In the supplemental materials, we provide the theoretical analysis in Appendix A, and the additional experimental details in Appendix B.

A. Theoretical Analysis

A.1. Optimization Consistency of Model Deviation

In the following, we provide the theorem related to optimization consistency of model deviations. When the aggregating weights, i.e., $\boldsymbol{p}$, achieve optimal, the model deviation rate is equally contributed to global updating.

Theorem 3 (Optimization consistency of model deviations)**.** *Rethinking the Lagrangian of dual form in Eq. (10),*

$$\boldsymbol{J} = \min_{\boldsymbol{p}} \frac{\eta_g^2}{2\phi} \|(\nabla \log \boldsymbol{u})^\top \boldsymbol{p}\|^2 + \lambda_E \boldsymbol{p}^\top \mathbf{1}, \tag{15}$$

it holds $\nabla \log(u_i\left(\boldsymbol{\theta}_g^t\right)) = \nabla \log(u_j\left(\boldsymbol{\theta}_g^t\right)), \forall i \neq j \in [K]$.

Proof. By deviating Eq. (15) with regarding to $\boldsymbol{p}$, we can obtain $(\nabla \log \boldsymbol{u}) \nabla \log(\boldsymbol{u})^\top \boldsymbol{p}^* = -\frac{\phi \lambda_E}{\eta_g^2} \boldsymbol{I}$. Remind that $\boldsymbol{d}^* = \frac{\eta_g}{\phi} (\nabla \log \boldsymbol{u})^\top \boldsymbol{p}^*$, for client $i \neq j \in [K]$ and $\eta_g \to 0$, we finally have consistent model deviation change rate as below:

$$\begin{aligned} &\lim_{\eta_g \to 0} c_i\left(\eta_g, \boldsymbol{d}^*\right) = \nabla \log(u_i\left(\boldsymbol{\theta}_g^*\right)) \boldsymbol{d}^* \\ &= \nabla \log(u_j\left(\boldsymbol{\theta}_g^*\right)) \boldsymbol{d}^* = \lim_{\eta_g \to 0} c_j\left(\eta_g, \boldsymbol{d}^*\right). \end{aligned} \tag{16}$$

□

Therefore, the global model updates with a direction that balances all model deviation change rates, obtaining consistent parameters for server and client models.

A.2. Bound of Client Model Divergence

In this part, we first introduce mild and general assumptions [23], and induct the model updating divergence bound for each client.

Assumption 5. *Let $F_k(\boldsymbol{\theta})$ be the expected model objective for client k, and assume $F_1, \cdots, F_K$ are all L-smooth, i.e., for all $\boldsymbol{\theta}_k$, $F_k(\boldsymbol{\theta}_k) \leq F_k(\boldsymbol{\theta}_k) + (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k)^\top \nabla F_k(\boldsymbol{\theta}_k) + \frac{L}{2} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k\|_2^2$.*

Assumption 6. *Let $F_1, \cdots, F_N$ are all μ-strongly convex: for all $\boldsymbol{\theta}_k$, $F_k(\boldsymbol{\theta}_k) \geq F_k(\boldsymbol{\theta}_k) + (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k)^\top \nabla F_k(\boldsymbol{\theta}_k) + \frac{\mu}{2} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k\|_2^2$.*

Assumption 7. *Let ξ_k^t be sampled from the k-th client's local data uniformly at random. The variance of stochastic gradients in each client is bounded: $\mathbb{E} \left\|\nabla F_k\left(\boldsymbol{\theta}_k^t, \xi_k^t\right) - \nabla F_k\left(\boldsymbol{\theta}_k^t\right)\right\|^2 \leq \sigma_k^2$.*

Assumption 8. *The expected squared norm of stochastic gradients is uniformly bounded, i.e., $\mathbb{E} \left\|\nabla F_k\left(\boldsymbol{\theta}_k^t, \xi_k^t\right)\right\|^2 \leq V^2$ for all $k = 1, \cdots, N$ and $t = 1, \cdots, T-1$*

Next, we introduce the lemma related to the bound of client model divergence.

