# AFL: A Single-Round Analytic Approach for Federated Learning with Pre-trained Models

## Supplementary Material

## A. Proof of Lemma 1

*Proof.* We prove this Lemma mainly based on the existing MP inverse partition result [4] as follows.

In [4], it has been demonstrated that the MP inverse of any matrix, $\boldsymbol{\mathcal{A}} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \end{bmatrix}$ can be written as

$$\boldsymbol{\mathcal{A}}^\dagger = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \end{bmatrix}^\dagger = \begin{bmatrix} \boldsymbol{U}^\dagger - \boldsymbol{U}^\dagger \boldsymbol{V}\boldsymbol{C}^\dagger - \boldsymbol{U}^\dagger \boldsymbol{V}(\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{K}\boldsymbol{V}^{*}\boldsymbol{U}^{\dagger *}\boldsymbol{U}^\dagger(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{C}^\dagger) \\ \boldsymbol{V}^\dagger - \boldsymbol{V}^\dagger \boldsymbol{U}\tilde{\boldsymbol{C}}^\dagger - \boldsymbol{V}^\dagger \boldsymbol{U}(\boldsymbol{I} - \tilde{\boldsymbol{C}}^\dagger \tilde{\boldsymbol{C}})\tilde{\boldsymbol{K}}\boldsymbol{U}^{*}\boldsymbol{V}^{\dagger *}\boldsymbol{V}^\dagger(\boldsymbol{I} - \boldsymbol{U}\tilde{\boldsymbol{C}}^\dagger) \end{bmatrix}, \tag{A.1}$$

where

$$\begin{cases} \boldsymbol{C} = (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\dagger)\boldsymbol{V} \\ \tilde{\boldsymbol{C}} = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^\dagger)\boldsymbol{U} \end{cases},$$

$$\begin{cases} \boldsymbol{K} = \left[\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{V}^{*}\boldsymbol{U}^{\dagger *}\boldsymbol{U}^\dagger \boldsymbol{V}(\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\right]^{-1} \\ \tilde{\boldsymbol{K}} = \left[\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{U}^{*}\boldsymbol{V}^{\dagger *}\boldsymbol{V}^\dagger \boldsymbol{U}(\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\right]^{-1} \end{cases}. \tag{A.2}$$

In the case of $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_v \end{bmatrix}$ with only real numbers, we substitute $\boldsymbol{U}$ with $\boldsymbol{X}_u^\top$, $\boldsymbol{V}$ with $\boldsymbol{X}_v^\top$. This rewrites (A.1), (A.2) into

$$\boldsymbol{X}^\dagger = \begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_v \end{bmatrix}^\dagger = \begin{bmatrix} \boldsymbol{X}_u^{\dagger\top} - \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{C}^\dagger \\ \boldsymbol{X}_v^{\dagger\top} - \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \tilde{\boldsymbol{C}}^\dagger \end{bmatrix}^\top - \begin{bmatrix} \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{K}\boldsymbol{X}_v \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top}(\boldsymbol{I} - \boldsymbol{X}_v^\top \boldsymbol{C}^\dagger) \\ \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top (\boldsymbol{I} - \tilde{\boldsymbol{C}}^\dagger \tilde{\boldsymbol{C}})\tilde{\boldsymbol{K}}\boldsymbol{X}_u \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top}(\boldsymbol{I} - \boldsymbol{X}_u^\top \tilde{\boldsymbol{C}}^\dagger) \end{bmatrix}^\top, \tag{A.3}$$

where

$$\begin{cases} \boldsymbol{C} = (\boldsymbol{I} - \boldsymbol{X}_u^\top \boldsymbol{X}_u^{\dagger\top})\boldsymbol{X}_v^\top \\ \tilde{\boldsymbol{C}} = (\boldsymbol{I} - \boldsymbol{X}_v^\top \boldsymbol{X}_v^{\dagger\top})\boldsymbol{X}_u^\top \end{cases},$$

$$\begin{cases} \boldsymbol{K} = \left[\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{X}_v \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\right]^{-1} \\ \tilde{\boldsymbol{K}} = \left[\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\boldsymbol{X}_u \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top (\boldsymbol{I} - \boldsymbol{C}^\dagger \boldsymbol{C})\right]^{-1} \end{cases}. \tag{A.4}$$

