0005	0.0005
Batch Size	128	128	128	128	128
Momentum	0.9	-	-	-	0.9
β_1	-	0.9	-	-	-
β_2	-	0.999	-	-	-

[4] The regularizer of SGDG follows the setting in [Cho and Lee, 2017].

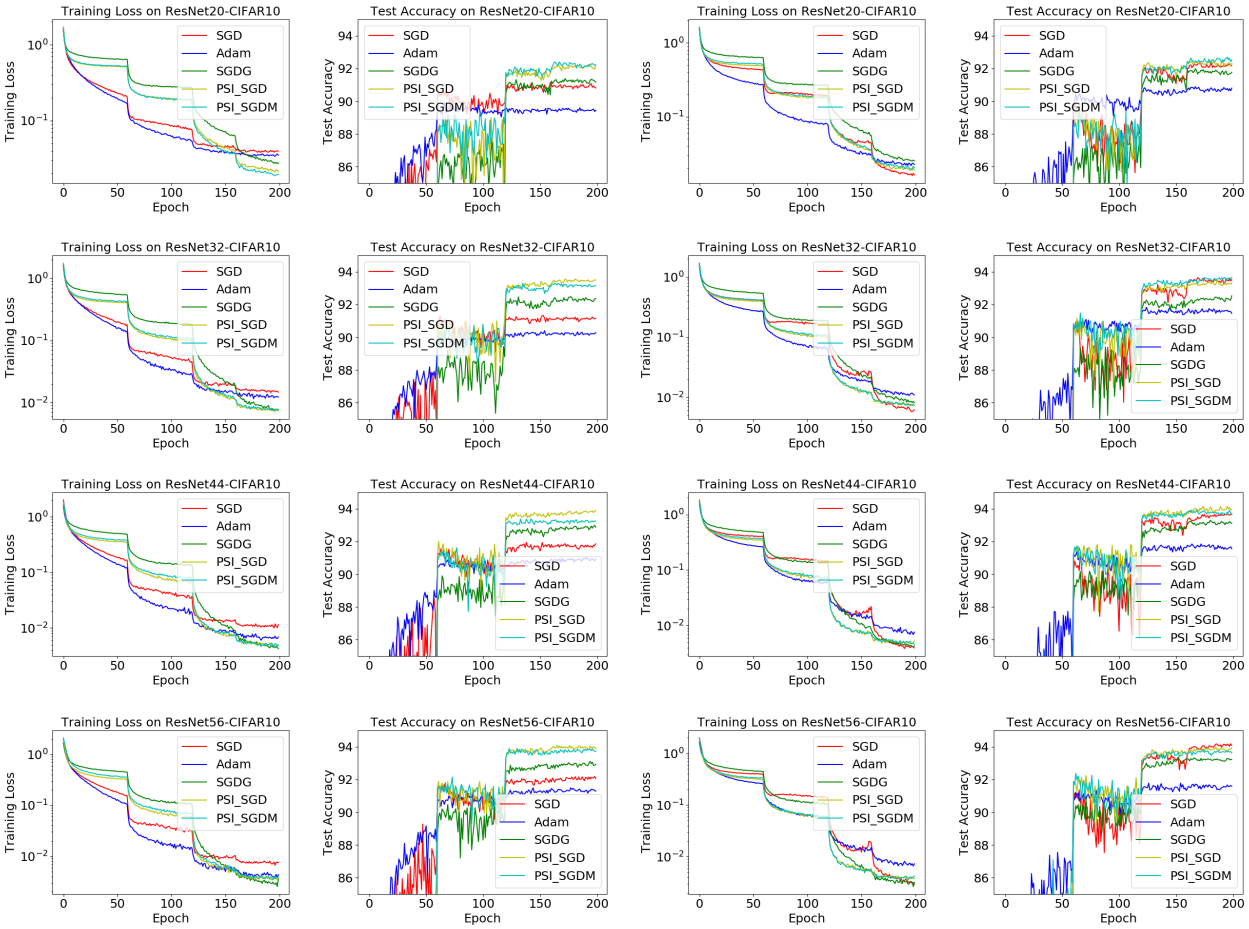

Figure 5: Results of `CIFAR10` on various structures of ResNet i.e. 20, 32, 44, 56. The first and the last two columns are respectively obtained with or without regularizer.

