Supplementary Material for Federated Label-Noise Learning with Local Diversity Product Regularization

I. Proof of Theorem 1

For the optimization problem

$$\min_{\mathbf{w},\tilde{\mathbf{T}}} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(\tilde{y}_i, \tilde{\mathbf{T}}^\top f(\mathbf{x}_i;\mathbf{w})) - \lambda \log\det(\mathbf{P}\mathbf{P}^\top + \mathbf{I}) \right\}, \tag{1}$$

its optimal solution $(\mathbf{w}^*, \mathbf{T})$ satisfies

$$\tilde{\mathbf{P}} = \mathbf{T}^\top \mathbf{P}^*,$$

where the matrix of noisy class posterior $\tilde{\mathbf{P}}$ is a constant matrix determined by the noisy dataset. Moreover, $\mathbf{P}^*$ satisfies

$$\mathbf{P}^* = \arg\max_{\mathbf{w}} \det(\mathbf{P}\mathbf{P}^\top), \tag{2}$$

where $\mathbf{P} = [f(\mathbf{x}_1;\mathbf{w}), ..., f(\mathbf{x}_N;\mathbf{w})]$.

According to the theory of transition matrix decomposition in [1], for any solution $(\mathbf{w}, \tilde{\mathbf{T}})$ to $\tilde{\mathbf{P}} = \tilde{\mathbf{T}}^\top \mathbf{P}$, its $\mathbf{P}$ can be written as

$$\mathbf{P} = \mathbf{V}^\top \mathbf{P}^*, \tag{3}$$

where $\mathbf{V}$ is a contraction matrix and each row of $\mathbf{V}$ is a probability vector. Based on (2), it can be obtained that

$$\begin{aligned} \det(\mathbf{P}^*\mathbf{P}^{*\top}) &\geq \det(\mathbf{P}\mathbf{P}^\top) \\ &= \det(\mathbf{V}^\top \mathbf{P}^*\mathbf{P}^{*\top}\mathbf{V}) \\ &= \det(\mathbf{V})^2 \det(\mathbf{P}^*\mathbf{P}^{*\top}). \end{aligned}$$

Thus, $\det(\mathbf{V}) \leq 1$.

Similar to

$$\mathbf{P}\mathbf{P}^\top = \sum_{m=1}^{M} \mathbf{P}_m \mathbf{P}_m^\top, \tag{4}$$

$\mathbf{P}^*\mathbf{P}^{*\top}$ can be decomposed as

$$\mathbf{P}^*\mathbf{P}^{*\top} = \sum_{m=1}^{M} \mathbf{P}_m^* \mathbf{P}_m^{*\top}. \tag{5}$$

Substitute equation (3) and equation (5) into equation (4), then

$$\mathbf{P}\mathbf{P}^\top = \sum_{m=1}^{M} \mathbf{P}_m \mathbf{P}_m^\top = \sum_{m=1}^{M} \mathbf{V}^\top \mathbf{P}_m^* \mathbf{P}_m^{*\top} \mathbf{V}.$$

If $\mathbf{P}_m^*\mathbf{P}_m^{*\top}$ is a full rank matrix, then

$$\begin{aligned} \det(\mathbf{P}_m\mathbf{P}_m^\top) &= \det(\mathbf{V}^\top \mathbf{P}_m^*\mathbf{P}_m^{*\top}\mathbf{V}) \\ &= \det(\mathbf{V})^2 \det(\mathbf{P}_m^*\mathbf{P}_m^{*\top}) \\ &\leq \det(\mathbf{P}_m^*\mathbf{P}_m^{*\top}). \end{aligned}$$

As a result, $\det(\mathbf{P}_m\mathbf{P}_m^\top + \mathbf{I}) \leq \det(\mathbf{P}_m^*\mathbf{P}_m^{*\top} + \mathbf{I})$, and

$$\prod_{m=1}^{M} \det(\mathbf{P}_m\mathbf{P}_m^\top + \mathbf{I}) \leq \prod_{m=1}^{M} \det(\mathbf{P}_m^*\mathbf{P}_m^{*\top} + \mathbf{I}). \tag{6}$$

If $\mathbf{P}_m^*\mathbf{P}_m^{*\top}$ is not a full rank matrix, then the product of its non-zero eigenvalues is larger than or equal to that of $\mathbf{P}_m\mathbf{P}_m^\top$. By adding an identity matrix to $\mathbf{P}_m^*\mathbf{P}_m^{*\top}$, the obtained matrix becomes a full rank matrix. The same result as equation (6) can then be obtained. Hence,

$$\mathbf{P}^* = \arg\max_{\mathbf{w}\in\mathcal{W}} \prod_{m=1}^{M} \det(\mathbf{P}_m\mathbf{P}_m^\top + \mathbf{I}),$$

and the proof is completed.

TABLE I
HYPERPARAMETER SETTINGS FOR FEDLNL.