As $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ are of full column ranks, we obtain an alternative formulation of the MP inverse, i.e.,

$$\boldsymbol{X}_u^\dagger = (\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1}\boldsymbol{X}_u^\top, \quad \boldsymbol{X}_v^\dagger = (\boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1}\boldsymbol{X}_v^\top. \tag{A.5}$$

Hence we have

$$\begin{aligned} \boldsymbol{C} &= (\boldsymbol{I} - \boldsymbol{X}_u^\top \boldsymbol{X}_u^{\dagger\top})\boldsymbol{X}_v^\top \\ &= (\boldsymbol{I} - \boldsymbol{X}_u^\top \boldsymbol{X}_u (\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1})\boldsymbol{X}_v^\top = \boldsymbol{0}. \end{aligned} \tag{A.6}$$

Similarly,

$$\begin{aligned} \tilde{\boldsymbol{C}} &= (\boldsymbol{I} - \boldsymbol{X}_v^\top \boldsymbol{X}_v^{\dagger\top})\boldsymbol{X}_u^\top \\ &= (\boldsymbol{I} - \boldsymbol{X}_v^\top \boldsymbol{X}_v (\boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1})\boldsymbol{X}_u^\top = \boldsymbol{0}. \end{aligned} \tag{A.7}$$

This simplifies $\boldsymbol{K}$ and $\tilde{\boldsymbol{K}}$ as

$$\begin{cases} \boldsymbol{K} = (\boldsymbol{I} + \boldsymbol{X}_v \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top)^{-1} \\ \tilde{\boldsymbol{K}} = (\boldsymbol{I} + \boldsymbol{X}_u \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top)^{-1} \end{cases}. \tag{A.8}$$

According to the Woodbury Matrix Identity, i.e., for conformable matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\boldsymbol{E} \in \mathbb{R}^{m\times m}$, and $\boldsymbol{D} \in \mathbb{R}^{m\times n}$,

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{E}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{E}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}, \tag{A.9}$$

we expand $\boldsymbol{K}$ by substituting $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{B} = \boldsymbol{X}_v$, $\boldsymbol{E} = \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top}$, and $\boldsymbol{D} = \boldsymbol{X}_v^\top$, leading to

$$\boldsymbol{K} = \boldsymbol{I} - \boldsymbol{X}_v(\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1}\boldsymbol{X}_v^\top. \tag{A.10}$$

Similarly,

$$\tilde{\boldsymbol{K}} = \boldsymbol{I} - \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1}\boldsymbol{X}_u^\top. \tag{A.11}$$

Thus,

$$\boldsymbol{X}^\dagger = \begin{bmatrix} \boldsymbol{X}_u^{\dagger\top} - \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{K}\boldsymbol{X}_v \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \\ \boldsymbol{X}_v^{\dagger\top} - \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \tilde{\boldsymbol{K}}\boldsymbol{X}_u \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \end{bmatrix}^\top. \tag{A.12}$$

Let $\boldsymbol{X}^\dagger = \begin{bmatrix} \bar{\boldsymbol{U}} & \bar{\boldsymbol{V}} \end{bmatrix}$, we have

$$\begin{aligned} \bar{\boldsymbol{U}} &= \left(\boldsymbol{X}_u^{\dagger\top} - \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{K}\boldsymbol{X}_v \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top}\right)^\top \\ &= \boldsymbol{X}_u^\dagger - \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{K}^\top \boldsymbol{X}_v \boldsymbol{X}_u^\dagger \\ \bar{\boldsymbol{V}} &= \left(\boldsymbol{X}_v^{\dagger\top} - \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \tilde{\boldsymbol{K}}\boldsymbol{X}_u \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top}\right)^\top \\ &= \boldsymbol{X}_v^\dagger - \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \tilde{\boldsymbol{K}}^\top \boldsymbol{X}_u \boldsymbol{X}_v^\dagger \end{aligned} \tag{A.13}$$

