Supplementary Document

The supplementary material is organized as follows: Section 1 provides the convergence of theory; Section 2 studies the effect of the amount of data that each client owns on its benefit from CL; in Section 3 additional experiments are provided to evaluate the effect of pacing function and its parameters in IID and non-IID FL; Section 4 presents an ablation study on the effect of the level of data heterogeneity; Section 5 studies the correlation between the ordering based learning and the level of statistical data heterogeneity on CIFAR-100 and finally, Section 6 presents the related work to this paper; and finally, Section 7 contains implementation details.

1. Convergence Theory

1.1. Proofs of the Main Theorems - Strongly Convex Problems

Lemma 1 *The stochastic gradient second moment satisfies:*

$$\mathbb{E}[\|g_k^{(t,j)}\|^2] \leq 2(3+2M)L(f(\theta_k^{(t,j)}) - f^*) + (3+2M)B^{(t,j)} + 3\sigma^2$$

Proof. Follows from,

$$\begin{aligned}\mathbb{E}[\|g_k^{(t,j)}\|^2] &\leq 3\|\nabla f(\theta_k^{(t,j)})\|^2 + 3\|b_k^{(t,j)}(\theta_k^{(t,j)})\|^2 \\ &\quad +3\mathbb{E}[\|n_k^{(t,j)}(\theta_k^{(t,j)}, \xi_k^{(t,j)})\|^2] \\ &\leq (3+2M)\|\nabla f(\theta_k^{(t,j)})\|^2 + (3+2M)B^{(t,j)} + 3\sigma^2 \\ &\leq 2(3+2M)L(f(\theta_k^{(t,j)}) - f^*) + (3+2M)B^{(t,j)} + 3\sigma^2\end{aligned}$$

where in the last line we used $\nabla f(\theta^*) = 0$ and L-smoothness. ■

Define,

$$g^{(t,j)} = \frac{1}{|S_t|}\sum_{k\in S_t} g_k^{(t,j)}, \;\; \bar{g}^{(t,j)} = \frac{1}{|S_t|}\sum_{k\in S_t} \nabla f(x_k^{(t,j)})$$

Note that Assumption 1, in particular (4) and (4) imply that $\mathbb{E}[g_k^{(t,j)}] = \bar{g}^{(t,j)}$.

The next Lemma is similar to [44, Lemma 5] with a simpler proof of simple adding the terms across agents up using the previous result.

Lemma 2

$$\mathbb{E}[\|g^{(t,k)} - \bar{g}^{(t,j)}\|^2] \leq \frac{(3+2M)}{Q^2}\sum_{k\in S^{(t)}}\Big[2L(f(\theta_k^{(t,j)}) - f^*) + B^{(t,j)}\Big] + \frac{3\sigma^2}{Q}$$

The next Lemma is similar to [44, Lemma 6] which in turn follows [45, Lemma 2.1]. Consider the sequence,

$$\hat{\theta}^{(t,j+1)} = \hat{\theta}^{(t,j)} - \alpha^{(t,j)} g^{(t,j)}$$

and note that by this construction $\hat{x}^{(t,J)} = \hat{x}^{(t+1,0)}$. The proof is a straightforward application of strong convexity. It holds that,

Lemma 3

$$\begin{aligned}\|\hat{\theta}^{(t,j)} - \alpha^{(t,j)}\bar{g}^{(t,j)} - \theta^*\|^2 &\leq \|\hat{\theta}^{(t,j)} - \theta^*\|^2 \\ &+\frac{2\alpha^{(t,j)}}{Q}\sum_{k\in S^{(t)}}\Big[(\alpha^{(t,j)}L - 1/2)(f(\theta_k^{(t,j)}) - f(\theta^*)) \\ &\qquad -\tfrac{\mu}{2}\|\theta_k^{(t,j)} - \theta^*\|^2\Big] \\ &+\frac{2\alpha^{(t,j)}L}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)} - \theta_k^{(t,j)}\|^2\end{aligned}$$

Finally we obtain our first derivation of expected convergence below.

