FedAlign: Differentially Private Distribution Alignment for Non-IID Federated Learning

Supplementary Material

A. Experimental Settings

A.1. Experimental Environments

All experiments were conducted on a workstation equipped with an Intel(R) Xeon(R) Bronze 3206R CPU @ 1.90 GHz (16 cores) and an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The system runs CentOS with Python 3.8. PyTorch (version 2.0.0) was used to implement *FedAlign* and all baseline models, ensuring a consistent and reproducible experimental setup.

A.2. Implementation Details

All experiments are conducted with a fixed set of 10 clients, where all clients participate in every communication round (participation ratio = 1). Each client performs 1 local epoch per round, and the training batch size is set to 10. Models are trained using the stochastic gradient descent (SGD) optimizer with a fixed learning rate of 0.01 and a momentum coefficient of 0.9.

The experiments are designed to evaluate the performance of *FedAlign* under both system and data heterogeneity conditions.

System Heterogeneity: To simulate variations in computational capabilities among clients, a hyperparameter delay is introduced. Larger values of delay correspond to higher disparities in client processing speeds. In the reported experiments, the delay is set to 0.1 to represent moderate system heterogeneity.

Data Heterogeneity: Two types of Non-IID scenarios are considered. First, label and quantity skew are induced by partitioning each clients data according to a Dirichlet distribution with concentration parameter $\alpha = 0.5$. Second, feature skew is modeled by adding Gaussian noise $\mathcal{N}(0, \beta)$ to the training and testing datasets, with $\beta \in \{0.05, 0.1\}$.

All reported results represent the mean and standard deviation over five independent runs, excluding the maximum and minimum performance values. The best performance in each setting is highlighted in bold. Other hyperparameters, including dropout rate and data partitioning strategy, are uniformly configured across all clients to ensure reproducibility.

B. Detailed Proofs

B.1. Proof of Theorem 1

Proof. Let $\tilde{g}_k$ be the stochastic gradient of client k, and let the global DP gradient be

$$\tilde{g}^{\mathrm{DP}} = \sum_{j=1}^{K} \frac{n_j}{N} \tilde{g}_j + \eta. \tag{15}$$

Using the identity

$$\|\tilde{g}_k - \tilde{g}^{\mathrm{DP}}\|^2 = \|\tilde{g}_k\|^2 + \|\tilde{g}^{\mathrm{DP}}\|^2 - 2\langle \tilde{g}_k, \tilde{g}^{\mathrm{DP}} \rangle, \tag{16}$$

and taking expectations yields

$$\mathbb{E}\|\tilde{g}_k - \tilde{g}^{\mathrm{DP}}\|^2 = \mathbb{E}\|\tilde{g}_k\|^2 - 2\mathbb{E}\langle \tilde{g}_k, \tilde{g}^{\mathrm{DP}} \rangle + \mathbb{E}\|\tilde{g}^{\mathrm{DP}}\|^2. \tag{17}$$

Inner product expansion. Substituting (15),

$$\mathbb{E}\langle \tilde{g}_k, \tilde{g}^{\mathrm{DP}} \rangle = \sum_{j=1}^{K} \frac{n_j}{N} \mathbb{E}\langle \tilde{g}_k, \tilde{g}_j \rangle + \mathbb{E}\langle \tilde{g}_k, \eta \rangle. \tag{18}$$

Since η is zero-mean and independent of $\tilde{g}_k$,

$$\mathbb{E}\langle \tilde{g}_k, \eta \rangle = 0. \tag{19}$$

For arbitrary random vectors A, B,

$$\mathbb{E}[A^\top B] = \mathbb{E}[A]^\top \mathbb{E}[B] + \mathrm{tr}(\mathrm{Cov}(A, B)). \tag{20}$$

Thus,

$$\mathbb{E}\langle \tilde{g}_k, \tilde{g}_j \rangle = \langle g_k, g_j \rangle + \mathrm{tr}(\mathrm{Cov}(\tilde{g}_k, \tilde{g}_j)). \tag{21}$$

Expected squared norms. The local gradient satisfies

$$\mathbb{E}\|\tilde{g}_k\|^2 = \|g_k\|^2 + \mathrm{tr}(\mathrm{Var}(\tilde{g}_k)). \tag{22}$$