Lemma 2 (Bound of Client Model Divergence)**.** *With assumption 8, η_t is non-increasing and $\eta_t < 2\eta_{t+E}$ (learning rate of t-th round and E-th epoch) for all $t \geq 0$, there exists $t_0 \leq t$, such that $t - t_0 \leq E - 1$ and $\boldsymbol{\theta}_k^{t_0} = \boldsymbol{\theta}^{t_0}$ for all $k \in [N]$. It follows that*

$$\mathbb{E}\left[\sum_k^K p_k \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2\right] \leq 4\eta_t^2 (E-1)^2 V^2. \tag{17}$$

Proof. Let E be the maximal local epoch. For any round $t > 0$, communication rounds from t to t_0 exist $t - t_0 < E - 1$. and the global model $\boldsymbol{\theta}^{t_0}$ and each local model $\boldsymbol{\theta}_k^{t_0}$ are same at round t_0.

$$\begin{aligned} &\mathbb{E}\left[\sum_k^K p_k \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2\right] \\ &= \mathbb{E}\left[\sum_k^K p_k \|(\boldsymbol{\theta}_k^t - \boldsymbol{\theta}^{t_0}) - (\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t_0})\|^2\right] && (17a) \\ &\leq \mathbb{E} \sum_k^K p_k \|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}^{t_0}\|^2 && (17b) \\ &= \mathbb{E} \sum_k^K p_k \left\|\sum_{t=t0}^{t-1} \eta_t \nabla F_k(\boldsymbol{\theta}_k^t, \xi_k^t)\right\|^2 && (17c) \\ &\leq \mathbb{E} \sum_k^K p_k (t - t_0) \sum_{t=t0}^{t-1} \eta_{t_0}^2 \left\|\nabla F_k(\boldsymbol{\theta}_t^k, \xi_k^t)\right\|^2 && (17d) \\ &\leq 4\eta_t^2 (E-1)^2 V^2, && (17e) \end{aligned}$$

where the Eq. (17b) holds since $\mathbb{E}(\boldsymbol{\theta}_k^t - \boldsymbol{\theta}^{t_0}) = \boldsymbol{\theta}^t - \boldsymbol{\theta}^{t_0}$, and $\mathbb{E}\|X - \mathbb{E}(X)\| \leq \mathbb{E}\|X\|$, and Eq. (17d) derives from Jensen inequality. □

A.3. Convergence Error Bound

Definition 1 (Heterogeneity Quantification [23])**.** Let F^* and F_k^* be the minimum values of F and F_k, respectively. We use the term $\Gamma = F^* - \sum_{k=1}^N p_k F_k^*$ for quantifying the degree of non-IID. If the data are IID, then Γ obviously goes to zero as the number of samples grows. If the data are non-IID, then Γ is nonzero, and its magnitude reflects the heterogeneity of the data distribution.

Theorem 4 (Convergence Error Bound)**.** *Let assumptions 5-8 hold, and L, μ, σ_k, V be defined therein. Let $\kappa = \frac{L}{\mu}, \gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. The FedU^2 with full client participation satisfies*

$$\mathbb{E}\left[F\left(\overline{\boldsymbol{\theta}}^t\right)\right] - F^* \leq \frac{\kappa}{\gamma + t} \left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right),$$

where $B = 4(E-1)^2V^2 + K + 2\Gamma$.

