Appendix of Paper
"Federated Learning with Data-Agnostic Distribution Fusion"

A Pseudo-code of `FedFusion`

Algorithm 1: The `FedFusion` algorithm.

Initialize $\mathbf{w}^0$.
for *each communication round* $t = 0, 1, \dots, T-1$ **do**
  $\mathbf{w}_k^{t+1}$:= the model received from client k
  $\mathbf{d_k} := (\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\sigma}}_k, \hat{\boldsymbol{\beta}}_k, \hat{\gamma}_k)$ // extracted from $\mathbf{w}_k^{t+1}$
  repeat
    Inference $\boldsymbol{\kappa}_m$, $\boldsymbol{\zeta}_{,}$, $\boldsymbol{\nu}_m$, $\boldsymbol{\varsigma}_m$, $\boldsymbol{\nu}^{'}_m$ and $\boldsymbol{\varsigma}^{'}_m$ based on encoder $\boldsymbol{\phi}$
    $\mathbf{b}_k$, $\boldsymbol{\lambda}_k$, $\mathbf{c}_k$:= sampling from distributions with Eq. 3, 5, 6
    $\tilde{\mathbf{z}}_m$:= sampling from $\mathcal{N}(\boldsymbol{\nu}^{'}_m, \boldsymbol{\varsigma}^{'}_m)$
    $\mathbf{z}_k := \sum_{m=1}^{M} b_{km} \cdot \tilde{\mathbf{z}}_m$
    Recover $\mathbf{z_k}$ to $\mathbf{d_k}$ based on decoder $\boldsymbol{\theta}$ with Eq. 10
  until *VAE converge*;
  $\mathbf{w}^{t+1} := \sum_{m=1}^{M} \pi_m \sum_{k=1}^{K} b_{km} \cdot c_{km} \cdot \mathbf{w}_k^{t+1}$ // model aggregation
  broadcast $\mathbf{w}^{t+1}$ to all clients

B Convergence of `FedFusion`

This section proof the convergence of the proposed `FedFusion` method. Before proposing the proof, we present the following assumptions and definitions.

Assumption 1 *(Bounded Taylor's Approximation): We assume that the loss function $f(\cdot)$ has L-smooth and τ-weak convexity, that is, for all $\mathbf{w}_i$ and $\mathbf{w}_j$:*

$$(\mathbf{w}_i - \mathbf{w}_j)^T \nabla f(\mathbf{w}_j) + \frac{\tau}{2}||\mathbf{w}_i - \mathbf{w}_j||^2 \le f(\mathbf{w}_i) - f(\mathbf{w}_j) \le (\mathbf{w}_i - \mathbf{w}_j)^T \nabla f(\mathbf{w}_j) + \frac{L}{2}||\mathbf{w}_i - \mathbf{w}_j||^2,$$

where $\tau \le L$ and $L > 0$.

Note that when $\tau < 0$, the above assumption covers non-convexity functions.

Assumption 2 *(Bounded Gradients Variance): Let $\hat{\mathbf{x}}_k$ is the sampled data from client-k and $\mathbf{g}_k = \nabla f(\mathbf{w}_k, \hat{\mathbf{x}}_k)$ is the gradient in regard to $\hat{\mathbf{x}}_k$. We assume the stochastic gradients $\mathbf{g}_k$ has the following upper-bounded variances in the whole training process: (1) $\mathbb{E}||\mathbf{g}_k - \mathbb{E}[\mathbf{g}_k]||^2 \le V$, $V \in \mathbb{R}$; and (2) $\mathbb{E}||\mathbf{g}_k||^2 \le G, G \in \mathbb{R}$.*

Definition 1 *(Diameter of Domain): Given a function $f(\mathbf{w})$, where $\mathbf{w} \in \mathbb{W}$, and $\mathbb{W}$ is f's domain of definition. The diameter of $\mathbb{W}$ is denoted by Γ: for every $\mathbf{w}_i, \mathbf{w}_j \in \mathbb{W}$: $||\mathbf{w}_i - \mathbf{w}_j|| \le \Gamma$.*