The global DP gradient yields

$$\begin{aligned} \mathbb{E}\|\tilde{g}^{\mathrm{DP}}\|^2 &= \|g\|^2 + \mathrm{tr}\left(\mathrm{Var}\left(\sum_{j=1}^{K} \frac{n_j}{N} \tilde{g}_j \right) \right) + \mathrm{tr}(\mathrm{Var}(\eta)) \\ &= \|g\|^2 + \mathrm{tr}\left(\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{n_i n_j}{N^2} \mathrm{Cov}(\tilde{g}_i, \tilde{g}_j) \right) \\ &\quad + \mathrm{tr}(\mathrm{Var}(\eta)). \end{aligned} \tag{23}$$

Collecting terms. The mean-gradient components simplify to

$$\|g_k\|^2 - 2\langle g_k, g\rangle + \|g\|^2 = \|g_k - g\|^2. \tag{24}$$

The variance–covariance terms combine to

$$\begin{aligned}\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) - 2\sum_{j=1}^{K}\frac{n_j}{N}\mathrm{tr}(\mathrm{Cov}(\tilde{g}_k, \tilde{g}_j)) \\ + \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{n_i n_j}{N^2}\mathrm{tr}(\mathrm{Cov}(\tilde{g}_i, \tilde{g}_j)).\end{aligned} \tag{25}$$

Adding $\mathrm{tr}(\mathrm{Var}(\eta))$ yields the decomposition

$$\begin{aligned}\mathbb{E}\|\tilde{g}_k - \tilde{g}^{\mathrm{DP}}\|^2 = \|g_k - g\|^2 + \mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) \\ - 2\sum_{j=1}^{K}\frac{n_j}{N}\mathrm{tr}(\mathrm{Cov}(\tilde{g}_k, \tilde{g}_j)) \\ + \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{n_i n_j}{N^2}\mathrm{tr}(\mathrm{Cov}(\tilde{g}_i, \tilde{g}_j)) \\ + \mathrm{tr}(\mathrm{Var}(\eta)),\end{aligned} \tag{26}$$

thereby proving the theorem. □

B.2. Proof of Proposition 1

Proof. Let $g_k = \mathbb{E}_{x\sim P_k}[\nabla\ell(w;x)]$ and $g = \mathbb{E}_{x\sim P}[\nabla\ell(w;x)]$ denote the local and global expected gradients. We adopt a standard smoothness assumption in which the gradient with respect to the input can be locally approximated by a first-order expansion,

$$\nabla\ell(w;x) \approx H(w)(x - \mu_0) + b(w),$$

where $H(w)$ approximates the Hessian and μ_0 is a reference point.

Expected-gradient expansion. Taking expectations under the local and global data distributions yields

$$g_k \approx H(w)(\mu_k - \mu_0) + b(w), \qquad g \approx H(w)(\mu - \mu_0) + b(w),$$

where μ_k and μ denote the local and global means. Subtracting the two expressions gives

$$g_k - g \approx H(w)(\mu_k - \mu),$$

and therefore

$$\|g_k - g\|^2 \le \|H(w)\|_F^2\,\|\mu_k - \mu\|^2.$$

Effect of DP noise on global mean estimation. Each client reports a perturbed mean

$$\tilde{\mu}_k = \mu_k + \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, \sigma_\mu^2 I).$$

The server aggregates them to form

$$\tilde{\mu} = \sum_{j=1}^{K}\frac{n_j}{N}\tilde{\mu}_j = \mu + \epsilon, \qquad \epsilon = \sum_{j=1}^{K}\frac{n_j}{N}\epsilon_j.$$

Since the noise across clients is independent and Gaussian,

$$\mathbb{E}\|\epsilon\|^2 = \sum_{j=1}^{K}\left(\frac{n_j}{N}\right)^2\mathbb{E}\|\epsilon_j\|^2 = \sum_{j=1}^{K}\left(\frac{n_j}{N}\right)^2 d\sigma_\mu^2,$$

where d is the input dimension.