Proof. By L-smooth assumption 5, we can obtain:

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{\theta}^t) - F(\boldsymbol{\theta}^*)\right] \\
&\leq \mathbb{E}\left[(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*)^\top \nabla F(\boldsymbol{\theta}^*) + \frac{L}{2}\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right\|^2\right] \\
&= \mathbb{E}\left[\frac{L}{2}\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right\|^2\right].
\end{aligned} \tag{18}$$

Since the updating in **EUA** is $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t \boldsymbol{d}^t$ for $\boldsymbol{d}_t^* = \frac{\eta_t}{\phi}(\nabla \log \boldsymbol{u}^t)^\top \boldsymbol{p}_t^* = \sum_k^K p_k \nabla F_k(\boldsymbol{\theta}_k^t)$, we can rewrite it as:

$$\begin{aligned}
&\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\|^2 \\
&= \|\boldsymbol{\theta}^t - \eta_t \boldsymbol{d}_t - \boldsymbol{\theta}^*\|^2 \\
&= \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2 - 2\left\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}^*, \eta_t \boldsymbol{d}_t \right\rangle + \eta \|\boldsymbol{d}_t\|^2.
\end{aligned} \tag{19}$$

Next, we induce the bound of the second term.

$$\begin{aligned}
&\left\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}^*, \eta_t \boldsymbol{d}_t \right\rangle \\
&= \sum_k^K p_k \eta_t \left\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t, \nabla F_k(\boldsymbol{\theta}_k^t) \right\rangle - \sum_k^K p_k \eta_t \left\langle \boldsymbol{\theta}_k^t - \boldsymbol{\theta}^*, \nabla F_k(\boldsymbol{\theta}_k^t) \right\rangle
\end{aligned} \tag{20}$$

By Cauchy-Schwarz inequality and AM-GM inequality, we have inequality of the first term of Eq. (20):

$$-2\left\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t, \nabla F_k(\boldsymbol{\theta}_k^t) \right\rangle \leq \frac{1}{\eta_t}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2 + \eta_t \|\nabla F_k(\boldsymbol{\theta}_k^t)\|^2. \tag{21}$$

By the $\mu-$strong convexity of $F_k(\cdot)$, we have

$$-\left\langle \boldsymbol{\theta}_k^t - \boldsymbol{\theta}^*, \nabla F_k(\boldsymbol{\theta}_k^t) \right\rangle \leq -\left(F_k(\boldsymbol{\theta}_k^t) - F_k(\boldsymbol{\theta}^*)\right) - \frac{\mu}{2}\|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}^*\|^2. \tag{22}$$

In Theorem 3, we get $(\nabla \log \boldsymbol{u})^\top \nabla \log(\boldsymbol{u}) \boldsymbol{p}^* = -\frac{\phi \lambda_E}{\eta_t^2} \boldsymbol{I}$, which indicates that:

$$\begin{aligned}
\|\boldsymbol{d}_t\|^2 &= \frac{\eta_t^2}{\phi}\|(\nabla \log \boldsymbol{u}^t)^\top \boldsymbol{p}\|^2 \\
&= \lambda_E \|\boldsymbol{p}^\top\|^2 \\
&\leq K,
\end{aligned} \tag{23}$$

where the last inequation holds due to $\lambda_E < 1$ and $\|\boldsymbol{p}\| \leq K$. By combining Eq. (21)-(23) and Lemma 2, it follows that

$$\begin{aligned}
&\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\|^2 \\
&= \|\boldsymbol{\theta}^t - \eta_t \boldsymbol{d}_t - \boldsymbol{\theta}^*\|^2 \\
&\leq \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2 + \eta_t \sum_k^K p_k \left(\frac{1}{\eta_t}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2 + \eta_t \|\nabla F_k(\boldsymbol{\theta}_k^t)\|^2\right) \\
&+ 2\eta_t \sum_k^K p_k \left(-\left(F_k(\boldsymbol{\theta}_k^t) - F_k(\boldsymbol{\theta}^*)\right) - \frac{\mu}{2}\|\boldsymbol{\theta}_k^t - \boldsymbol{\theta}^*\|^2\right) + \eta_t^2 \|\boldsymbol{d}_t\|^2 \\
&= (1 - \mu\eta_t)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2 + \sum_k^K p_k \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2 + \eta_t^2 K + 2\eta_t^2 \Gamma \\
&\leq (1 - \mu\eta_t)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2 + 4\eta_t^2 (E-1)^2 V^2 + \eta_t^2 K + 2\eta_t^2 \Gamma.
\end{aligned} \tag{24}$$

Lastly, let $D_t = \mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2$, it follows that

$$D_{t+1} \leq (1 - \eta_t \mu) D_t + \eta_t^2 B, \tag{25}$$

where $B = 4(E-1)^2V^2 + K + 2\Gamma$.