Lemma 4 *Let $\bar{L} := (L + (3+2M)2L/Q)$ and assume that $\alpha^{(t,j)} \leq \frac{1}{4\bar{L}}$. It holds that the expected distance of the average parameter to the solution satisfies the recursion,*

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_k^{(t,j+1)} - \theta^*\|^2] &\leq \left(1 - \alpha^{(t,j)}\mu\right)\|\hat{\theta}^{(t,j)} - \theta^*\|^2 \\ &-\frac{\alpha^{(t,j)}}{2}\Big(f(\hat{\theta}^{(t,j)} - f(\theta^*)\Big) \\ &+\frac{2\alpha^{(t,j)}L}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)} - \theta_k^{(t,j)}\|^2 \\ &+\frac{(3+2M)(\alpha^{(t,j)})^2B^{(t,j)}}{Q} + \frac{3\sigma^2(\alpha^{(t,j)})^2}{Q}\end{aligned}$$

Proof. Using the previous set of results,

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_k^{(t,j+1)} - \theta^*\|^2] &\leq \|\hat{\theta}^{(t,j)} - \alpha^{(t,j)}\bar{g}^{(t,j)} - \theta^*\|^2 \\ &\quad +(\alpha^{(t,j)})^2\mathbb{E}[\|g^{(t,j)} - \bar{g}^{(t,j)}\|^2] \\ &\leq \|\hat{\theta}^{(t,j)} - \theta^*\|^2 \\ &+\frac{2\alpha^{(t,j)}}{Q}\sum_{k\in S^{(t)}}\Big[(\alpha^{(t,j)}L - 1/2)(f(\theta_k^{(t,j)}) - f(\theta^*)) \\ &\qquad -\tfrac{\mu}{2}\|\theta_k^{(t,j)} - \theta^*\|^2\Big] \\ &+\frac{2\alpha^{(t,j)}L}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)} - \theta_k^{(t,j)}\|^2 \\ &+\frac{(3+2M)(\alpha^{(t,j)})^2}{Q^2}\sum_{k\in S^{(t)}}\Big[2L(f(\theta_k^{(t,j)} - f(\theta^*)) + B^{(t,j)}\Big] \\ &+\frac{3\sigma^2(\alpha^{(t,j)})^2}{Q} \\ &\leq \|\hat{\theta}^{(t,j)} - \theta^*\|^2 \\ &+\frac{2\alpha^{(t,j)}}{Q}\sum_{k\in S^{(t)}}\Big[(\alpha^{(t,j)}\bar{L} - 1/2)(f(\theta_k^{(t,j)}) - f(\theta^*)) \\ &\qquad -\tfrac{\mu}{2}\|\theta_k^{(t,j)} - \theta^*\|^2\Big] \\ &+\frac{2\alpha^{(t,j)}L}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)} - \theta_k^{(t,j)}\|^2 \\ &+\frac{(3+2M)(\alpha^{(t,j)})^2B^{(t,j)}}{Q} + \frac{3\sigma^2(\alpha^{(t,j)})^2}{Q}\end{aligned}$$

By assumption $\alpha^{(t,j)}\bar{L} - 1/2 \leq -\frac{1}{4}$ and then applying Jensen's inequality we have $\frac{1}{Q}\sum_{k\in S^{(t)}}\Big[-\frac{1}{4}(f(\theta_k^{(t,j)}) - f(\theta^*)) - \frac{\mu}{2}\|\theta_k^{(t,j)} - \theta^*\|^2\Big] \leq -\Big(\frac{1}{4}(f(\hat{\theta}^{(t,j)}) - f(\theta^*)) + \frac{\mu}{2}\|\hat{\theta}^{(t,j)} - \theta^*\|^2\Big)$. Plugging this expression into the last displayed equation, the conclusion follows. ■

Next, from Lemma 1 we can conclude that

$$\begin{aligned}&\tfrac{1}{Q}\sum_{k\in S^{(t)}}\mathbb{E}\left[\left\|g_k^{(t,j)}-g^{(t,j)}\right\|^2\right]\\&\leq\tfrac{2(3+2M)L}{Q}\sum_{k\in S^{(t)}}\left(f(\theta_k^{(t,j)})-f(\theta^*)\right)\\&\quad+(3+2M)B^{(t,j)}+3\sigma^2\end{aligned}\tag{5}$$

Next we derive a recursion on the average parameter deviation.