Theorem 1 *(Convergence Bound): If Assumption 1 and 2 hold, with learning epoch T, local epoch E, diameter of domain Γ, and learning rate η, we have the following convergence bound for the proposed `FedFusion` algorithm:*

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \le \frac{L}{E+T}\left(\frac{A}{\tau} + \frac{E+1}{2}\Gamma^2\right), \tag{1}$$

where $\mathbf{w}^*$ are the optimal parameters; $\mathbf{w}^T$ are the parameters at the T-th learning epoch, and A is a constant:

$$A = 4\eta(E-1)^2 G + \eta \sum_{m=1}^{M} \pi_m^2 V + 2\eta G + 2\Gamma G + L\Gamma^2.$$

Corollary 1 *(Convergence Rate): If Assumption 1 and 2 hold, with learning epoch T and local epoch E, we have the following convergence rate for the proposed `FedFusion` algorithm:*

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \mathcal{O}\left(\frac{1}{E+T}\right). \tag{2}$$

The proof is based on the convergence rate analyzed in [1]. We denote the optimal solution for $\mathbf{w}$ as $\mathbf{w}^*$. Notice that $\bar{\mathbf{v}}^{t+1} = \bar{\mathbf{w}}^t - \eta \mathbf{g}^t$ where $\mathbf{g}^t$ is the gradient. According to Assumption 1, 2 we have:

$$||\bar{\mathbf{v}}^{t+1} - \mathbf{w}^*||^2 = ||\bar{\mathbf{w}}^t - \mathbf{w}^* - \eta \mathbf{g}^t||^2 + \eta^2 ||\mathbf{g}^t - \bar{\mathbf{g}}^t||^2, \tag{3}$$

and

$$||\bar{\mathbf{w}}^t - \mathbf{w}^* - \eta \mathbf{g}^t||^2 = ||\bar{\mathbf{w}}^t - \mathbf{w}^*||^2 - 2\eta \langle \bar{\mathbf{w}}^t - \mathbf{w}^*, \bar{\mathbf{g}}^t \rangle + \eta^2 ||\bar{\mathbf{g}}^t||^2. \tag{4}$$

Let $B_1 = -2\eta \langle \bar{\mathbf{w}}^t - \mathbf{w}^*, \bar{\mathbf{g}}^t \rangle$ and $B_2 = \eta^2 ||\bar{\mathbf{g}}^t||^2$. With Assumption 2, we have

$$B_2 = \eta^2 ||\bar{\mathbf{g}}^t||^2 \leq \eta^2 G. \tag{5}$$

We rewrite B_1 by

$$\begin{aligned} B_1 &= -2\eta \langle \bar{\mathbf{w}}^t - \mathbf{w}^*, \bar{\mathbf{g}}^t \rangle \\ &= -2\eta \sum_{m=1}^{M} \pi_m \langle \bar{\mathbf{w}}^t - \mathbf{w}_m^t, \nabla f_m(\mathbf{w}_m^t) \rangle - 2\eta \sum_{m=1}^{M} \pi_m \langle \mathbf{w}_m^t - \mathbf{w}^*, \nabla f_m(\mathbf{w}_m^t) \rangle. \end{aligned}$$

By Cauchy-Schwarz inequality[1] and AM-GM inequality[2], we have

$$-2\langle \bar{\mathbf{w}}^t - \mathbf{w}_m^t, \nabla f_m(\mathbf{w}_m^t) \rangle \leq \frac{1}{\eta} ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 + \eta ||\nabla f_m(\mathbf{w}_m^t)||^2. \tag{6}$$