For a diminishing stepsize, $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. For $v = \max\{\frac{\beta^2 B}{\beta\mu - 1}, (\gamma+1)D_1\}$, by definition, it holds $D_t \leq \frac{v}{\gamma+t}$ for $t = 1$. Assume $D_t \leq \frac{v}{\gamma+t}$ holds, then we expand as below:

$$\begin{aligned}
D_{t+1} &\leq (1 - \eta_t \mu) D_t + \eta_t^2 B \\
&\leq \left(1 - \frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2} v + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\mu - 1}{(t+\gamma)^2} v\right] \\
&\leq \frac{v}{t+\gamma+1}.
\end{aligned} \tag{26}$$

Recall Eq. (18), we finally catch:

$$\mathbb{E}\left[F(\theta^t) - F(\theta^*)\right] \leq \frac{L}{2} D_t \leq \frac{L}{2}\frac{v}{\gamma+t}. \tag{27}$$

Following the specific case of [23], we can choose $\beta = \frac{2}{\mu}, \gamma = \max\left\{8\frac{L}{\mu}, E\right\} - 1$ and denote $\kappa = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu}\frac{1}{\gamma+t}$. One can verify that the choice of η_t satisfies $\eta_t \leq 2\eta_{t+E}$ for $t \geq 1$. Then, we have

$$\begin{aligned}
v &= \max\left\{\frac{\beta^2 B}{\beta\mu - 1}, (\gamma+1)\Delta_1\right\} \\
&\leq \frac{\beta^2 B}{\beta\mu - 1} + (\gamma+1)\Delta_1 \\
&\leq \frac{4B}{\mu^2} + (\gamma+1)D_1
\end{aligned} \tag{28}$$

and

$$\mathbb{E}\left[F\left(\overline{\boldsymbol{\theta}}^t\right)\right] - F^* \leq \frac{L}{2}\frac{v}{\gamma+t} \leq \frac{\kappa}{\gamma+t}\left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2} D_1\right). \tag{29}$$

As we can see, FedU^2 similarly converges to a generalization error bound as the FedAvg-like FL model with non-IID data. Discriminatively, benefiting from the optimization of **EUA**, the communication round multiplies with a smaller B. □

B. Experimental Supplementary

B.1. Hyper-parameter Sensitivity Analysis

In the following, we study the sensitivity of remaining highly relevant hyper-parameters, i.e., the effect of client numbers and local epochs. Specifically, we compare FedU^2-SimCLR and its runner-up method, i.e., FedX-SimCLR, on CIFAR10 $\alpha = 0.1$, by varying the local epochs $E = \{5, 10, 20, 50\}$ in Fig. 7 and the number of clients $K = \{5, 10, 20, 50, 100\}$ in Fig. 8. We train all models until converge to obtain fairly comparable results. As we can see: (1) With the increase of local epochs, each client of FedX-SimCLR obtains a better-performing model, while

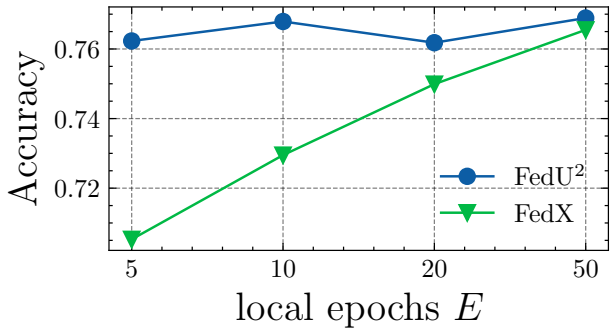


Figure 7. The effect of local epochs E (on CIFAR10 $\alpha = 0.1$).