Lemma 5 *Let $\mu>0$. The average iterate deviation satisfies the bound,*

$$\begin{aligned}&\mathbb{E}\left[\tfrac{1}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j+1)}-\theta_k^{(t,j+1)}\|^2\right]\\&\leq(1-\alpha^{(t,j)}\mu/2)\mathbb{E}\left[\tfrac{1}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)}-\theta_k^{(t,j)}\|^2\right]\\&+\tfrac{2(3+2M)L(\alpha^{(t,j)})^2}{Q}\sum_{k\in S^{(t)}}\left(f(\theta_k^{(t,j)})-f(\theta^*)\right)\\&+\alpha^{(t,j)}\left(\alpha^{(t,j)}(3+2M)+B^{(t,j)}/\mu\right)B^{(t,j)}+3(\alpha^{(t,j)})^2\sigma^2\end{aligned}$$

Proof. Indeed, compute directly,

$$\begin{aligned}&\mathbb{E}\left[\tfrac{1}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j+1)}-\theta_k^{(t,j+1)}\|^2\right]\\&\leq\mathbb{E}\left[\tfrac{1}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)}-\theta_k^{(t,j)}\|^2\right]\\&\quad+\tfrac{(\alpha^{(t,j)})^2}{Q}\sum_{k\in S^{(t)}}\mathbb{E}\left[\|g^{(t,j)}-g_k^{(t,j)}\|^2\right]\\&\quad-\tfrac{2\alpha^{(t,j)}}{Q}\sum_{k\in S^{(t)}}\mathbb{E}\left[\left\langle\theta_k^{(t,j)}-\hat{\theta}^{(t,j)},g_k^{(t,j)}-g^{(t,j)}\right\rangle\right]\end{aligned}$$

For the second term in the above expression we can apply (5). For the third, we note that,

$$\begin{aligned}&-\sum_{k\in S^{(t)}}\mathbb{E}\left[\left\langle\theta_k^{(t,j)}-\hat{\theta}^{(t,j)},g_k^{(t,j)}-g^{(t,j)}\right\rangle\right]\\&=-\sum_{k\in S^{(t)}}\left\langle\theta_k^{(t,j)}-\hat{\theta}^{(t,j)},\nabla f(\theta_k^{(t,j)})+b_k^{(t,j)}(\theta_k^{(t,j)})\right\rangle\\&+\sum_{k\in S^{(t)}}\left\langle\theta_k^{(t,j)}-\hat{\theta}^{(t,j)},\tfrac{1}{Q}\sum_{k\in S^{(t)}}\left[\nabla f(\theta_k^{(t,j)})+b_k^{(t,j)}(\theta_k^{(t,j)})\right]\right\rangle\\&=\sum_{k\in S^{(t)}}\left\langle\hat{\theta}^{(t,j)}-\theta_k^{(t,j)},\nabla f(\theta_k^{(t,j)})+b_k^{(t,j)}(\theta_k^{(t,j)})\right\rangle\\&\leq\sum_{k\in S^{(t)}}\left[f(\hat{\theta}^{(t,j)})-f(\theta_k^{(t,j)})-\tfrac{\mu}{2}\|\theta_k^{(t,j)}-\theta^{(t,j)}\|^2\right]\\&\quad+\sum_{k\in S^{(t)}}B^{(t,j)}\|\theta_k^{(t,j)}-\theta^{(t,j)}\|\end{aligned}$$

where we used strong convexity in the inequality. Applying Young's inequality to obtain $B^{(t,j)}\|\theta_k^{(t,j)}-\theta^{(t,j)}\|\leq\frac{\mu}{4}\|\theta_k^{(t,j)}-\theta^{(t,j)}\|^2+\frac{1}{\mu}(B^{(t,j)})^2$ yields the final result. ■