Substitute $\boldsymbol{K}$ and $\tilde{\boldsymbol{K}}$ with (A.10) and (A.11), we may rewrite (A.13) into

$$\begin{aligned} \bar{\boldsymbol{U}} &= \boldsymbol{X}_u^\dagger - \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \left(\boldsymbol{I} - \boldsymbol{X}_v(\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1}\boldsymbol{X}_v^\top\right)^\top \boldsymbol{X}_v \boldsymbol{X}_u, \\ \bar{\boldsymbol{V}} &= \boldsymbol{X}_v^\dagger - \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \left(\boldsymbol{I} - \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1}\boldsymbol{X}_u^\top\right)^\top \boldsymbol{X}_u \boldsymbol{X}_v^\dagger. \end{aligned} \tag{A.14}$$

That is,

$$\begin{aligned} \bar{\boldsymbol{U}} =& \boldsymbol{X}_u^\dagger - \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{X}_v \boldsymbol{X}_u^\dagger \\ & - \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \boldsymbol{X}_v^\top \boldsymbol{X}_v (\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1} \boldsymbol{X}_v^\top \boldsymbol{X}_v \boldsymbol{X}_u^\dagger \\ \bar{\boldsymbol{V}} =& \boldsymbol{X}_v^\dagger - \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \boldsymbol{X}_u \boldsymbol{X}_v^\dagger \\ & - \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \boldsymbol{X}_u^\top \boldsymbol{X}_u (\boldsymbol{X}_u^\top \boldsymbol{X}_u + \boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1} \boldsymbol{X}_u^\top \boldsymbol{X}_u \boldsymbol{X}_v^\dagger. \end{aligned} \tag{A.15}$$

Let

$$\begin{cases} \boldsymbol{C}_u = \boldsymbol{X}_u^\top \boldsymbol{X}_u \\ \boldsymbol{C}_v = \boldsymbol{X}_v^\top \boldsymbol{X}_v \end{cases}, \text{and} \begin{cases} \boldsymbol{R}_u = \boldsymbol{R}_u^{-1} \\ \boldsymbol{R}_v = \boldsymbol{R}_v^{-1} \end{cases}. \tag{A.16}$$

we have

$$\begin{cases} \bar{\boldsymbol{U}} = \left[\boldsymbol{I} - \boldsymbol{R}_u\boldsymbol{C}_v - \boldsymbol{R}_u\boldsymbol{C}_v(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_v\right]\boldsymbol{X}_u^\dagger \\ \bar{\boldsymbol{V}} = \left[\boldsymbol{I} - \boldsymbol{R}_v\boldsymbol{C}_u - \boldsymbol{R}_v\boldsymbol{C}_u(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_u\right]\boldsymbol{X}_v^\dagger \end{cases}. \tag{A.17}$$

Thus,

$$\boldsymbol{X}^\dagger = \begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_v \end{bmatrix}^\dagger = \begin{bmatrix} \bar{\boldsymbol{U}} & \bar{\boldsymbol{V}} \end{bmatrix}, \tag{A.18}$$

which completes the proof. □

## B. Proof of Theorem 1

*Proof.* As indicated in Lemma 1, we have

$$\boldsymbol{X}^\dagger = \begin{bmatrix} \bar{\boldsymbol{U}} & \bar{\boldsymbol{V}} \end{bmatrix} \tag{A.19}$$

where

$$\begin{cases} \bar{\boldsymbol{U}} = \left[\boldsymbol{I} - \boldsymbol{R}_u\boldsymbol{C}_v - \boldsymbol{R}_u\boldsymbol{C}_v(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_v\right]\boldsymbol{X}_u^\dagger \\ \bar{\boldsymbol{V}} = \left[\boldsymbol{I} - \boldsymbol{R}_v\boldsymbol{C}_u - \boldsymbol{R}_v\boldsymbol{C}_u(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_u\right]\boldsymbol{X}_v^\dagger \end{cases}, \tag{A.20}$$