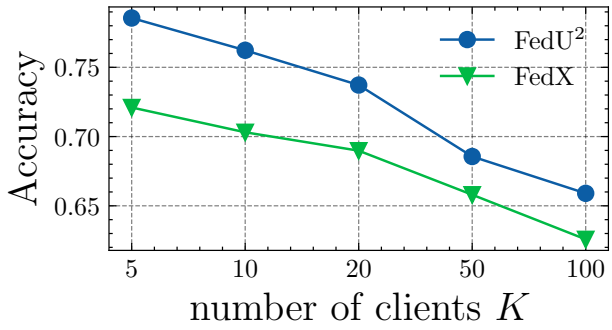


Figure 8. The effect of client number K (on CIFAR10 $\alpha = 0.1$).

each client of FedU^2-SimCLR is insensitive. This states that FedU^2 balances the client model deviation change rate in **EUA**, bringing the benefits of quick convergence. (2) The performance of all methods decreases when the number of clients increases, but FedU^2-SimCLR consistently outperforms FedX-SimCLR. It validates that enhancing uniform and unified representations will make FUSL methods more generalizable to the cases of various participants amounts.

B.2. Enlarged Figures in Visualization

In our main paper, we depict the top-k singular values of covariance matrix representations in Fig. (3), the corresponding 3-D representation in Fig. (4), and the distribution of data representation in Fig. (5) between global model and randomly sampled local models. The purpose of the above figures is to illustrate the representation enhancement of FedU^2. In Fig. 9-11, we enlarge these figures to explore the detailed comparisons. In terms of Fig. 11, FedU^2 keeps the unified representation between global and local models as well as clearer decision boundary for each class.

