Appendix of ”The Orthogonality of Weight Vectors: the Secret Key to Normalization and Residual Connections”

1 Proof of Theorem 2

Proof. When using the SGD to update parameters,

$$\mathbf{w}^l_{i,t+1} = \mathbf{w}^l_{i,t} - \eta_t \nabla \mathbf{w}^l_{i,t} \tag{1}$$

η_t is the learning rate. Let $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ be any two vectors of the weight W^l_t then

$$\begin{aligned} \mathbf{w}^l_{i,t+1} &= \mathbf{w}^l_{i,t} - \eta_t \nabla \mathbf{w}^l_{i,t} \\ \mathbf{w}^l_{j,t+1} &= \mathbf{w}^l_{j,t} - \eta_t \nabla \mathbf{w}^l_{j,t} \end{aligned} \tag{2}$$

$\mathbf{w}^l_{i,t+1}$ and $\mathbf{w}^l_{j,t+1}$ are the translation transformations of $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ along the $\eta_t \nabla \mathbf{w}^l_{i,t}$ and $\eta_t \nabla \mathbf{w}^l_{j,t}$ directions, in particular, when $\nabla \mathbf{w}^l_{i,t} = \nabla \mathbf{w}^l_{j,t}$, $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ move in the same direction. Therefore, the angle between vectors remains unchanged under translation, and in this case, the translation transformation is a conformal mapping. In the same way, when updating parameters using the Adam,

$$\begin{aligned} \mathbf{w}^l_{i,t+1} &= \mathbf{w}^l_{i,t} - \alpha_t \frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v_{t+1}}} + \epsilon} b_{t+1} \\ \mathbf{m}_{t+1} &= \beta_1 \mathbf{m}_t + (1-\beta_1) \nabla \mathbf{w}^l_t \\ \mathbf{v}_{t+1} &= \beta_2 \mathbf{v}_t + (1-\beta_2) \nabla \mathbf{w}^l_t \odot \nabla \mathbf{w}^l_t \\ b_{t+1} &= \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}} \end{aligned} \tag{3}$$

Where $\odot$ is the multiplication of corresponding elements. For any two vectors $\mathbf{w}^l_{i,t}$ and$\mathbf{w}^l_{j,t}$, $\mathbf{m}_{i,t+1} = \mathbf{m}_{j,t+1}$ and $\mathbf{v}_{i,t+1} = \mathbf{v}_{j,t+1}$ when $\nabla \mathbf{w}^l_{i,t} = \nabla \mathbf{w}^l_{j,t}$, so the Adam is also a conformal mapping.

However, $\nabla \mathbf{w}^l_{i,t} \neq \nabla \mathbf{w}^l_{j,t}$ in general. Let $\kappa = \left\| \nabla \mathbf{w}^l_{i,t} - \nabla \mathbf{w}^l_{j,t} \right\|_2$ be the degree of difference between $\nabla \mathbf{w}^l_{i,t}$ and $\nabla \mathbf{w}^l_{j,t}$. Obviously, $\kappa \geq 0$. As κ increases, the degree of difference between $\nabla \mathbf{w}^l_{i,t}$ and $\nabla \mathbf{w}^l_{j,t}$ becomes larger. As κ decreases, the degree of difference becomes smaller. When κ approaches 0, the optimizer approaches a conformal mapping.

For a fixed κ, the degree of shift difference between parameter vectors $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ is $\eta_t \kappa$ when the optimizer is SGD. If the optimizer is the Adam, The degree of difference in translation transformations between $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ is determined by $\mathbf{m}_{i,t+1}$, $\mathbf{m}_{j,t+1}$, $\mathbf{v}_{i,t+1}$, and $\mathbf{v}_{j,t+1}$. Where $\mathbf{m}_{i,t+1}$ and $\mathbf{m}_{j,t+1}$ represent translation transformations of $\beta_1 \mathbf{m}_{i,t}$ and $\beta_1 \mathbf{m}_{j,t}$, and $\mathbf{v}_{i,t+1}$ and $\mathbf{v}_{j,t+1}$ represent translation transformations of $\beta_2 \mathbf{v}_{i,t}$ and $\beta_2 \mathbf{v}_{j,t}$, and $\left\| \nabla \mathbf{w}^l_{i,t} \odot \nabla \mathbf{w}^l_{i,t} - \nabla \mathbf{w}^l_{i,t} \odot \nabla \mathbf{w}^l_{i,t} \right\|_2 \geq \kappa$. Therefore, the degree of difference in translation vectors between $\mathbf{w}^l_{i,t}$ and $\mathbf{w}^l_{j,t}$ will be significantly greater than $\eta_t \kappa$. The ability of the SGD optimizer for conformal mapping surpasses that of the Adam optimizer.