Now we want to use the previous Lemma in order to bound the contribution of the average iterate discrepancy to the overall descent appearing in Lemma 4. Taking a sum for a given t, for $j=1,...,J$, we can see that

$$\begin{aligned}&\sum_{j=0}^{J}\mathbb{E}\left[\tfrac{1}{Q}\sum_{k\in S^{(t)}}\|\hat{\theta}^{(t,j)}-\theta_k^{(t,j)}\|^2\right]\\&\leq\sum_{j=0}^{J}\tfrac{2\alpha^{(t,j)}L}{Q}\prod_{l=j}^{J}(1-\alpha^{(t,l)}\mu/2)\\&\quad\Bigg[2\alpha^{(t,j)}(3+2M)L\sum_{k\in S^{(t)}}\left(f(\theta_k^{(t,j)})-f(\theta^*)\right)\\&\quad+\left(\alpha^{(t,j)}(3+2M)+B^{(t,j)}/\mu\right)B^{(t,j)}\\&\quad+3\alpha^{(t,j)}\sigma^2\Bigg]\end{aligned}\tag{6}$$

With that, we proceed with the main result:

Proof. of Theorem 1 From Lemma 4 and (6)

$$\begin{aligned}&\mathbb{E}[\|\hat{\theta}^{(t+1,0)}-\theta^*\|^2]\leq\prod_{j=0}^{J}(1-\alpha^{(t,j)}\mu/2)\|\hat{\theta}^{(t,0)}-\theta^*\|^2\\&+\sum_{j=0}^{J}\tfrac{2\alpha^{(t,j)}L}{Q}\prod_{l=j}^{J}(1-\alpha^{(t,l)}\mu/2)\\&\quad\Bigg[\left(2\alpha^{(t,j)}(3+2M)L-\tfrac{1}{2}\right)\sum_{k\in S^{(t)}}\left(f(\theta_k^{(t,j)})-f(\theta^*)\right)\\&\quad+\left(2\alpha^{(t,j)}(3+2M)+B^{(t,j)}/\mu\right)B^{(t,j)}\\&\quad+6\alpha^{(t,j)}\sigma^2\Bigg]\end{aligned}$$

Noting that the assumption on the Theorem implies that the term involving the objective value difference is negative, we obtain the statement of the main result. ■

1.2. Nonconvex Objectives

Proof. of Theorem 2 As standard, we begin by applying the Descent Lemma across subsequent averaging steps.

$$\begin{aligned}f(\theta^{(t+1,0)})-f(\theta^{(t,0)})&\leq\left\langle\nabla f(\theta^{(t,0)}),\theta^{(t+1,0)}-\theta^{(t,0)}\right\rangle\\&\quad+\tfrac{L}{2}\|\theta^{(t+1,0)}-\theta^{(t,0)}\|^2\\&\leq-\left\langle\nabla f(\theta^{(t,0)}),\tfrac{1}{Q}\sum_{k\in S^{(t)}}\sum_{j=0}^{J}\alpha^{(t,j)}g_k^{(t,j)}\right\rangle\\&\quad+\tfrac{L}{2}\left\|\tfrac{1}{Q}\sum_{k\in S^{(t)}}\sum_{j=0}^{J}\alpha^{(t,j)}g_k^{(t,j)}\right\|^2\end{aligned}$$

Now, we consider the discrepancy of $g_k^{(t,j)}$ to $\nabla f(\theta^{(t,0)})$ to obtain a perturbation from the decrease we expect to get, that we wish to eventually bound relative to said decrease. Specifically, taking total expectations (and implicitly using the tower property):