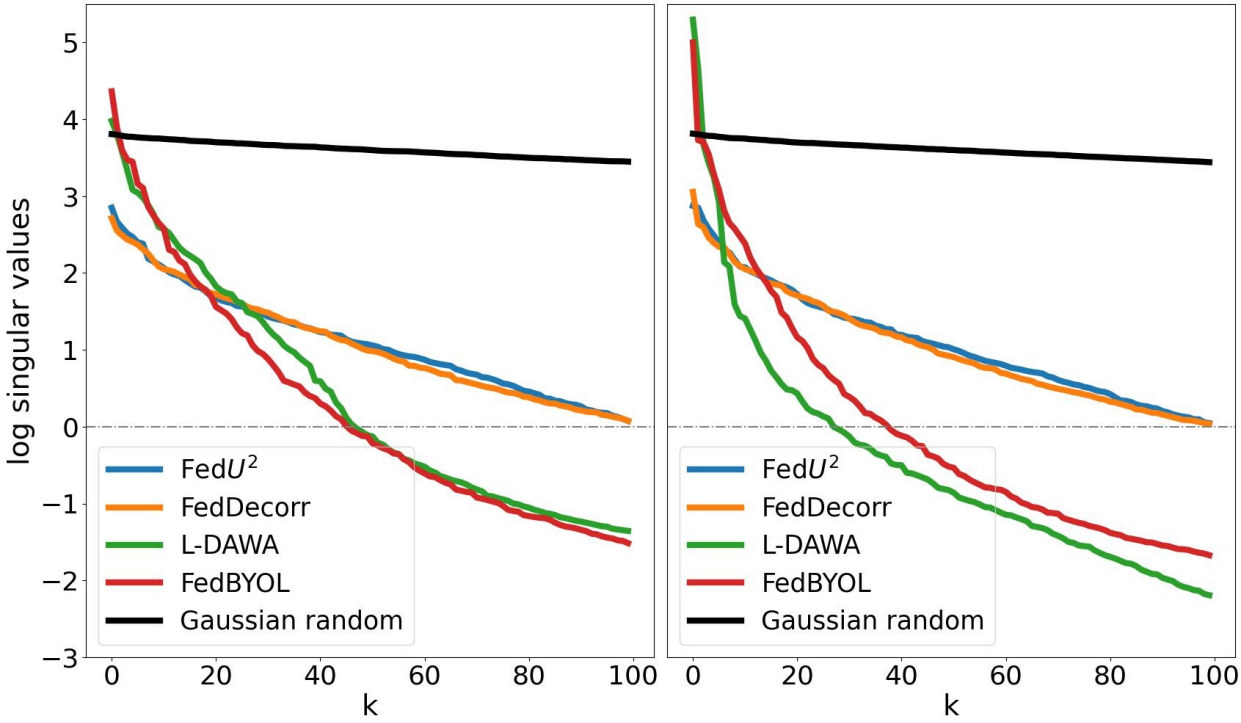
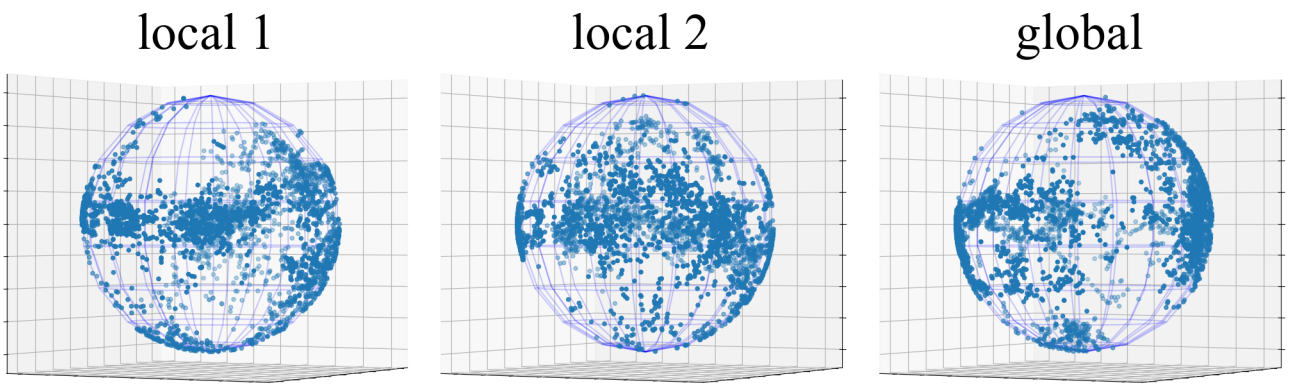
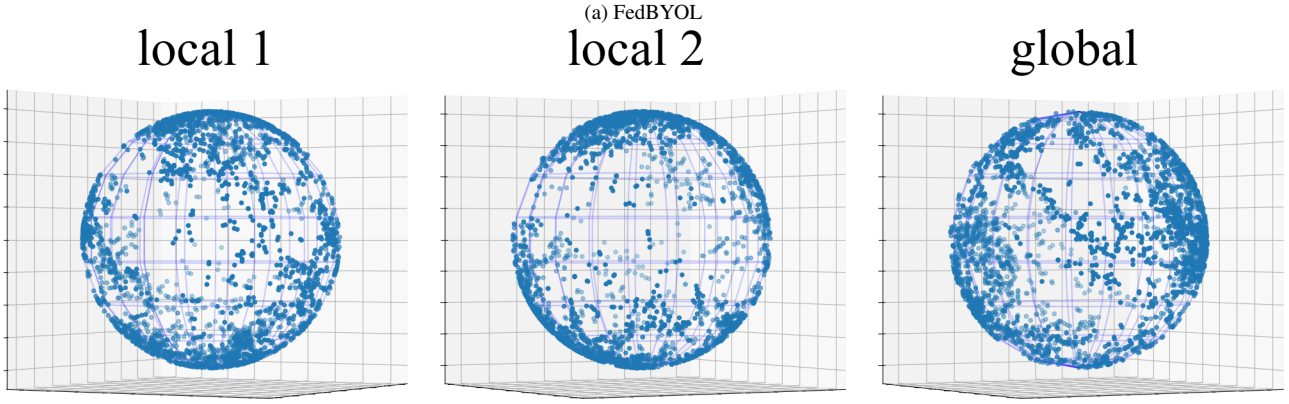