and

$$\begin{cases} \boldsymbol{R}_u = (\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1} = \boldsymbol{X}_u^\dagger \boldsymbol{X}_u^{\dagger\top} \\ \boldsymbol{R}_v = (\boldsymbol{X}_v^\top \boldsymbol{X}_v)^{-1} = \boldsymbol{X}_v^\dagger \boldsymbol{X}_v^{\dagger\top} \end{cases}$$

$$\begin{cases} \boldsymbol{C}_u = \boldsymbol{R}_u^{-1} = \boldsymbol{X}_u^\top \boldsymbol{X}_u \\ \boldsymbol{C}_v = \boldsymbol{R}_v^{-1} = \boldsymbol{X}_v^\top \boldsymbol{X}_v \end{cases}. \tag{A.21}$$

Hence,

$$\begin{aligned} \boldsymbol{W} = \boldsymbol{X}^\dagger \boldsymbol{Y} &= \begin{bmatrix} \bar{\boldsymbol{U}} & \bar{\boldsymbol{V}} \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}_u \\ \boldsymbol{Y}_v \end{bmatrix} \\ &= \bar{\boldsymbol{U}}\boldsymbol{Y}_u + \bar{\boldsymbol{V}}\boldsymbol{Y}_v. \end{aligned} \tag{A.22}$$

By substituting $\bar{\boldsymbol{U}}$ and $\bar{\boldsymbol{V}}$ with those in (A.20), we rewrite (A.22) into

$$\begin{aligned} \hat{\boldsymbol{W}} = &\left[\boldsymbol{I} - \boldsymbol{R}_u\boldsymbol{C}_v - \boldsymbol{R}_u\boldsymbol{C}_v(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_v\right]\boldsymbol{X}_u^\dagger\boldsymbol{Y}_u \\ &+ \left[\boldsymbol{I} - \boldsymbol{R}_v\boldsymbol{C}_u - \boldsymbol{R}_v\boldsymbol{C}_u(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_u\right]\boldsymbol{X}_v^\dagger\boldsymbol{Y}_v. \end{aligned} \tag{A.23}$$

As $\hat{\boldsymbol{W}}_u = \boldsymbol{X}_u^\dagger \boldsymbol{Y}_u$ and $\hat{\boldsymbol{W}}_v = \boldsymbol{X}_v^\dagger \boldsymbol{Y}_v$, (A.23) can be rewritten as

$$\begin{aligned} \hat{\boldsymbol{W}} = &\left[\boldsymbol{I} - \boldsymbol{R}_u\boldsymbol{C}_v - \boldsymbol{R}_u\boldsymbol{C}_v(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_v\right]\hat{\boldsymbol{W}}_u \\ &+ \left[\boldsymbol{I} - \boldsymbol{R}_v\boldsymbol{C}_u - \boldsymbol{R}_v\boldsymbol{C}_u(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_u\right]\hat{\boldsymbol{W}}_v. \end{aligned} \tag{A.24}$$

That is,

$$\hat{\boldsymbol{W}} = \boldsymbol{\mathcal{W}}_u\hat{\boldsymbol{W}}_u + \boldsymbol{\mathcal{W}}_v\hat{\boldsymbol{W}}_v, \tag{A.25}$$

where

$$\begin{cases} \boldsymbol{\mathcal{W}}_u = \boldsymbol{I} - \boldsymbol{R}_u\boldsymbol{C}_v - \boldsymbol{R}_u\boldsymbol{C}_v(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_v \\ \boldsymbol{\mathcal{W}}_v = \boldsymbol{I} - \boldsymbol{R}_v\boldsymbol{C}_u - \boldsymbol{R}_v\boldsymbol{C}_u(\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}\boldsymbol{C}_u \end{cases},$$

$$\begin{cases} \boldsymbol{C}_u = \boldsymbol{X}_u^\top \boldsymbol{X}_u \\ \boldsymbol{C}_v = \boldsymbol{X}_v^\top \boldsymbol{X}_v \end{cases} \text{and} \begin{cases} \boldsymbol{R}_u = \boldsymbol{C}_u^{-1} \\ \boldsymbol{R}_v = \boldsymbol{C}_v^{-1} \end{cases}. \tag{A.26}$$