Mean estimation error. We decompose

$$\|\mu_k - \tilde{\mu}\|^2 = \|\mu_k - \mu - \epsilon\|^2 = \|\mu_k - \mu\|^2 + \|\epsilon\|^2 - 2\langle\mu_k - \mu, \epsilon\rangle.$$

The cross-term has zero expectation because ϵ is independent of $\mu_k - \mu$ and has zero mean. Thus,

$$\mathbb{E}\|\mu_k - \tilde{\mu}\|^2 = \|\mu_k - \mu\|^2 + \mathbb{E}\|\epsilon\|^2.$$

Rearranging yields

$$\|\mu_k - \mu\|^2 = \mathbb{E}\|\mu_k - \tilde{\mu}\|^2 - \sum_{j=1}^{K}\left(\frac{n_j}{N}\right)^2 d\sigma_\mu^2.$$

Combining with gradientmean relationship. Substituting into the gradient bound leads to

$$\|g_k - g\|^2 \le \|H(w)\|_F^2\left(\mathbb{E}\|\mu_k - \tilde{\mu}\|^2 - \sum_{j=1}^{K}\left(\frac{n_j}{N}\right)^2 d\sigma_\mu^2\right),$$

which completes the proof. □

B.3. Proof of Proposition 2

Proof. We first relate the variance of the stochastic gradient to input second moments via the Lipschitz property of the sample-wise gradient. Let $\ell(\cdot)$ be the per-sample loss and assume its gradient with respect to model parameters is L-Lipschitz in the input, i.e.

$$\|\nabla_w\ell(w;x) - \nabla_w\ell(w;x')\| \le L\|x - x'\|. \tag{27}$$

For the stochastic gradient $\tilde{g}_k$ computed on client k we then have

$$\mathbb{E}\big[\|\tilde{g}_k - g_k\|^2\big] \le L^2\,\mathbb{E}\big[\|x\|^2\big], \tag{28}$$

where the expectation is taken over the local data sampling on client k and $g_k = \mathbb{E}[\tilde{g}_k]$ denotes the corresponding mean gradient.

Decomposing the second moment of the input around the local mean μ_k gives

$$\mathbb{E}\big[\|x\|^2\big] = \mathbb{E}\big[\|x-\mu_k\|^2\big] + \|\mu_k\|^2. \tag{29}$$

Using the identity $\mathbb{E}[\|x-\mu_k\|^2] = \mathrm{tr}(\Sigma_k)$, where Σ_k is the covariance matrix of client k's data, we obtain

$$\mathbb{E}\big[\|\tilde{g}_k - g_k\|^2\big] \le L^2\big(\mathrm{tr}(\Sigma_k) + \|\mu_k\|^2\big). \tag{30}$$

Since $\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) = \mathbb{E}\|\tilde{g}_k - g_k\|^2$, the above inequality yields the desired bound

$$\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) \le L^2\big(\mathrm{tr}(\Sigma_k) + \|\mu_k\|^2\big), \tag{31}$$

which is Eq. (9).

We now account for DP noise injected into the covariance estimation used for normalization/alignment. Let each client release a perturbed covariance

$$\tilde{\Sigma}_k = \Sigma_k + \xi_k, \qquad \xi_k \sim \mathcal{N}(0, \sigma^2_{\mathrm{DP},k} I), \tag{32}$$

and let the server aggregate these perturbed covariances by weighted averaging:

$$\tilde{\Sigma} = \sum_{j=1}^{K} \frac{n_j}{N} \tilde{\Sigma}_j = \sum_{j=1}^{K} \frac{n_j}{N} \Sigma_j + \eta_{\mathrm{agg}}, \tag{33}$$

where $\eta_{\mathrm{agg}} = \sum_{j=1}^{K} (n_j/N)\xi_j$ denotes the aggregated noise. Because the ξ_j are zero-mean, independent Gaussian perturbations, the expected squared Frobenius norm of the aggregated noise satisfies

$$\mathbb{E}\|\eta_{\mathrm{agg}}\|_F^2 = \sum_{j=1}^{K} \left(\frac{n_j}{N}\right)^2 d_\Sigma \sigma^2_{\mathrm{DP},j}, \tag{34}$$

where d_Σ is the dimension of the vectorized covariance matrix (i.e. the number of entries after flattening).