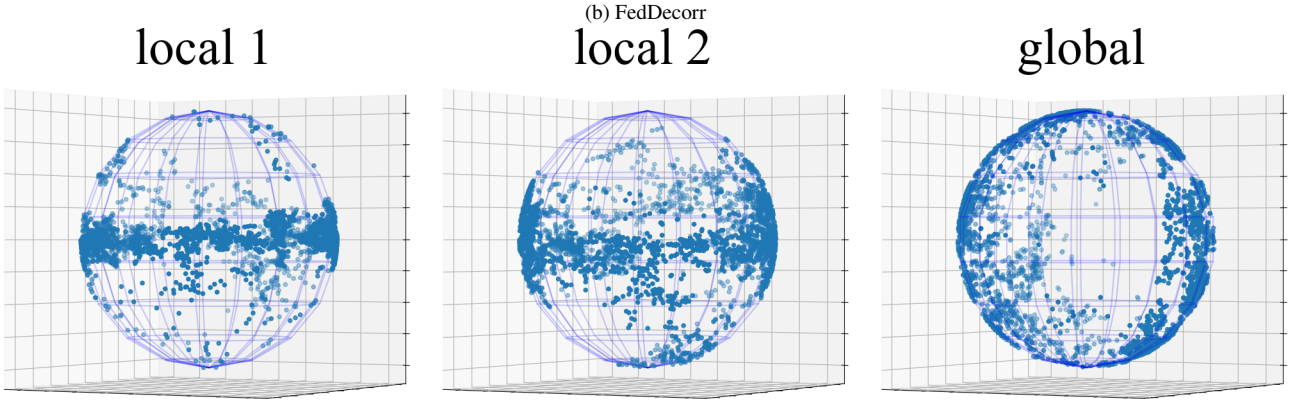
Figure 9. Top k log singular values of the covariance matrix of global model (left) and local model (right) representations.