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{Q}\sum_{k\in S^{(t)}}\sum_{j=0}^{J} g_k^{(t,j)}\right] \\
&= \frac{1}{Q}\sum_{k\in S^{(t)}}\sum_{j=0}^{J}\alpha^{(t,j)}\mathbb{E}\Big[\nabla f(\theta^{(t,0)}) + b_k^{(t,0)}(\theta^{(t,0)}) \\
&\quad -f(\theta^{(t,0)}) - b_k^{(t,0)}(\theta^{(t,0)}) + \nabla f(\theta^{(t,0)},\xi_k^{(t,j)}) \\
&\quad - \nabla f(\theta^{(t,0)},\xi_k^{(t,j)}) + \nabla f(\theta^{(t,j)},\xi_k^{(t,j)})\Big] \\
&= \frac{1}{Q}\sum_{k\in S^{(t)}}\sum_{j=0}^{J}\alpha^{(t,j)}\Big[\nabla f(\theta^{(t,0)}) \\
&\quad -\mathbb{E}\Big[\nabla f(\theta^{(t,0)},\xi_k^{(t,j)}) + \nabla f(\theta^{(t,j)},\xi_k^{(t,j)})\Big]\Big]
\end{aligned}
$$

and so, combining the previous two sets of equations,

$$
\begin{aligned}
f(\theta^{(t+1,0)}) - f(\theta^{(t,0)}) \leq & -\frac{\sum_{j=0}^{J}\alpha^{(t,j)}}{Q}\|\nabla f(\theta^{(t,0)})\|^2 \\
& + \frac{\sum_{j=0}^{J}\left(\alpha^{(t,j)}\sum_{l=j}^{J}\alpha^{(t,l)}\right)}{Q}LG^2 \\
& + \frac{\sum_{j=0}^{J}\left(\alpha^{(t,j)}\right)^2 LG^2}{2Q}
\end{aligned}
$$

from which we obtain the final result. ■

1.3. Convergence Rate Advantages of Curriculum Learning for Centralized Learning

Consider a standard loss function of the least squares form,

$$\mathcal{L}(\theta,\{x_i,y_i\}) = \frac{1}{2N}\sum_{i=1}^{N}(f(\theta,x_i)-y_i)^2$$

Compute the generic form of the Hessian,

$$
\begin{aligned}
\nabla^2_{\theta\theta}\mathcal{L}(\theta,\{x_i,y_i\}) = & \frac{1}{N}\sum_{i=1}^{N}\big(\nabla_\theta f(\theta,x_i)\nabla_\theta f(\theta,x_i)^T \\
& + \nabla^2_{\theta\theta} f(\theta,x_i)(f(\theta,x_i)-y_i)\big)
\end{aligned}
$$

Note that the Fisher information matrix, or Gauss-Newton term $\nabla_\theta f(\theta,x_i)\nabla_\theta f(\theta,x_i)^T$ is expected to be positive definite and independent of each samples loss value, however, the greater the magnitude of the overall loss $(f(\theta,x_i)-y_i)$ the greater the potential influence of the Hessian of the neural network model $\nabla^2_{\theta\theta}f(\theta,x_i)$ on the overall Hessian of the objective function. Thus, inherently, curriculum training makes the initial objective function more convex than otherwise. This has the clear consequence of enabling faster optimization trajectories at the beginning of the training process. This, however, does not explain why the advantage is more pronounced in the case of non-iid data, which is the most striking finding in our experiments.

1.4. Notes of Convergence Insights from Alternative Formulations

In this section, we give a review of the literature on Federated Averaging and the associated convergence guarantees, presenting an analysis of how we expect these to be modified by the introduction of curriculum learning. As such we round out our theoretical understanding of the procedure.

The score of the data samples is based on the server's parameter vector θ_g. Naturally, this approach creates a significant association between the degree of statistical dissimilarity of the data at each client with the training difficulty score used to rank data samples for curriculum learning. So we can safely purport that CL, for non-iid data, results in a level of dissimilarity that increases with the iteration t.