Replacing Σ_k by its noisy counterpart in the previous Lipschitz-based bound and taking expectation over the DP noise gives

$$\begin{aligned}\mathrm{tr}\big(\mathrm{Var}(\tilde{g}_k^{\mathrm{DP}})\big) &\le L^2\Big(\mathrm{tr}(\Sigma_k) + \|\mu_k\|^2 + \mathbb{E}[\mathrm{tr}(\eta_{\mathrm{agg}})]\Big) \\ &= L^2\Big(\mathrm{tr}(\Sigma_k) + \|\mu_k\|^2 \\ &\quad + \sum_{j=1}^{K} \left(\frac{n_j}{N}\right)^2 d_\Sigma \sigma^2_{\mathrm{DP},j}\Big),\end{aligned}$$

which is exactly the bound stated in Eq. (10). This completes the proof. □

B.4. Proof of Proposition 3

Proof. Let $g_k = \mathbb{E}_{x\sim P_k}[\nabla \ell(w;x)]$ and $g = \sum_{j=1}^{K}(n_j/N) g_j$ denote the local and global gradients. Using the Edgeworth expansion, the expected gradient under distribution P_k admits the approximation

$$g_k \approx g^{\mathrm{Gauss}} + \frac{\gamma_k}{6} M(w),$$

where $M(w)$ is the third-order derivative term of the loss. Similarly,

$$g \approx g^{\mathrm{Gauss}} + \frac{\bar{\gamma}}{6} M(w).$$

Subtracting gives the skewness-induced deviation

$$g_k - g \approx \frac{\gamma_k - \bar{\gamma}}{6} M(w),$$

and therefore

$$\|g_k - g\|^2 \propto |\gamma_k - \bar{\gamma}|^2.$$

Effect of DP noise. Let the DP-perturbed gradients be

$$g_k^{\mathrm{DP}} = g_k + \xi_k, \qquad g^{\mathrm{DP}} = g + \xi,$$

where ξ_k, ξ are zero-mean Gaussian and independent of skewness effects. Then

$$g_k^{\mathrm{DP}} - g^{\mathrm{DP}} = (g_k - g) + (\xi_k - \xi).$$

The cross term vanishes in expectation, so the deterministic skewness contribution remains dominant:

$$\mathbb{E}\|g_k^{\mathrm{DP}} - g^{\mathrm{DP}}\|^2 = \|g_k - g\|^2 + \mathcal{O}(\sigma^2_{\mathrm{DP}}) \propto |\gamma_k - \bar{\gamma}|^2.$$

This completes the proof. □

B.5. Proof of Proposition 4

Proof. Let $x \sim P_k$ with mean μ_k, covariance Σ_k, and kurtosis κ_k. For smooth loss $\ell(w;x)$, linearizing the stochastic gradient around μ_k gives

$$\nabla \ell(w;x) \approx A(w)(x - \mu_k) + b(w),$$

so the Gaussian reference variance satisfies

$$\mathrm{Var}^{\mathrm{Gauss}}(\nabla \ell(w;x)) \propto \mathrm{tr}(\Sigma_k).$$

Kurtosis correction. Applying GramCharlier expansion and retaining the leading non-Gaussian term gives the variance adjustment

$$\mathrm{Var}(\nabla \ell(w;x)) \approx \mathrm{Var}^{\mathrm{Gauss}}(\nabla \ell(w;x)) \left(1 + \frac{\kappa_k - 3}{4}\right).$$

Taking traces yields

$$\mathrm{tr}(\mathrm{Var}(\nabla \ell(w;x))) \propto \left(1 + \frac{\kappa_k - 3}{4}\right) \mathrm{tr}(\Sigma_k).$$

Effect of DP noise. With DP-perturbed gradient

$$\tilde{g}_k^{\mathrm{DP}} = g_k + \epsilon_{\mathrm{DP}},$$

the additive isotropic noise contributes independent variance

$$\mathrm{tr}(\mathrm{Var}(\epsilon_{\mathrm{DP}})) = d\sigma_{\mathrm{DP}}^2.$$

Combining the kurtosis-induced multiplicative factor with the DP noise term gives the desired relationship. □

B.6. Proof of Theorem 2

Proof. The proof follows the standard descent analysis under L-smoothness, combined with a decomposition of the stochastic update direction into its bias and variance components, and an explicit upper bound on the statistical disparity term.

Descent inequality under L-smoothness. For any L-smooth non-convex objective F, the update $w_{t+1} = w_t - \eta d_t$ satisfies

$$F(w_{t+1}) \le F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{L}{2}\|w_{t+1} - w_t\|^2. \tag{35}$$

Substituting the update rule gives

$$F(w_{t+1}) \le F(w_t) - \eta \langle \nabla F(w_t), d_t \rangle + \frac{L\eta^2}{2}\|d_t\|^2. \tag{36}$$