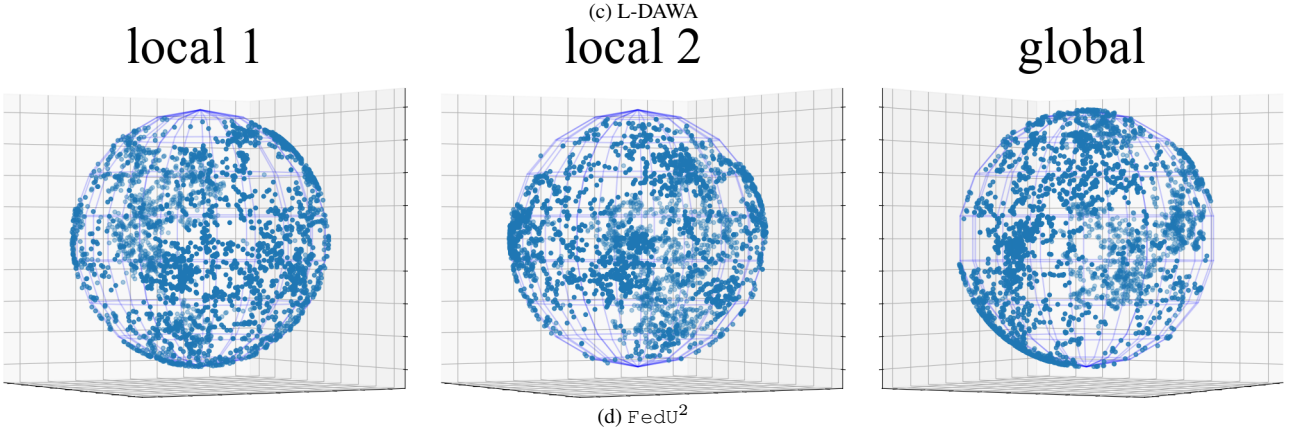
(a) FedBYOL



(b) FedDecorr

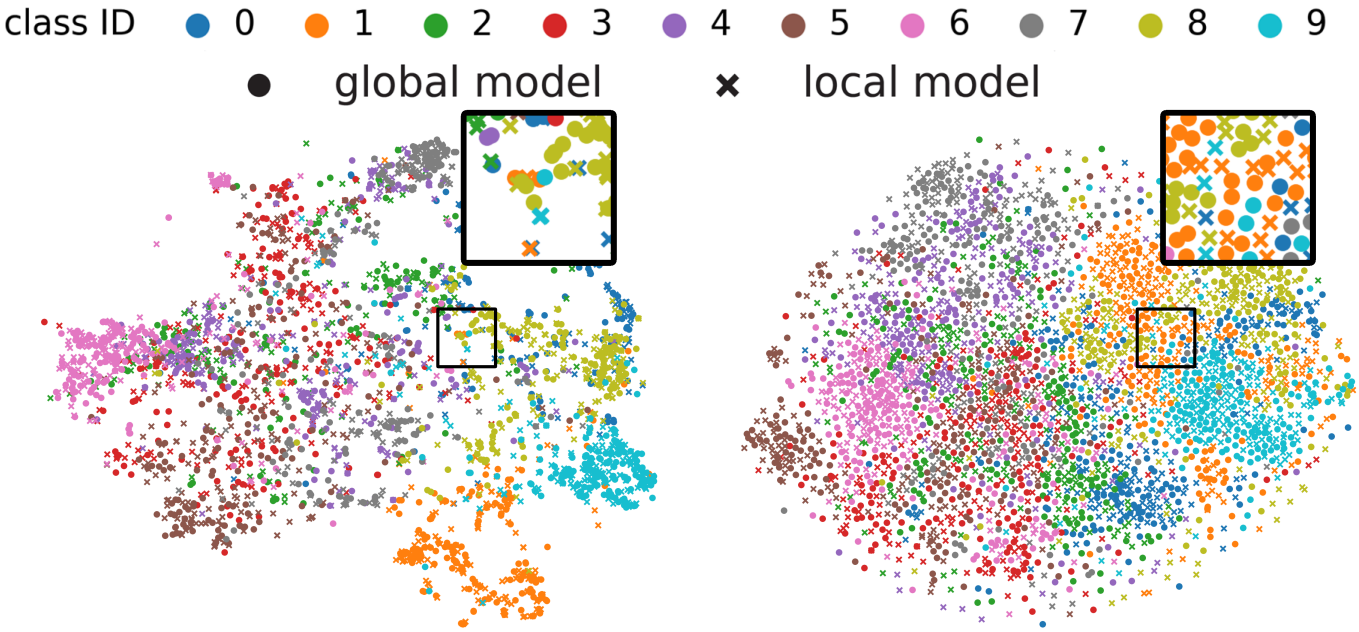


(c) L-DAWA



(d) FedU^2

Figure 10. The representations collapse issue on the sphere using BYOL model (on CIFAR10 $\alpha = 0.1$ Cross-silo). The more blank representation space, the more severe collapse issue is.



(b) FedBYOL

(c) FedDecorr

(d) L-DAWA

(e) FedU^2

Figure 11. The distributions of data representations using global and local BYOL model (on CIFAR10 $\alpha = 0.1$ Cross-silo).