To understand how this affects the convergence, we review a few standard works and study how increasing heterogeneity with the iteration number affects the convergence guarantees.

To begin with, the state of the art in convergence theory of Federated Averaging (or Local SGD) for convex objectives is given, to the best of our knowledge, in [44].

Here the main result of interest is [44, Theorem 5], for which the objective optimality gap is bounded by,

$$\mathbb{E}[f(\theta_T) - f(\theta^*)] \leq \frac{C_1}{\gamma T} + C_2\gamma\sigma^2 + C_3\gamma^2\sigma^2$$

where the variance σ is proportional to the heterogeneity. It can be seen from the convergence theory that the bound changes to, with σ_t iteration dependent,

$$\mathbb{E}[f(\theta_T) - f(x^*)] \leq \frac{C_1}{\gamma T} + C_2\gamma\sum_{t=0}^{T}\sigma_t^2 + C_3\gamma^2\sum_{t=0}^{T}\sigma_t^2$$

suggesting an overall better convergence quality for any given iteration, since we expect $\sigma_t < \sigma$ up until $t = T$, i.e., early iterations generate better accuracy than otherwise.

In regards to nonconvex objectives, which are of course more faithful to the practice of training neural networks, to the best of our knowledge the state of the art in theoretical convergence guarantees for local SGD is given in [39]. There, a notion of gradient similarity is presented,

$$\Lambda(\theta,q) = \frac{\sum_{m=1}^{M} q_m\|\nabla f_m(\theta)\|^2}{\left\|\sum_{m=1}^{M} q_m\nabla f_m(\theta)\right\|^2}$$

and assuming a bound λ on this term, λ does not appear directly in the convergence bounds in [39, Theorem 4.2 and Theorem 4.4] (respectively for the objective satisfying the PL condition and the general case). However, the number of local steps, which they denote as E, (i.e. the number of SGD steps in `ClientUpdate` in Algorithm 1) depends on $E \propto 1/\lambda$, meaning the greater the dissimilarity and the fewer local iterations are permitted to ensure convergence, a net increase in the total number of communications necessary.

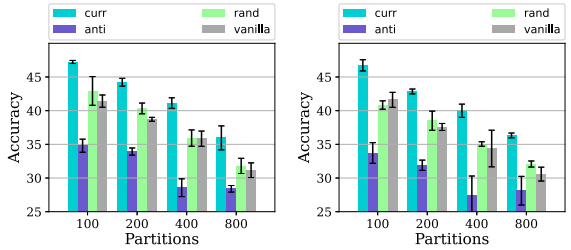

Figure 10: There is no correlation between the amount of data on the client's end and the benefit they gain from ordered learning. The accuracy decreases when the amount of data each client owns is reduced, but it gains the same amount of benefit from curriculum learning with more data. Evaluating the impact of the amount of data each client owns on the accuracy when the clients employ curriculum, anti-curriculum or random ordering during their local training on CIFAR-10 with Non-IID (2) for FedAVg (left), and with Dir(0.05) for Fedprox (right). All curricula use the linear pacing functions with a = 0.8 and b = 0.2. Each experiment is repeated three times for a total of 100 communication rounds with ten local epochs, and the mean and standard deviation for global test accuracy are reported.

The use of the FedProx objective can also be analyzed through the lens of iterate-varying dissimilarity. Considering [46, Theorem 4] we have that with,

$$\rho_t = \frac{1}{\mu} - \bar{\rho}(B_t, \gamma, \mu),\ \bar{\rho}(B_t, \gamma, \mu) = O(B_t)O(\gamma)O(1/\mu)$$

and, with S_t devices chosen at iteration t

$$\mathbb{E}[f(x_{t+1}|S_t] - f(x_t) \leq -\rho_t \|\nabla f(x_t)\|^2$$

and thus with Curriculum training, we see increasing B_t and thus decreasing ρ_t, and thus, again, we shall expect to see initial faster and then gradually slower convergence.