□

## C. Proof of Theorem 2

*Proof.* First we consider the aggregation of two clients. Directly substituting $\hat{\boldsymbol{W}}_u$ as $\hat{\boldsymbol{W}}_u^{\mathrm{r}}$ and changing $\boldsymbol{C}_u$, $\boldsymbol{C}_v$ to $\boldsymbol{C}_u^{\mathrm{r}} = (\boldsymbol{X}_u^\top \boldsymbol{X}_u + \gamma\boldsymbol{I})$, $\boldsymbol{C}_v^{\mathrm{r}} = (\boldsymbol{X}_v^\top \boldsymbol{X}_v + \gamma\boldsymbol{I})$ in Theorem B, we have

$$\hat{\boldsymbol{W}}^{\mathrm{r}} = \boldsymbol{\mathcal{W}}_u^{\mathrm{r}}\hat{\boldsymbol{W}}_u^{\mathrm{r}} + \boldsymbol{\mathcal{W}}_v^{\mathrm{r}}\hat{\boldsymbol{W}}_v^{\mathrm{r}}, \tag{A.27}$$

where

$$\begin{cases} \boldsymbol{\mathcal{W}}_u^{\mathrm{r}} = \boldsymbol{I} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}(\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}\boldsymbol{C}_v^{\mathrm{r}} \\ \boldsymbol{\mathcal{W}}_v^{\mathrm{r}} = \boldsymbol{I} - \boldsymbol{R}_v^{\mathrm{r}}\boldsymbol{C}_u^{\mathrm{r}} - \boldsymbol{R}_v^{\mathrm{r}}\boldsymbol{C}_u^{\mathrm{r}}(\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}\boldsymbol{C}_u^{\mathrm{r}} \end{cases},$$

$$\begin{cases} \boldsymbol{C}_u^{\mathrm{r}} = (\boldsymbol{X}_u^\top \boldsymbol{X}_u + \gamma\boldsymbol{I}) \\ \boldsymbol{C}_v^{\mathrm{r}} = (\boldsymbol{X}_v^\top \boldsymbol{X}_v + \gamma\boldsymbol{I}) \end{cases} \text{and} \begin{cases} \boldsymbol{R}_u^{\mathrm{r}} = \boldsymbol{C}_u^{\mathrm{r}-1} \\ \boldsymbol{R}_v^{\mathrm{r}} = \boldsymbol{C}_v^{\mathrm{r}-1} \end{cases}. \tag{A.28}$$

Since $\hat{\boldsymbol{W}}_u^{\mathrm{r}} = (\boldsymbol{X}_u^\top \boldsymbol{X}_u + \gamma\boldsymbol{I})^{-1}\boldsymbol{X}_u^\top \boldsymbol{Y}_u$, then

$$\begin{aligned} \boldsymbol{\mathcal{W}}_u^{\mathrm{r}}\hat{\boldsymbol{W}}_u^{\mathrm{r}} &= [\boldsymbol{I} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}(\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}\boldsymbol{C}_v^{\mathrm{r}}]\boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{X}_u^\top \boldsymbol{Y}_u \\ &= [\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}(\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}\boldsymbol{C}_v^{\mathrm{r}}\boldsymbol{R}_u^{\mathrm{r}}]\boldsymbol{X}_u^\top \boldsymbol{Y}_u. \end{aligned} \tag{A.29}$$

According to the Woodbury Matrix Identity in (A.9), let $\boldsymbol{B} = \boldsymbol{I}, \boldsymbol{D} = \boldsymbol{I}$, we have

$$(\boldsymbol{A} + \boldsymbol{E})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1} + \boldsymbol{E}^{-1})^{-1}\boldsymbol{A}^{-1}. \tag{A.30}$$

Then we have

$$\boldsymbol{A}^{-1}(\boldsymbol{A}^{-1} + \boldsymbol{E}^{-1})^{-1}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} - (\boldsymbol{A} + \boldsymbol{E})^{-1}. \tag{A.31}$$

