# Appendix

## A ResNet Experiments

We perform experiments for LP, FT, FedNCM and FedNCM+FT using the ResNet18 model using the FedAvg algorithm and the CIFAR-10 and Flowers102 datasets. Results are summarized in Table A where we observe that FedNCM performance is better by almost 13% compared to squeezenet, while FT performance is degraded compared to squeezenet. We hypothesis this is due to the challenges of deeper networks in heterogenous federated learning. For the Flowers102 dataset, FedNCM and FedNCM+FT produce the best results by far. Additionally, for flowers FedNCM outperformed all other methods. The variance between runs using ResNet18 is much higher than was observed for SqueezeNet, FedNCM appears to help stabilize the results since it provides the most consistency by for both datasets.

## B Hyperparameter Settings

For CIFAR-10, CUB, Stanford Cars and Eurosat datasets the learning rates for the FedAvg algorithm were tuned via a grid search over learning rates $\{0.1, 0.07, 0.05, 0.03, 0.01, 0.007, 0.005, 0.003, 0.001\}$. For Flowers102, based on preliminary analysis we used lower learning rates were tuned over learning rates $\{0.01, 0.007, 0.005, 0.003, 0.001, 0.0007, 0.0005, 0.0003, 0.0001\}$.

Prior work on federated learning with pre-trained models has indicated that for FedADAM lower global learning rates and higher client learning rates were more effective. As a result for CIFAR-10 and Flowers the client learning rate was tuned over $\{1, 0.1, 0.01, 0.001, 0.0001\}$ and the server learning rate was tuned over $\{0.001, 0.0001, 0.00001, 0.000001\}$, each combination of server and client learning rates were tried. For

## C Extended Accuracy Comparison Figures

Figure 1 is the extended version of Figure 1 in the main body of the paper. In the paper we truncate the number of round displayed for the random setting since random requires many more rounds to converge than the other methods. Figure 1 shows these same figures with the entirety of the training rounds displayed for the random setting.

## D Compute

We use a combination of NVIDIA A100-SXM4-40GB, NVIDIA RTX A4500, Tesla V100-SXM2-32GB and Tesla P100-PCIE-12GB GPUs for a total of 1.1 GPU years . In addition to the experiments reported in the paper, this includes preliminary experiments and hyperparameter searches.

| Dataset | FedNCM | FedNCM + FT | FT+Pretrain | LP+Pretrain |
|---|---|---|---|---|
| CIFAR-10 | $77.74 \pm 0.05$ | $79.05 \pm 1.31$ | $77.87 \pm 4.07$ | $74.73 \pm 3.03$ |
| Flowers102 | $74.13 \pm 0.31$ | $74.1 \pm 0.26$ | $34.41 \pm 10.16$ | $25.35 \pm 2.59$ |

Table 1: ResNet18 model performance for FedAvg. As with Squeezenet, FedNCM+FT continues to outperforms in all cases.

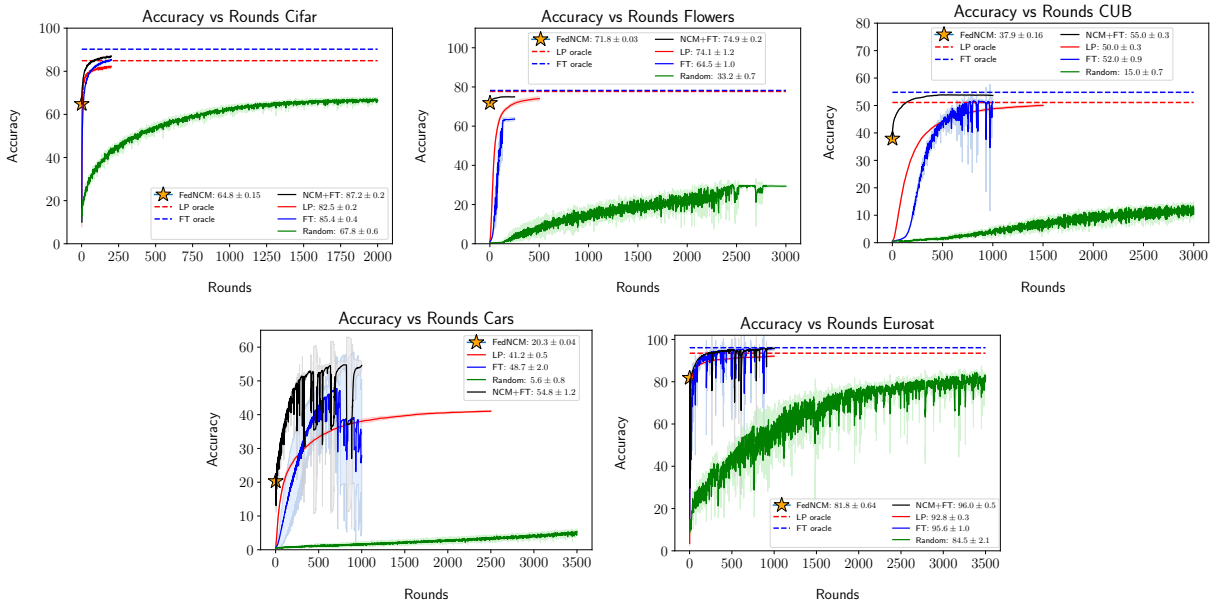


Figure 1: The full training of Random baseline corresponding to Figure 1 in the paper is shown. We observer Random is always very far from the other baseliens and converges slowly.