Hyperparameters	**CIFAR-10**							**CIFAR-100**			**Clothing1M** [11]
	flip-0.2	flip-0.4	flip-0.45	sym-0.2	sym-0.4	sym-0.5	asym-0.4	flip-0.2	flip-0.4	flip-0.45	real-world
λ	0.020	0.010	0.010	0.010	0.010	0.010	0.020	0.020	0.020	0.020	0.050
α_1	0.990	0.990	0.990	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.990
α_2	0.010	0.010	0.020	0.030	0.030	0.080	0.010	0.100	0.100	0.100	0.030

TABLE II
TEST ACCURACIES (%) OF THE SELECTED SCHEMES OVER **CIFAR-10** DATASET UNDER DIFFERENT SETTINGS OF LOCAL TRAINING SAMPLES (PAIR-FLIPPING NOISE, NOISE RATE 0.4).

Schemes	5000 samples/clients	500 samples/clients
FedLSR	82.7	78.3
FedCorr	86.3	75.3
RoFL	89.9	80.4
VolMinNet-FL	90.4	69.1

II. DETAILED EXPERIMENTAL SETUP

A. *Synthesis noise patterns*

Three types of label-noise are used in the experiments: pair flipping (denoted as flip)[2], symmetry (denoted as sym) [3], and asymmetry (denoted as asym) [4]. Pair flipping noise is generated by replacing the clean label i with the noisy label $(i+1)$ for a percentage of training data[1], where the percentage is determined by the noise rate. Symmetric noise is generated by randomly replacing the clean label i with all possible labels but i for a percentage of training data. Asymmetric noise is designed by simulating the structure of real-world label-noise, where a clean label is only replaced by a noisy label of similar classes (e.g. dog↔cat and deer→horse).

The noise rate of pair flipping noise and symmetry noise is selected from $\{0.2, 0.4, 0.45\}$ and $\{0.2, 0.4, 0.45\}$, respectively, while the noise rate of asymmetry noise is set to 0.4. To insert label-noise into the clean training data, an NTM is first generated, based on the selected type of label-noise and the selected value of noise rate. According to the NTM, the training data of **CIFAR-10** and **CIFAR-100** are then manually corrupted.

B. *Hyperparameter settings*

The compared schemes can be divided into two groups. The first group represents the schemes extended to the FL framework, i.e., S-adaptation-FL, VolMinNet-FL, and DivideMix-FL. The second group represents the original FL schemes that tackles label-noise issues, i.e, RoFL [5], FedLSR [6], and FedCorr [7].

For the schemes in the first group, they have two types of hyperparameters: the first type is their original hyperparameters proposed in the CL setting, and the second type is the hyperparameters introduced by the FL framework. For the first type of hyperparameters, we refer to their original papers [8], [9], [10] to determine the settings for their hyperparameters.

As for the second type of hyperparameters, the number of local iterations is set to 3, 3, and 5 for S-adaptation-FL, VolMinNet-FL, and DivideMix-FL, respectively. The rest hyperparameters related to the FL framework are the same as that of FedLNL, e.g., the learning rate and the batch size.

Since the schemes in the second group are designed for the FL framework, we refer to their original papers [5], [6], [7] to determine the settings for their hyperparameters.

In our scheme FedLNL, except for the hyperparameters related to the FL framework, there are three hyperparameters: the weight λ for the local diversity product (LDP) regularizer and the two weights α_1 and α_2 in the update rule of the local concentration matrix $\mathbf{D}_m$

$$\mathbf{D}_m \leftarrow \alpha_1 \mathbf{D}_m + \alpha_2 \mathbf{C}_m.$$

The values of these three hyperparameter are presented in Table I.

[1] For label 9 in **CIFAR-10** and label 99 in **CIFAR-100**, they are replaced by label 0

TABLE III
TEST ACCURACIES (%) OF THE SELECTED SCHEMES OVER **CIFAR-10** DATASET UNDER DIFFERENT SETTINGS OF LOCAL TRAINING SAMPLES (PAIR-FLIPPING NOISE, NOISE RATE 0.4).

Schemes	5000 samples/clients	500 samples/clients
RoFL	89.9	80.4
VolMinNet-FL	90.4	69.1
Accurate NTM	90.7	89.2

REFERENCES

[1] Y. Zhang, G. Niu, and M. Sugiyama, "Learning noise transition matrix from only noisy labels via total variation regularization," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 12 501–12 512.

[2] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2233–2241.

[4] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] S. Yang, H. Park, J. Byun, and C. Kim, "Robust federated learning with noisy labels," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 35–43, 2022.

[6] X. Jiang, S. Sun, Y. Wang, and M. Liu, "Towards federated learning against noisy labels via local self-regularization," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 862–873.

[7] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "FedCorr: Multi-stage federated learning for label noise correction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 184–10 193.

[8] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proceedings of the International Conference on Learning Representations*, 2017.

[9] X. Li, T. Liu, B. Han, G. Niu, and M. Sugiyama, "Provably end-to-end label-noise learning without anchor points," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 6403–6413.

[10] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *Proceedings of the International Conference on Learning Representations*, 2020.

[11] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2691–2699.