Swapping $\boldsymbol{A}^{-1}$ with $\boldsymbol{C}_v^{\mathrm{r}}$ and $\boldsymbol{E}^{-1}$ with $\boldsymbol{C}_u^{\mathrm{r}}$, we have

$$\boldsymbol{C}_v^{\mathrm{r}}(\boldsymbol{C}_v^{\mathrm{r}} + \boldsymbol{C}_u^{\mathrm{r}})^{-1}\boldsymbol{C}_v^{\mathrm{r}} = \boldsymbol{C}_v^{\mathrm{r}} - (\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_v^{\mathrm{r}})^{-1}. \tag{A.32}$$

Similarly,

$$\boldsymbol{R}_u^{\mathrm{r}}(\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_v^{\mathrm{r}})^{-1}\boldsymbol{R}_r^{\mathrm{r}} = \boldsymbol{R}_u^{\mathrm{r}} - (\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}. \tag{A.33}$$

By substituting (A.32) into (A.29),

$$\boldsymbol{\mathcal{W}}_u^{\mathrm{r}}\hat{\boldsymbol{W}}_u^{\mathrm{r}} = [\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_u^{\mathrm{r}}\boldsymbol{C}_v^{\mathrm{r}}\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}(\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_v^{\mathrm{r}})^{-1}\boldsymbol{R}_u^{\mathrm{r}}]\boldsymbol{X}_u^\top \boldsymbol{Y}_u. \tag{A.34}$$

Then

$$\boldsymbol{\mathcal{W}}_u^{\mathrm{r}}\hat{\boldsymbol{W}}_u^{\mathrm{r}} = [\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}(\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_v^{\mathrm{r}})^{-1}\boldsymbol{R}_u^{\mathrm{r}}]\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u. \quad \text{(A.35)}$$

Further substituting (A.33) into (A.35), we have

$$\begin{aligned}\boldsymbol{\mathcal{W}}_u^{\mathrm{r}}\hat{\boldsymbol{W}}_u^{\mathrm{r}} &= [\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}}(\boldsymbol{R}_u^{\mathrm{r}} + \boldsymbol{R}_v^{\mathrm{r}})^{-1}\boldsymbol{R}_u^{\mathrm{r}}]\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u \\ &= [\boldsymbol{R}_u^{\mathrm{r}} - \boldsymbol{R}_u^{\mathrm{r}} + (\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}]\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u \\ &= (\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u \\ &= (\boldsymbol{C}_u + \boldsymbol{C}_v + 2\gamma\boldsymbol{I})^{-1}\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u. \quad \text{(A.36)}\end{aligned}$$

Similarly,

$$\boldsymbol{\mathcal{W}}_v^{\mathrm{r}}\hat{\boldsymbol{W}}_v^{\mathrm{r}} = (\boldsymbol{C}_u + \boldsymbol{C}_v + 2\gamma\boldsymbol{I})^{-1}\boldsymbol{X}_v^{\top}\boldsymbol{Y}_v. \quad \text{(A.37)}$$

Thus equation (A.27) can be converted to

$$\hat{\boldsymbol{W}}^{\mathrm{r}} = (\boldsymbol{C}_u + \boldsymbol{C}_v + 2\gamma\boldsymbol{I})^{-1}(\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u + \boldsymbol{X}_v^{\top}\boldsymbol{Y}_v). \quad \text{(A.38)}$$

Since $\hat{\boldsymbol{W}} = \boldsymbol{X}^{\dagger}\boldsymbol{Y}$ and $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_u \\ \boldsymbol{X}_v\end{bmatrix}$, $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{Y}_u \\ \boldsymbol{Y}_v\end{bmatrix}$, with full-column rank of $\boldsymbol{X}$,