□

2 Proof of Theorem 3

Proof. For a FNN, the forward propagation process of the neural network is as follows:

$$\mathbf{z}^l = W^l \mathbf{x}^{l-1} + \mathbf{b}^l, \mathbf{x}^l = f\left(\mathbf{z}^l\right) \tag{4}$$

After adding residual connections and normalization (Here, taking batch normalization as an example, the result of layer normalization are similar.),

$$\mathbf{x}^l = BN\left(\mathbf{x}^{l-1} + f\left(\mathbf{z}^l\right)\right) \tag{5}$$

where $\hat{\mathbf{x}}^l = \mathbf{x}^{l-1} + f\left(\mathbf{z}^l\right)$, $\mathbf{x}^l = BN\left(\hat{\mathbf{x}}^l\right)$. Let

$$\mu = \frac{1}{N} \sum_i \hat{x}^l_i, \sigma^2 = \frac{1}{N} \sum_i \left(\hat{x}^l_i - \mu\right)^2. \tag{6}$$

Then, $x^l_i = \frac{\hat{x}^l_i - \mu}{\sigma}$.. Next, we calculate the relevant partial derivatives.

$$\frac{\partial \mu}{\partial \hat{x}^l_j} = \frac{1}{N} \tag{7}$$

$$\begin{aligned} \frac{\partial \sigma}{\partial \hat{x}^l_j} =& \frac{1}{2\sigma N} \sum_i \left[2\left(\hat{x}^l_i - \mu\right)\left(\delta_{ij} - \frac{1}{N}\right)\right] \\ =& \frac{1}{N\sigma} \sum_i \left[\left(\hat{x}^l_i - \mu\right)\delta_{ij}\right] - \frac{1}{N^2\sigma} \sum_i \left(\hat{x}^l_i - \mu\right) \\ =& \frac{\hat{x}^l_j - \mu}{N\sigma} - 0 = \frac{x^l_j}{N} \end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial x_i^l}{\partial \hat{x}_j^l} &= \frac{\left(\delta_{ij} - \frac{\partial \mu}{\partial \hat{x}_j^l}\right)\sigma - \left(\hat{x}_j^l - \mu\right)\frac{\partial \sigma}{\partial \hat{x}_j^l}}{\sigma^2} \\ &= \frac{\left(\delta_{ij}-\frac{1}{N}\right) - x_i^l \frac{\partial\sigma}{\partial \hat{x}_j^l}}{\sigma} = \frac{\left(\delta_{ij}-\frac{1}{N}\right) - \frac{x_i^l x_j^l}{N}}{\sigma} \\ &= \frac{\left(N\delta_{ij}-1\right) - x_i^l x_j^l}{N\sigma}.\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial L}{\partial \hat{x}_j^l} &= \sum_i \frac{\partial L}{\partial x_i^l}\frac{\partial x_i^l}{\partial \hat{x}_j^l} \\ &= \frac{\sum_i \left(\left(x_i^l\right)' \left[\left(N\delta_{ij}-1\right) - x_i^l x_j^l\right]\right)}{N\sigma} \\ &= \frac{N\left(x_j^l\right)' - \sum_i \left(x_i^l\right)' - x_j^l \sum_i \left(\left(x_i^l\right)' x_i^l\right)}{N\sigma} \\ &= \frac{\left(x_j^l\right)'}{\sigma} - \frac{\sum_i \left(x_i^l\right)'}{N\sigma} - \frac{x_j^l \sum_i \left(\left(x_i^l\right)' x_i^l\right)}{N\sigma} \\ &= \frac{1}{\sigma}\left[\left(x_j^l\right)' - \frac{\sum_i \left(x_i^l\right)'}{N} - \frac{x_j^l \sum_i \left(\left(x_i^l\right)' x_i^l\right)}{N}\right]\end{aligned} \tag{10}$$