By Assumption 1, we have τ-weak convexity of loss function $f(\cdot)$:

$$-\langle \mathbf{w}_m^t - \mathbf{w}^*, \nabla f_m(\mathbf{w}_m^t) \rangle \leq -(f_m(\mathbf{w}_m^t) - f_m(\mathbf{w}^*)) - \frac{\tau}{2} ||\mathbf{w}_m^t - \mathbf{w}^*||^2. \tag{7}$$

Applying Eq. (5)-(7) on Eq. (4), we have

$$\begin{aligned} ||\bar{\mathbf{w}}^t - \mathbf{w}^* - \eta \mathbf{g}^t||^2 \leq\ & ||\bar{\mathbf{w}}^t - \mathbf{w}^*||^2 + \eta^2 G + \\ & \eta \sum_{m=1}^{M} \pi_m \left(\frac{1}{\eta} ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 + \eta ||\nabla f_m(\mathbf{w}_m^t)||^2 \right) - \\ & 2\eta \sum_{m=1}^{M} \left(f_m(\mathbf{w}_m^t) - f_m(\mathbf{w}^*) + \frac{\tau}{2} ||\mathbf{w}_m^t - \mathbf{w}^*||^2 \right). \end{aligned}$$

The above inequality can be rewrote as:

$$\begin{aligned} ||\bar{\mathbf{w}}^t - \mathbf{w}^* - \eta \mathbf{g}^t||^2 &\leq (1 - \tau\eta) ||\mathbf{w}_m^t - \mathbf{w}^*||^2 + \sum_{m=1}^{M} \pi_m ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 + 2\eta^2 G - 2\eta \sum_{m=1}^{M} (f_m(\mathbf{w}_m^t) - f_m(\mathbf{w}^*)) \\ &= (1 - \tau\eta) ||\mathbf{w}_m^t - \mathbf{w}^*||^2 + \sum_{m=1}^{M} \pi_m ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 + 2\eta^2 G + 2\eta \sum_{m=1}^{M} \pi_m (f_m(\mathbf{w}^*) - f_m(\mathbf{w}_m^t)). \end{aligned} \tag{8}$$

With Assumption 1, we have L-smooth of loss function $f(\cdot)$:

$$f_m(\mathbf{w}^*) - f_m(\mathbf{w}_m^t) \leq \langle \mathbf{w}^* - \mathbf{w}_m^t, \nabla f_m(\mathbf{w}_m^t) \rangle + \frac{L}{2} ||\mathbf{w}^* - \mathbf{w}_m^t||^2. \tag{9}$$

[1] https://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality

[2] https://en.wikipedia.org/wiki/Inequality_of_arithmetic_and_geometric_means

Using Assumption 1 and Defination 1, we have:

$$f_m(\mathbf{w}^*) - f_m(\mathbf{w}_m^t) \leq \Gamma G + \frac{L}{2}\Gamma^2. \tag{10}$$

Applying Eq. (8) and Eq. (10), we can rewrite Eq. (3) as

$$\begin{aligned} ||\bar{\mathbf{v}}^{t+1} - \mathbf{w}^*||^2 \quad &\leq \quad (1-\tau\eta)||\mathbf{w}_m^t - \mathbf{w}^*||^2 + \sum_{m=1}^{M} \pi_m ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 + \\ &\quad\quad \eta^2 ||\mathbf{g}^t - \bar{\mathbf{g}}^t||^2 + 2\eta^2 G + 2\eta\Gamma G + \eta L \Gamma^2. \end{aligned}$$

With Assumption 2, we have:

$$\begin{aligned} \mathbb{E}||\mathbf{g}^t - \bar{\mathbf{g}}^t||^2 \quad &= \quad \mathbb{E}||\sum_{m=1}^{M} \pi_m (\nabla f_m(\mathbf{w}_m^t, \mathbf{x}_m) - \nabla f_m(\mathbf{w}_m^t))||^2 \\ &= \quad \sum_{m=1}^{M} \pi_m^2 ||\nabla f_m(\mathbf{w}_m^t, \mathbf{x}_m) - \nabla f_m(\mathbf{w}_m^t)||^2 \\ &\leq \quad \sum_{m=1}^{M} \pi_m^2 V. \end{aligned}$$