Expectation and bias–variance decomposition. Taking expectation of (36) and expanding $d_t = \mathbb{E}[d_t] + (d_t - \mathbb{E}[d_t])$ yields

$$\begin{aligned}\mathbb{E}[F(w_{t+1})] \le \mathbb{E}[F(w_t)] &- \eta \langle \nabla F(w_t), \mathbb{E}[d_t] \rangle \\ &+ \frac{L\eta^2}{2}\Big(\|\mathbb{E}[d_t]\|^2 + \mathbb{E}\|d_t - \mathbb{E}[d_t]\|^2\Big).\end{aligned} \tag{37}$$

Introduce and subtract $\eta\|\nabla F(w_t)\|^2$ to reveal a descent term:

$$\begin{aligned}\mathbb{E}[F(w_{t+1})] \le \mathbb{E}[F(w_t)] &- \eta\|\nabla F(w_t)\|^2 \\ &+ \eta \langle \nabla F(w_t), \nabla F(w_t) - \mathbb{E}[d_t] \rangle \\ &+ \frac{L\eta^2}{2}\Big(\|\mathbb{E}[d_t]\|^2 + \mathbb{E}\|d_t - \mathbb{E}[d_t]\|^2\Big).\end{aligned} \tag{38}$$

Applying the Cauchy–Schwarz and Young inequalities gives

$$\begin{aligned}\langle \nabla F(w_t), \nabla F(w_t) - \mathbb{E}[d_t] \rangle \le \frac{1}{2}\|\nabla F(w_t)\|^2 \\ + \frac{1}{2}\|\nabla F(w_t) - \mathbb{E}[d_t]\|^2.\end{aligned} \tag{39}$$

Combining terms yields the per-iteration inequality

$$\begin{aligned}\mathbb{E}[F(w_{t+1})] \le \mathbb{E}[F(w_t)] &- \frac{\eta}{2}\|\nabla F(w_t)\|^2 \\ &+ \frac{\eta}{2}\|\nabla F(w_t) - \mathbb{E}[d_t]\|^2 \\ &+ \frac{L\eta^2}{2}\Big(\|\mathbb{E}[d_t]\|^2 + \mathbb{E}\|d_t - \mathbb{E}[d_t]\|^2\Big).\end{aligned} \tag{40}$$

Bounding the bias term via statistical disparity. Using Jensen's inequality,

$$\|\nabla F(w_t) - \mathbb{E}[d_t]\|^2 \le \frac{1}{K}\sum_{k=1}^{K}\|\nabla F(w_t) - g_k(w_t)\|^2. \tag{41}$$

The deviation $\|\nabla F - g_k\|$ can be bounded by discrepancy of the empirical statistics of client k, giving

$$\begin{aligned}\|\nabla F(w_t) - \mathbb{E}[d_t]\|^2 &\le \frac{1}{K}\sum_{k=1}^{K}\Big(L_\mu^2\Delta_{\mu,k} + L_\sigma^2\|\Sigma_k - \Sigma\|_F^2 \\ &\quad + L_\gamma|\gamma_k - \gamma| + L_\kappa|\kappa_k - \kappa|\Big) \\ &= \Gamma_{\mathrm{stat}}.\end{aligned} \tag{42}$$

Bounding the variance term and DP noise. The variance of d_t decomposes as

$$\mathbb{E}\|d_t - \mathbb{E}[d_t]\|^2 = \mathbb{E}\|\tilde{g}_t - \mathbb{E}[\tilde{g}_t]\|^2 + \mathbb{E}\|\varepsilon_t\|^2, \tag{43}$$

where $\tilde{g}_t$ is the stochastic mini-batch gradient and ε_t is DP Gaussian noise.

Define the minimal client-side variance by $\sigma_{\min}^2$, and note that DP noise satisfies

$$\mathbb{E}\|\varepsilon_t\|^2 = \sigma_{\mathrm{DP}}^2 = \frac{8C^2\log(1.25/\delta)}{\epsilon^2}. \tag{44}$$

Thus

$$\mathbb{E}\|d_t - \mathbb{E}[d_t]\|^2 \le \sigma_{\min}^2 + \sigma_{\mathrm{DP}}^2. \tag{45}$$

Final convergence rate. Summing (40) over $t = 0, \ldots, T-1$ and telescoping the left-hand side gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w_t)\|^2 \le \frac{2\Delta_F}{\eta T} + \eta L\big(\sigma_{\min}^2 + \Gamma_{\mathrm{stat}} + \sigma_{\mathrm{DP}}^2\big), \tag{46}$$

which matches the claim in Theorem 2. □