Let $\mathbf{x}^l = \left\{x_1^l, x_2^l, \cdots, x_N^l\right\}$, then

$$\begin{aligned}\left(\hat{\mathbf{x}}^l\right)' &= \frac{\partial L}{\partial \hat{\mathbf{x}}^l} \\ &= \frac{1}{\sigma}\left[\left(\mathbf{x}^l\right)' - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{1}\right)}{N}\mathbf{1} - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{x}^l\right)}{N}\mathbf{x}^l\right]\end{aligned} \tag{11}$$

where $(\cdot, \cdot)$ is the inner product in N dimensions. Science

$$\begin{gathered}\left\|\vec{1}\right\|^2 = N \\ \left\|\mathbf{x}^l\right\|^2 = \sum_i \left(x_i^l\right)^2 = \sum_i \frac{N\left(\hat{x}_i^l - \mu\right)^2}{\sum_j \left(\hat{x}_j^l - \mu\right)^2} = N\end{gathered} \tag{12}$$

$$\begin{aligned}\left(\hat{\mathbf{x}}^l\right)' &= \frac{\partial L}{\partial \hat{\mathbf{x}}^l} \\ &= \frac{1}{\sigma}\left[\left(\mathbf{x}^l\right)' - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{1}\right)}{(\mathbf{1}, \mathbf{1})}\mathbf{1} - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{x}^l\right)}{\left(\mathbf{x}^l, \mathbf{x}^l\right)}\mathbf{x}^l\right]\end{aligned} \tag{13}$$

For neurons in layer l, $\mathbf{z}^l = W^l\mathbf{x}^{l-1} + \mathbf{b}^l$, then

$$\begin{aligned}\nabla W^l &= \frac{\partial L}{\partial \mathbf{z}^l}\frac{\partial \mathbf{z}^l}{\partial \mathbf{w}^l} = \frac{\partial L}{\partial \hat{\mathbf{x}}^l}\frac{\partial \hat{\mathbf{x}}^l}{\partial \mathbf{z}^l}\frac{\partial \mathbf{z}^l}{\partial \mathbf{w}^l} \\ &= \frac{1}{\sigma}f'\left(\mathbf{z}^l\right) \odot \left[\left(\mathbf{x}^l\right)' - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{1}\right)}{(\mathbf{1}, \mathbf{1})}\mathbf{1} - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{x}^l\right)}{\left(\mathbf{x}^l, \mathbf{x}^l\right)}\mathbf{x}^l\right]\left(\mathbf{x}^{l-1}\right)^T\end{aligned} \tag{14}$$

$\odot$ is the element-wise multiplication. $\left(\mathbf{x}^l\right)' - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{1}\right)}{(\mathbf{1}, \mathbf{1})}\mathbf{1} - \frac{\left(\left(\mathbf{x}^l\right)', \mathbf{x}^l\right)}{\left(\mathbf{x}^l, \mathbf{x}^l\right)}\mathbf{x}^l$ is the projection transformation of vector $\left(\mathbf{x}^l\right)'$ in the directions of vectors $\mathbf{1}$ and $\mathbf{x}^l$, abbreviated as $P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l\perp}}\left(\mathbf{x}^l\right)'$. Let $\delta_{NR}^l = f'\left(\mathbf{z}^l\right) \odot \left(W^{l+1}\delta^{l+1}\right)$, NR represents the addition of normalization and residual connections.

When the batch normalization and residual connections are also introduced to the neurons in the $l+1$ layer,

$$\hat{\mathbf{x}}^{l+1} = \mathbf{x}^l + f\left(W^{l+1}\mathbf{x}^l + b^{l+1}\right), \mathbf{x}^{l+1} = BN\left(\hat{\mathbf{x}}^{l+1}\right) \tag{15}$$