$$\begin{aligned}\hat{\boldsymbol{W}} = \boldsymbol{X}^{\dagger}\boldsymbol{Y} &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}Y \quad \text{(A.39)} \\ &= (\begin{bmatrix}\boldsymbol{X}_u^{\mathrm{T}} & \boldsymbol{X}_v^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_u \\ \boldsymbol{X}_v\end{bmatrix})^{-1}\begin{bmatrix}\boldsymbol{X}_u^{\mathrm{T}} & \boldsymbol{X}_v^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}_u \\ \boldsymbol{Y}_v\end{bmatrix}) \\ &= (\boldsymbol{X}_u^{\top}\boldsymbol{X}_u + \boldsymbol{X}_v^{\top}\boldsymbol{X}_v)^{-1}(\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u + \boldsymbol{X}_v^{\top}\boldsymbol{Y}_v) \\ &= (\boldsymbol{C}_u + \boldsymbol{C}_v)^{-1}(\boldsymbol{X}_u^{\top}\boldsymbol{Y}_u + \boldsymbol{X}_v^{\top}\boldsymbol{Y}_v).\end{aligned}$$

By comparing with (A.27) and (A.39), we can obtain the relation between $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{W}}^{\mathrm{r}}$ as follows.

$$\hat{\boldsymbol{W}}^{\mathrm{r}} = (\boldsymbol{C}_u^{\mathrm{r}} + \boldsymbol{C}_v^{\mathrm{r}})^{-1}(\boldsymbol{C}_u + \boldsymbol{C}_v)\hat{\boldsymbol{W}}. \quad \text{(A.40)}$$

By extending to the multi-client scenario, we have

$$\hat{\boldsymbol{W}}_{\mathrm{agg},k}^{\mathrm{r}} = (\boldsymbol{C}_{\mathrm{agg},k}^{\mathrm{r}})^{-1}\boldsymbol{C}_{\mathrm{agg},k}\hat{\boldsymbol{W}}_{\mathrm{agg},k}, \quad \text{(A.41)}$$

where

$$\boldsymbol{C}_{\mathrm{agg},k}^{\mathrm{r}} = \boldsymbol{C}_{\mathrm{agg},k} + k\gamma\boldsymbol{I} = \sum_i^k \boldsymbol{C}_i^{\mathrm{r}},$$

$$\boldsymbol{C}_i^{\mathrm{r}} = \boldsymbol{X}_i^{\top}\boldsymbol{X}_i + \gamma\boldsymbol{I}, \quad \text{(A.42)}$$

which complete the proof. □

## D. Validating AA Laws on Dummy Dataset

Here we validate the AA laws in AFL, whether the aggregated weight $\hat{\boldsymbol{W}}_{\mathrm{agg},K}$ equals to $\hat{\boldsymbol{W}}$ trained on a centralized dataset. This is done by measuring the deviation (i.e., $\Delta\boldsymbol{W} = \|\hat{\boldsymbol{W}} - \hat{\boldsymbol{W}}_{\mathrm{agg},K}\|_1$) between the joint-trained weight and the aggregated one on a dummy dataset.

**Dummy Dataset.** We randomly generate a 512-dimension and 10,000-sample dummy dataset. This dataset has 10 classes, with each class containing an identical number of samples. The samples in the dummy dataset are randomly but evenly distributed to $K$ clients (we set $K = 2, 10, 20, 50, 100, 200$).

**Results.** As indicated in Table A.1, without the RI process, the deviation is negligible for $K = 2, 10$, but it grows with an increasing $K$ and could become unacceptable (e.g., $3.67 \times 10^{12}$ for $K = 200$). This is because the full-column rank assumption might not hold anymore for large $K$. By adopting the RI process, the deviations become negligible (around $10^{-10}$) for various $K$ values as shown in the second row of Table A.1. The RI process introduces $\gamma$ (we adopt $\gamma = 1$ in this case, but any value would suffice) to satisfy full-column rank condition, which is later removed in (16) to restore the AA law's optimality. This experiment has well demonstrated AFL's invariance to data partitioning with empirical evidence. The codes for the dummy data validation can be found in the file *App_Dummy.ipynb* in the released code.

## E. Implementation Details of Experiments

For the compared methods, we set the local epoch to 1 and all the clients are selected to participate each round after local training. The batch size is set to 64 and we employ SGD optimizer with learning rate of 0.05. The number of global communication rounds is set to be 500 since there is little or no performance gain with more rounds. We report the average and standard deviation of best top-1 accuracy in three runs. All the experiments are conducted on a NVIDIA GeForce RTX 4090 GPU. All the compared methods except FedDisco are implemented with PFLlib [45] and FedDisco is implemented upon FedAvg via official codes.