From Lemma 3 of [1], we have:

$$\sum_{m=1}^{M} \pi_m ||\bar{\mathbf{w}}^t - \mathbf{w}_m^t||^2 \leq 4\eta^2 (E-1)^2 G. \tag{11}$$

Let $\Delta^t = \mathbb{E}||\mathbf{w}_m^t - \mathbf{w}^*||^2$, with Assumption 1, we have

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \langle \mathbf{w}^T - \mathbf{w}^*, \nabla f(\mathbf{w}^*) \rangle + \frac{L}{2}||\mathbf{w}^T - \mathbf{w}^*||^2. \tag{12}$$

For the optimal solution $\mathbf{w}^*$, $\nabla f(\mathbf{w}^*) = 0$. So

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \frac{L}{2}||\mathbf{w}^T - \mathbf{w}^*||^2 = \frac{L}{2}\Delta^t, \tag{13}$$

and

$$\Delta^{t+1} \leq (1-\tau\eta)\Delta^t + \eta A, \tag{14}$$

where

$$A = 4\eta(E-1)^2 G + \eta \sum_{m=1}^{M} \pi_m^2 V + 2\eta G + 2\Gamma G + L\Gamma^2. \tag{15}$$

Let $\eta = \frac{\beta}{t+E}$ and $v = \max\left\{\frac{\beta A}{\beta\tau - 1}, (E+1)\Delta_1\right\}$, we have

$$\begin{aligned} \Delta^{t+1} \quad &\leq \quad (1-\tau\eta)\Delta^t + \eta A \\ &\leq \quad (1 - \tau\frac{\beta}{t+E})\frac{v}{t+E} + \frac{\beta}{t+E}A \\ &= \quad \frac{t+E-1}{(t+E)^2}v + \left(\frac{\beta A}{(t+E)^2} - \frac{\tau\beta - 1}{(t+E)^2}v\right) \\ &\leq \quad \frac{v}{t+E+1}. \end{aligned} \tag{16}$$

Substituting Eq. (16) to Eq. (13), we get

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \frac{L}{2}\frac{v}{E+t}. \tag{17}$$

Taking $\beta = \frac{2}{\tau}$, the upper bound of v can be given by

$$v \leq \frac{\beta A}{\beta\tau - 1} + (E+1)\Delta_1 \leq \frac{2A}{\tau} + (E+1)\Delta_1, \tag{18}$$

where $\Delta_1 = ||\mathbf{w}^0 - \mathbf{w}^*||^2 \leq \Gamma^2$.

With Eq. (17) and Eq. (18), we have

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \frac{L}{E+t}(\frac{A}{\tau} + \frac{E+1}{2}\Gamma^2), \tag{19}$$

where

$$A = 4\eta(E-1)^2 G + \eta \sum_{m=1}^{M} \pi_m^2 V + 2\eta G + 2\Gamma G + L\Gamma^2. \tag{20}$$

Theorem 1 is proved.

To prove Corollary 1, we take $\eta = \frac{2}{\tau(T+E)}$. With Theorem 1, we have

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \frac{L}{E+T}\left(\frac{C_1}{\tau} + \frac{C_2}{\tau^2(T+E)} + C_3\right), \tag{21}$$

where $C_1 = 2\Gamma G + L\Gamma^2$ $C_2 = 8(E-1)^2 G + 2\sum_{m=1}^{M} \pi_m^2 V + 4G$ and $C_3 = \frac{E+1}{2}\Gamma^2$ are constants. So we have:

$$\mathbb{E}[f(\mathbf{w}^T)] - f(\mathbf{w}^*) \leq \mathcal{O}\left(\frac{1}{E+T}\right), \tag{22}$$

which proves Corollary 1.

References

[1] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on Non-IID data. In *ICLR*, 2020.