For the equation (14),

$$\begin{aligned}&\left(\mathbf{x}^l\right)' = \frac{\partial L}{\partial \mathbf{x}^l} = \frac{\partial \hat{\mathbf{x}}^{l+1}}{\partial \mathbf{x}^l}\frac{\partial L}{\partial \hat{\mathbf{x}}^{l+1}} \\ &= \frac{1}{\sigma}\left(\mathbf{1} + f'\left(\mathbf{z}^{l+1}\right)\right) \\ &\odot W^{l+1}\left[\left(\mathbf{x}^{l+1}\right)' - \frac{\left(\left(\mathbf{x}^{l+1}\right)', \mathbf{1}\right)}{(\mathbf{1}, \mathbf{1})}\mathbf{1} - \frac{\left(\left(\mathbf{x}^{l+1}\right)', \mathbf{x}^{l+1}\right)}{\left(\mathbf{x}^{l+1}, \mathbf{x}^{l+1}\right)}\mathbf{x}^{l+1}\right] \\ &= \frac{1}{\sigma}\left(\mathbf{1} + f'\left(\mathbf{z}^{l+1}\right)\right) \odot W^{l+1}P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l+1\perp}}\left(\mathbf{x}^{l+1}\right)'\end{aligned} \tag{16}$$

So,

$$\begin{aligned}&\nabla W^l = \frac{1}{\sigma^l}f'\left(\mathbf{z}^l\right) \odot P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l\perp}} \\ &\left[\frac{1}{\sigma^{l+1}}\left(\mathbf{1} + f'\left(\mathbf{z}^{l+1}\right)\right) \odot W^{l+1}P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l+1\perp}}\left(\mathbf{x}^{l+1}\right)'\right]\left(\mathbf{x}^{l-1}\right)^T\end{aligned} \tag{17}$$

At the same time,

$$\begin{aligned}&\delta_{NR}^l = \frac{1}{\sigma^l}f'\left(\mathbf{z}^l\right) \odot P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l\perp}} \\ &\left[\frac{1}{\sigma^{l+1}}\left(\mathbf{1} + f'\left(\mathbf{z}^{l+1}\right)\right) \odot W^{l+1}P_{\mathbf{1}^\perp}P_{\mathbf{x}^{l+1\perp}}\left(\mathbf{x}^{l+1}\right)'\right]\end{aligned} \tag{18}$$

If the neural network does not incorporate the normalization and residual connections,

$$\begin{aligned}\mathbf{x}^l = \hat{\mathbf{x}}^l &= f\left(\mathbf{z}^l\right) = f\left(W^l\mathbf{x}^{l-1} + \mathbf{b}^l\right) \\ \nabla W^l &= \delta^l\left(\mathbf{x}^{l-1}\right)^T \\ \delta^l &= f'\left(\mathbf{z}^l\right) \odot \left(W^{l+1}\delta^{l+1}\right)\end{aligned} \tag{19}$$

If $\nabla\mathbf{w}_i^l$ and $\nabla\mathbf{w}_j^l$ are any two row vectors in ∇W^l, so without adding normalization and residual connections,

$$\begin{aligned}\nabla\mathbf{w}_i^{c,l} &= x_i^{l-1}\delta^l \\ \nabla\mathbf{w}_j^{c,l} &= x_j^{l-1}\delta^l\end{aligned} \tag{20}$$

After adding normalization and residual connections,

$$\begin{aligned}\nabla\mathbf{w}_{NR,i}^{c,l} &= x_{NR,i}^{l-1}\delta_{NR}^l \\ \nabla\mathbf{w}_{NR,j}^{c,l} &= x_{NR,j}^{l-1}\delta_{NR}^l\end{aligned} \tag{21}$$

Normalization makes $x_{NR,i}^{l-1}$ and $x_{NR,j}^{l-1}$ belong to a distribution with an expected value of 0 and a variance of 1. Compared to x_i^{l-1} and x_j^{l-1}, $x_{NR,i}^{l-1}$ and $x_{NR,j}^{l-1}$ are closer, resulting in smaller differences between $\nabla \mathbf{w}_{NR,i}^{c,l}$ and $\nabla \mathbf{w}_{NR,j}^{c,l}$. This makes the parameter updates closer to translational transformations, thereby improving the optimizer ability for conformal mappings.In the same way, if $\nabla \mathbf{w}_i^l$ and $\nabla \mathbf{w}_j^l$ are any two column vectors in ∇W^l, then

$$\begin{aligned} \nabla \mathbf{w}_{NR,i}^{r,l} &= \delta_{NR,i}^{l} \left(\mathbf{x}_{NR}^{l-1}\right)^T \\ \nabla \mathbf{w}_{NR,j}^{r,l} &= \delta_{NR,j}^{l} \left(\mathbf{x}_{NR}^{l-1}\right)^T \end{aligned} \tag{22}$$

$$\begin{aligned} \nabla \mathbf{w}_{i}^{r,l} &= \delta_{i}^{l} \left(\mathbf{x}^{l-1}\right)^T \\ \nabla \mathbf{w}_{j}^{r,l} &= \delta_{j}^{l} \left(\mathbf{x}^{l-1}\right)^T \end{aligned} \tag{23}$$