For the specified hyperparameters in compared methods, we tune the parameters via grid search. For FedProx [18], we tune the hyperparameter $\mu$ from $\{0.0001, 0.001, 0.01, 0.1\}$. For MOON [16], we tune the hyperparameter $\mu$ from $\{0.1, 1, 5, 10\}$. For FedDyn [1], we tune the hyperparameter $\alpha$ from $\{0.001, 0.01, 0.1, 1.0\}$. For FedNTD [15], we tune the hyperparameters $\tau$ from $\{0.1, 0.5, 1.0\}$ and $\beta$ from $\{0.1, 0.5, 1.0, 2.0\}$. For FedDisco [43], we tune the hyperparameters $a$ from $\{0.01, 0.05, 0.1, 0.5\}$ and $b$ from $\{0.005, 0.01, 0.05, 0.1\}$. The best parameters we adopted are, $\mu = 0.001$ in FedProx [18], $\mu = 1$ in MOON [16], $\alpha = 1.0$ in FedDyn [1], $\tau = 0.5, \beta = 1.0$ in FedNTD [15] and $a = 0.05, b = 0.01$ in FedDisco [43].

## F. Necessity of FL with Pre-trained Model

To validate the necessity of FL with pre-trained backbone, we train the models locally without aggregation under the setting of $\alpha$=0.1, K=100. We report the average and maximum test accuracy of local training among all the clients. As shown in Table. A.2, without aggregation, the results of local training (12.04% and 16.36%) fall behind FedAvg (56.57%)

Table A.1. Deviation $\Delta \boldsymbol{W}$ between the joint-trained weight and the aggregated one (average of 5 runs).

| Difference | $K = 2$ | $K = 10$ | $K = 20$ | $K = 50$ | $K = 100$ | $K = 200$ |
|---|---|---|---|---|---|---|
| $\Delta \boldsymbol{W}$ (w/o RI) | $7.83 \times 10^{-14}$ | $1.76 \times 10^{-12}$ | $9.86 \times 10^{-1}$ | $5.90$ | $5.93 \times 10^{4}$ | $3.67 \times 10^{12}$ |
| $\Delta \boldsymbol{W}$ (w/ RI) | $4.94 \times 10^{-14}$ | $1.74 \times 10^{-12}$ | $5.09 \times 10^{-10}$ | $8.45 \times 10^{-10}$ | $7.57 \times 10^{-10}$ | $7.81 \times 10^{-10}$ |

and AFL (58.56%) with large gaps. Training local models without FL could suffer from the data heterogeneity and the collaboration among clients is beneficial. This pattern is also validated in previous studies with pre-trained backbone [7].

Table A.2. Comparison between FL teniques including FedAvg and AFL with local training when utilizing pre-trained backbone.

| Methods | Local Max | Local Avg | FedAvg | AFL |
|---|---|---|---|---|
| Acc.(%) | 16.36 | 12.04 | 56.57 | **58.56** |

## G. Comparative Study with Single-Round FL

In this section, we provide a comparative study with another single-round FL technique FedFisher [11] to further validate our proposed AFL. We compare AFL with FedFisher under the setting of $\alpha = 0.1, K = 50$ (larger K will lead to out-of-memory in FedFisher) with the same pre-trained ResNet-18 provided by the repository of FedFisher. As shown in Table A.3, AFL outperforms FedFisher with considerable distance (35.87% v.s. 19.31%). FedFisher utilizes iterative gradient-descent to aggregate local weights and MSE loss is established by the difference of global and local weights to preventing drifting between them during aggregation. However, this technique could still suffer from the data heterogeneity, while AFL formulates the AA law to achieve the invariance to data partitioning, enabling outperforming result when compared with FedFisher.

Table A.3. Comparative study between AFL and FedFisher.

| Methods | FedFisher | AFL |
|---|---|---|
| Acc.(%) | 19.31 | 35.87 |