2. Effect of amount of data on clients end

In this section, we are interested in understanding whether the previous conclusions we made for CIFAR10 generalize to both high and low data regimes on the client's end. In particular, we divide the larger dataset into multiples of the number of clients and randomly assign M of those data partitions to the M clients. The larger the number of partitions, the smaller the amount of data on each of the clients. As can be seen from Fig. 10 the amount of data that each client owns has no relationship with the benefit it gains from curriculum learning. In fact, CL ameliorates the classification accuracy performance equally under both lower and higher data regimes on the clients' end.

3. Effect of pacing function and its parameters in IID and Non-IID FL

This subsection complements subsection 3.2 of the main paper, where we evaluated the effect of pacing function and its hyperparameter a when clients train on CIFAR-10 with FedAvg under IID data. Here, we report the results for FedAvg under Non-IID Dir(0.05). The conclusion is similar– Fig. 11 shows that bigger values of a provide better accuracy performance for most of the pacing function families on both extreme IID and Non-IID setting. It is noteworthy that, the observations generalize to other baselines, as discussed in different sections of the paper.

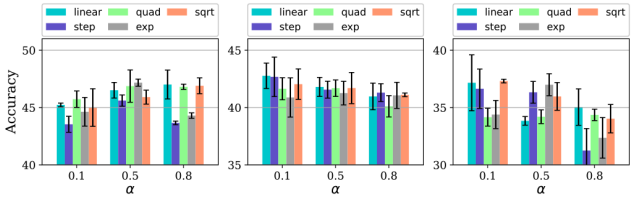

Figure 11: Bigger a values provide better accuracy performance for most of pacing function families and on both IID and Non-IID setting for curriculum learning. But a notable contrast can be seen with random-/anti ordering. The effect of using different pacing function families and their hyperparameter a on the accuracy when the clients employ curriculum, anti-curriculum or random ordering during their local training on CIFAR-10 with Non-IID Dir(0.05) data. The figures from left to right are for curriculum, random, and anti ones.

4. Ablation study on the effect of level of data heterogeneity

In this section, we report the ablation studies on the pathological case of extreme heterogeneity. We have carefully examined the results for the specific pathological where one client has only images of one of the classes (such as cats) and another has only images of one other class (such as dogs). In this severe heterogeneous scenario, the results that reported in Table 3 demonstrate a significant improvement achieved by CL compared to the vanilla approach. CL led to a remarkable improvement of 43%, and 29% respectively over the "`Vanilla`" baseline, while is higher than the Dir (.) Non-IID (less heterogeneous) setup reported in the main paper of about 16%, and 17% for FedProx and FedAvg. These findings further reinforce the claim that "the more heterogeneous the more benefit from CL". Here, we are referring to the relative improvement over the baseline ("`Vanilla`").

Table 3: The benefit of ordered learning under the pathological case of one client having only images of one of the classes (such as cats) and another having only images of one other class (such as dogs) (very high heterogeneity) when clients are trained on CIFAR-10.

Algorithm	Curriculum	Anti	Random	Vanilla
FedProx	21.83 ± 0.36	11.28 ± 0.82	16.21 ± 0.17	15.25 ± 0.81
FedAvg	19.69 ± 0.46	10.59 ± 0.10	13.62 ± 0.38	15.29 ± 0.79

5. Effect of level of heterogeneity

This subsection complements subsection 3.3 of the main paper. In this section, we present further experimental results showing the relationship between ordering-based learning and the level of statistical data heterogeneity. Herein, we are interested in investigating whether the previous conclusions we made for CIFAR-10 generalize to other datasets such as CIFAR-100. The performance of an "expert" model with the same network architecture trained in a standard IID centralized non-federated setting on the dataset is about 53%. Fig. 12 shows the same trend as in CIFAR-10, i.e., *again, we see that as the data from the clients becomes more heterogeneous, the global model benefits more from curriculum learning, resulting in higher performance accuracy when compared to "vanilla" and "anti-/random" learning.* We provided rigorous analysis to explain this phenomenon.