Formula (18) indicates that, under the influence of vectors $\mathbf{1}$ and variance σ, the values of $\delta_{NR,i}^{l}$ are closer to $\delta_{NR,j}^{l}$ compared to δ_i^l and δ_j^l. Therefore, the optimizer exhibits better conformal capabilities. In conclusion, normalization and residual links generally enhance the optimizer conformal mapping ability □

3 The explanation for LoRA maintaining orthogonality of parameter vectors.

For a pre-trained weight matrix $W_0 \in R^{d \times k}$, the core improvement of LoRA is that constrain the update of W_0 by representing the latter with a low-rank decomposition

$$W_0 + \varDelta W = W_0 + BA \tag{24}$$

where $B \in R^{d \times r}$, $A \in R^{r \times k}$, and the rank $r \ll \min{(d,k)}$. For any two vectors in the parameter matrix W_0, the essence of Equation (24) remains a translation transformation of vectors. In Equation (24), $\varDelta W$ is decomposed into the product of two low-rank matrices, $B$and A, yet matrices B and A are still computed from the gradients of parameters. The normalization and residual links reduce the differences between gradient vectors, hence diminishing the distinctions between any vectors within matrices B and A. Consequently, the results of Equation (24) for any two vectors in W_0 tend to undergo translation transformations along the same direction, ensuring that the orthogonality between parameter vectors is not significantly altered in the network fine-tuning with LoRA.

4 Notations Introduced Previously and Some Additional Experiments

Notation	Type	Definition
l	Integer	Number of network layers
N	Integer	Batch size
M_l	Integer	Number of neurons in layer l
k	Integer	Output dimension
W^l	$R^{M_l \times M_{l-1}}$	Weight from layer $l-1$ to l
$\mathbf{b}^l$	R^{M_l}	Bias from layer $l-1$ to l
$\mathbf{x}^l$	R^{M_l}	Neurons of layer l
$\mathbf{z}^l$	R^{M_l}	Neurons of layer l not activated by the activation funtion
I_k	$R^{k \times k}$	Identity matrix of size k
$\mathbf{1}_k$	R^n	All-ones vector

Table 1: Notations introduced previously.

Optimizers	PII Layer_1↑	PII Layer_2↓	PII Layer_3↓
SGD	1.0/0.858	0.0.070/0.121	0.050/0.062
Adam	1.0/0.333	0.070/0.324	0.050/0.152
AdamW	1.0/0.352	0.070/0.311	0.050/0.166
RMSprop	1.0/0.384	0.070/0.309	0.050/0.154

Table 2: Conformal capabilities of optimizers in CNNs. The datasets is CIFAR-100. Similar to the results on CIFAR-10, the conformal mapping capability of the SGD optimizer is the best, while the conformal mapping capabilities of Adam and AdamW are comparable

Functions/	PII	PII ↓	ACC %	ACC % ↑
Optimizers	Original	Norm_Res	Original	Norm_Res
GELU/Adam	0.212	0.154	39.38	49.43
GELU/SGD	0.194	0.166	51.59	53.34

Table 3: The influence of the combination of normalization and residual connections on the orthogonality of parameter vectors and testing accuracy in CNNs. The datasets is CIFAR-100.For Adam and SGD, normalization and residual connections can significantly reduce PII.Therefore, it enhances the orthogonality of parameter vectors, simultaneously improving the network's capability.

Optimizers	Hidden layer_1	Hidden layer_2	Hidden layer_3
Original	0.131	0.148	0.153
Normalization	0.100	0.107	0.128
Residual	0.111	0.137	0.147
Norm_Residual	0.079	0.090	0.107

Table 4: Changes in PII after separately adding normalization and residual connections."Original" represents the results without the addition of normalization and residual connections; "Normalization" represents the results with only normalization added; 'Residual' represents the results with only residual connections added, and "Norm_Residual" represents the results with both normalization and residual connections added simultaneously. Experimental results indicate that adding only normalization or residual connections can also reduce PII, but the most significant reduction occurs when both normalization and residual connections are added simultaneously, resulting in the lowest PII.