## A Analysis of W in MoSLoRA

### A.1 Vanilla MoSLoRA

For an arbitrary input $x$, we have:

$$y = x\mathbf{W}_{merge}; \mathbf{W}_{merge} = \mathbf{W}_0 + \mathbf{AWB}, \tag{2}$$

where the $\mathbf{W}_0$ is frozen during training. Then we have:

$$\frac{\partial y}{\partial \mathbf{A}} = \frac{\partial y}{\partial \mathbf{W}_{merge}}\mathbf{B}^T\mathbf{W}^T; \frac{\partial y}{\partial \mathbf{W}} = \mathbf{A}^T\frac{\partial y}{\partial \mathbf{W}_{merge}}\mathbf{B}^T; \frac{\partial y}{\partial \mathbf{B}} = \mathbf{W}^T\mathbf{A}^T\frac{\partial y}{\partial \mathbf{W}_{merge}} \tag{3}$$

Denote the learning rate as $\eta$, the updating process is:

$$\mathbf{A} \leftarrow \mathbf{A} - \eta\frac{\partial y}{\partial \mathbf{A}} = \mathbf{A} - \eta\frac{\partial y}{\partial \mathbf{W}_{merge}}\mathbf{B}^T\mathbf{W}^T \tag{4}$$

The process is similar for $\mathbf{W}$ and $\mathbf{B}$. Let $\Delta = \frac{\partial y}{\partial \mathbf{W}_{merge}}$. Thus, the weight of the updated LoRA branch would be:

$$\begin{aligned} \mathbf{W}_{LoRA} &= (\mathbf{A} - \eta\Delta\mathbf{B}^T\mathbf{W}^T)(\mathbf{W} - \eta\mathbf{A}^T\Delta\mathbf{B}^T)(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \\ &= (\mathbf{AW} - \eta\mathbf{AA}^T\Delta\mathbf{B}^T - \eta\Delta\mathbf{B}^T\mathbf{W}^T\mathbf{W} + \eta^2\Delta\mathbf{B}^T\mathbf{W}^T\mathbf{A}^T\Delta\mathbf{B}^T)(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \end{aligned} \tag{5}$$

### A.2 Merge A and W

Denote $\hat{\mathbf{A}} = \mathbf{AW}$. It means that we initialize $\hat{\mathbf{A}}$ as the same as $\mathbf{AW}$. The output is the same:

$$y = x\mathbf{W}_{merge}; \mathbf{W}_{merge} = \mathbf{W}_0 + \hat{\mathbf{A}}\mathbf{B} = \mathbf{W}_0 + \mathbf{AWB}. \tag{6}$$

However, the corresponding gradients would be:

$$\frac{\partial y}{\partial \hat{\mathbf{A}}} = \frac{\partial y}{\partial \mathbf{W}_{merge}}\mathbf{B}^T = \Delta\mathbf{B}^T; \frac{\partial y}{\partial \mathbf{B}} = \hat{\mathbf{A}}^T\frac{\partial y}{\partial \mathbf{W}_{merge}} = \hat{\mathbf{A}}^T\Delta \tag{7}$$

Based on that, we can get the updated LoRA after updating the parameters:

$$\begin{aligned} \hat{\mathbf{W}}_{LoRA} &= (\hat{\mathbf{A}} - \eta\Delta\mathbf{B}^T)(\mathbf{B} - \eta\hat{\mathbf{A}}^T\Delta) \\ &= (\mathbf{AW} - \eta\Delta\mathbf{B}^T)(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \end{aligned} \tag{8}$$

### A.3 Comparison

Comparing Equation 5 and 8, we can conclude that the updated weights are not the same, since

$$\begin{aligned} \hat{\mathbf{W}}_{LoRA} - \mathbf{W}_{LoRA} &= (-\eta\Delta\mathbf{B}^T + \eta\mathbf{AA}^T\Delta\mathbf{B}^T + \eta\Delta\mathbf{B}^T\mathbf{W}^T\mathbf{W} - \eta^2\Delta\mathbf{B}^T\mathbf{W}^T\mathbf{A}^T\Delta\mathbf{B}^T)(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \\ &= (\eta(\mathbf{A} - \eta\Delta\mathbf{B}^T\mathbf{W}^T)\mathbf{A}^T\Delta\mathbf{B}^T + \eta\Delta\mathbf{B}^T(\mathbf{W}^T\mathbf{W} - \mathbf{I}))(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \neq \mathbf{0}. \end{aligned} \tag{9}$$

### A.4 Fix W as Orthogonal Matrix

If we fix $\mathbf{W}$ as **orthogonal matrix and do not update** (i.e., $\mathbf{WW}^T = \mathbf{I}$), the updated LoRA would be:

$$\begin{aligned} \mathbf{W}^{\mathbf{I}}_{LoRA} &= (\mathbf{A} - \eta\Delta\mathbf{B}^T\mathbf{W}^T)\mathbf{W}(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \\ &= (\mathbf{AW} - \eta\Delta\mathbf{B}^T\mathbf{W}^T\mathbf{W})(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) \\ &= (\mathbf{AW} - \eta\Delta\mathbf{B}^T)(\mathbf{B} - \eta\mathbf{W}^T\mathbf{A}^T\Delta) = \hat{\mathbf{W}}_{LoRA} \end{aligned} \tag{10}$$

### A.5 Conclusion

*Though mathematically equivalent initialized, the optimization process would be different if $\mathbf{W}$ is learnable. Specifically, the optimization process would be the same i.i.f $\mathbf{W}$ is a fixed orthogonal matrix.*