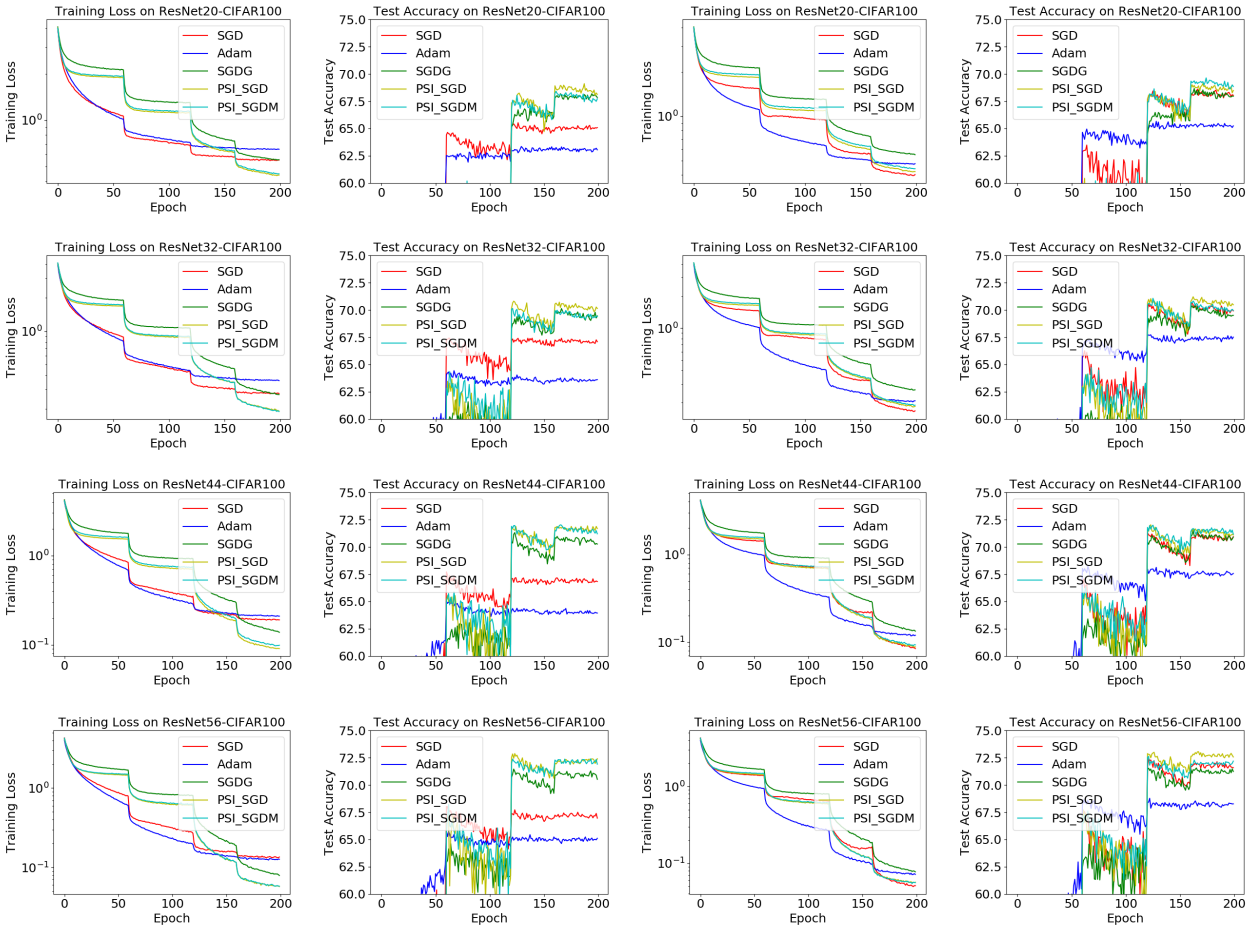

Figure 6: Results of `CIFAR100` on various structures of ResNet i.e. 20, 32, 44, 56. The first and the last two columns are respectively obtained with or without regularizer.