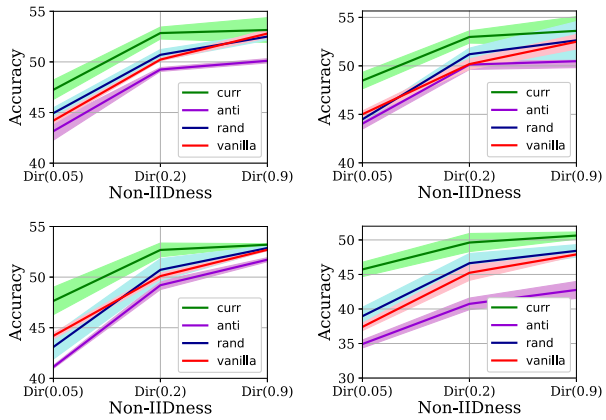

Figure 12: Curriculum-learning helps more when training with more severe data heterogeneity across clients on CIFAR-100. Test accuracy of different baselines when sweeping from extremely Non-IID setting, Dir (0.05) to highly IID setting, Dir(0.9). For each baseline, the average of final global test accuracy is reported. We run each baseline 3 times for 100 communication rounds with 10 local epochs. The figures from left to right, are for FedAvg, Fedprox, Scaffold, and FedNova baselines.

6. Related Work

Early CL formulated the easy-to-hard training paradigm in the context of deep learning [1]. CL determines a sequence of training instances, which in essence corresponds to a list of samples ranked in ascending order of learning difficulty [2]. Samples are ranked according to per-sample loss [21]. In the early steps of training, samples with smaller loss (higher score) are selected, and gradually the subset size over time is increased to cover all the training data. [22] proposed to manually sort the samples using human annotators. Self-paced learning (SPL) [2] chooses the curriculum based on hardness (e.g., per-sample loss) during training. [20] proposes using a consistency score (c-score) calculated based on the consistency of a model in correctly predicting a particular example's label trained on i.i.d. draws of the training set. [47] determines the difficulty of learning an example by the metric of the earliest training iteration, after which the model predicts the ground truth class for that example in all subsequent iterations.

7. Implementation Details

We begin by splitting the dataset into K partitions, and these partitions are distributed among the N clients in the federation. For most experiments $M = 100$ and the partitions are constructed with an input Non-IID Dirichlet distribution with parameter β and using Algorithm 2 with $f_{ord} = 0$, unless otherwise specified. The merits of the Algorithm 2 are detailed in Section 4.3.

At the client, we use an SGD optimizer for training with an exponentially decaying learning $\eta = \eta_0(1 + \alpha * i)^{-b}$, with parameters $\eta_0 = 0.001$, $\alpha = 0.001$, $b = 0.75$ and i is the step index, and a momentum $\rho = 0.9$ and weight decay of $\omega = 5 * 10^{-4}$. The step count i is a parameter local to the clients and is reset at the beginning of each federation round thereby resetting the learning rate back to η_0 for each round of federation. For the ResNet models however, we do not use the exponential decay learning rate and set $b = 0$ with $\eta_0 = 0.01$, and weight decay $\omega = 0$, due to our observation that these values empirically work well.

A small batch size of $bs_{data} = 10$ is used on the server. At each client, we use the local epochs $n_{epoch} = 10$, which, together with the client data partition size, determines the number of local steps at the clients between two global model averaging steps of the federation algorithm. The number of communication rounds of federation is $R = 100$ and the client participation rate is $f = 0.1$, unless otherwise specified. Similarly, when performing client curriculum, we use a client batch size of $bs_{client} = 10$.

Certain federated learning algorithms require additional algorithm specific parameters; these are chosen to match the best values reported by the authors in their respective papers. For reproducibility of the experiments, we seed our random number generator with a seed of 202207 at the beginning of each experiment. Each experiment consists of 3 trials, and we report the mean and variance of the results.