Appendix

A Limitations and Future Work

Non-IID distributions. This work only focuses on certain categories of non-identical data, which limit our evaluation of WAFFLE. Further categories like Prior probability shift, Concept drift, and Covariate shift should be tested.

Dataset selection. The benchmark could also have been more realistic with data sets created for this purpose, like FEMNIST or real world data sets, which would benefit from private personalised learning such as medical images.

Hyperparameters. Hyperparameters were also a limitation, as some can be very sensitive (Weight Erosion and WAFFLE) and more rigorous testing methods are necessary to find the optimal settings.

Computational cost. A major limitation was the number of runs we were able to perform due to computational time limitations, especially for CIFAR10. This only allowed us to report the results of one seed which reduces the confidence of the results.

Ω and Ψ. The WAFFLE benchmark is also limited by the functions we chose for Ω and Ψ. Both of these functions were designed to offer versatility in WAFFLE, so that they could be optimized for various contexts. Depending on the category of non-identical data, it might be that some type of Ω and Ψ functions gives better results in general. Therefore, more research should be done to find better functions depending on the problem.

Benchmark. WAFFLE must still be benchmarked against more personalized FL methods. This work only shows two methods (Weight Erosion and APFL). However, the performance of APFL was lower than anticipated. It is possible that this is because APFL adapts poorly to the tasks proposed in our setting and further experimentation is recommended.

Robustness. Finally, we assume honest participation of agents which may not be the case in reality. While tests should be done to verify the strength of WAFFLE against a data poisoning attack, we assume that WAFFLE should not be vulnerable by design as the weighting mechanism should mitigate vulnerability by ignoring malicious participants.

B Figures, Tables and Algorithms

In Algorithm 2, changes w.r.t. SCAFFOLD are written in red ink [Karimireddy et al., 2020].

Table 2: Repartition of labels for the distribution C between our ten agents

	Label									
	0	1	2	3	4	5	6	7	8	9
Agent 0	0	0	0	0.1	0.2	0.4	0.2	0.1	0	0
Agent 1	0	0	0.1	0.2	0.4	0.2	0.1	0	0	0
Agent 2	0	0.1	0.2	0.4	0.2	0.1	0	0	0	0
Agent 3	0.1	0.2	0.4	0.2	0.1	0	0	0	0	0
Agent 4	0.2	0.4	0.2	0.1	0	0	0	0	0	0.1
Agent 5	0.4	0.2	0.1	0	0	0	0	0	0.1	0.2
Agent 6	0.2	0.1	0	0	0	0	0	0.1	0.2	0.4
Agent 7	0.1	0	0	0	0	0	0.1	0.2	0.4	0.2
Agent 8	0	0	0	0	0	0.1	0.2	0.4	0.2	0.1
Agent 9	0	0	0	0	0.1	0.2	0.4	0.2	0.1	0

Algorithm 2. `WAFFLE`: **W**eighted **A**veraging **f**or Personalized **F**ederated **Le**arning

server input : Initial global model $\boldsymbol{x}$ and global control variate $\boldsymbol{c}$, global step-size η_g, and index $i^\star$ of the requesting agent
agent i's input : Initial local control variate $\boldsymbol{c}_i$, and local step-size η_l
output : Model $\boldsymbol{x}$ optimized for agent $i^\star$

$(\alpha^{r-2}, \alpha^{r-1}) \leftarrow (\frac{1}{N}\mathbf{1}_{N\times 1}, \frac{1}{N}\mathbf{1}_{N\times 1})$
for round $r \leftarrow 1$ **to** R **do**
  communicate $(\boldsymbol{x}, \boldsymbol{c})$ to all agents $i \in \mathcal{S}$
  on each agent $i \in \{1, \ldots, N\}$ **in parallel do**
    initialize local model $\boldsymbol{y}_i \leftarrow \boldsymbol{x}$
    for $k \leftarrow 1$ **to** K **do**
      compute mini-batch gradient $g_i(\boldsymbol{y}_i)$
      $\boldsymbol{y}_i \leftarrow \boldsymbol{y}_i - \eta_l(g_i(\boldsymbol{y}_i)) - \boldsymbol{c}_i + \boldsymbol{c})$
    $\boldsymbol{c}_i^+ \leftarrow \boldsymbol{c}_i - \boldsymbol{c} + \frac{1}{K\eta_l}(\boldsymbol{x} - \boldsymbol{y}_i)$
    $(\Delta\boldsymbol{y}_i, \Delta\boldsymbol{c}_i) \leftarrow (\boldsymbol{y}_i - \boldsymbol{x}, \boldsymbol{c}_i^+ - \boldsymbol{c}_i)$
    $\boldsymbol{c}_i \leftarrow \boldsymbol{c}_i^+$
    communicate $(\Delta\boldsymbol{y}_i, \Delta\boldsymbol{c}_i)$ to the server
  $(\bar{\alpha}^r, \alpha^r) \leftarrow$ `CalcWeights(`N, R, $i^\star$, r, $\{\Delta\boldsymbol{y}_i\}$, α^{r-1}, α^{r-2}`)`
  $\alpha^{r-2} \leftarrow \alpha^{r-1}$ and $\alpha^{r-1} \leftarrow \alpha^r$
  $(\Delta\boldsymbol{x}, \Delta\boldsymbol{c}) \leftarrow \sum_{i\in\{1,\ldots,N\}} \bar{\alpha}_i^r\,(\Delta\boldsymbol{y}_i, \Delta\boldsymbol{c}_i)$
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + \eta_g\Delta\boldsymbol{x}$ and $\boldsymbol{c} \leftarrow \boldsymbol{c} + \Delta\boldsymbol{c}$

Table 3: Hyperparameters for personalized FL methods

	MNIST					CIFAR10				
Distr.	A	B	C	A*	B*	A	B	C	A*	B*
`WE` (p_d)	0.006	0.0067	0.006	0.0088	0.0088	1.1	1.1	0.9	0.6	0.6
`WAFFLE` ($\Delta\Omega$)	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2

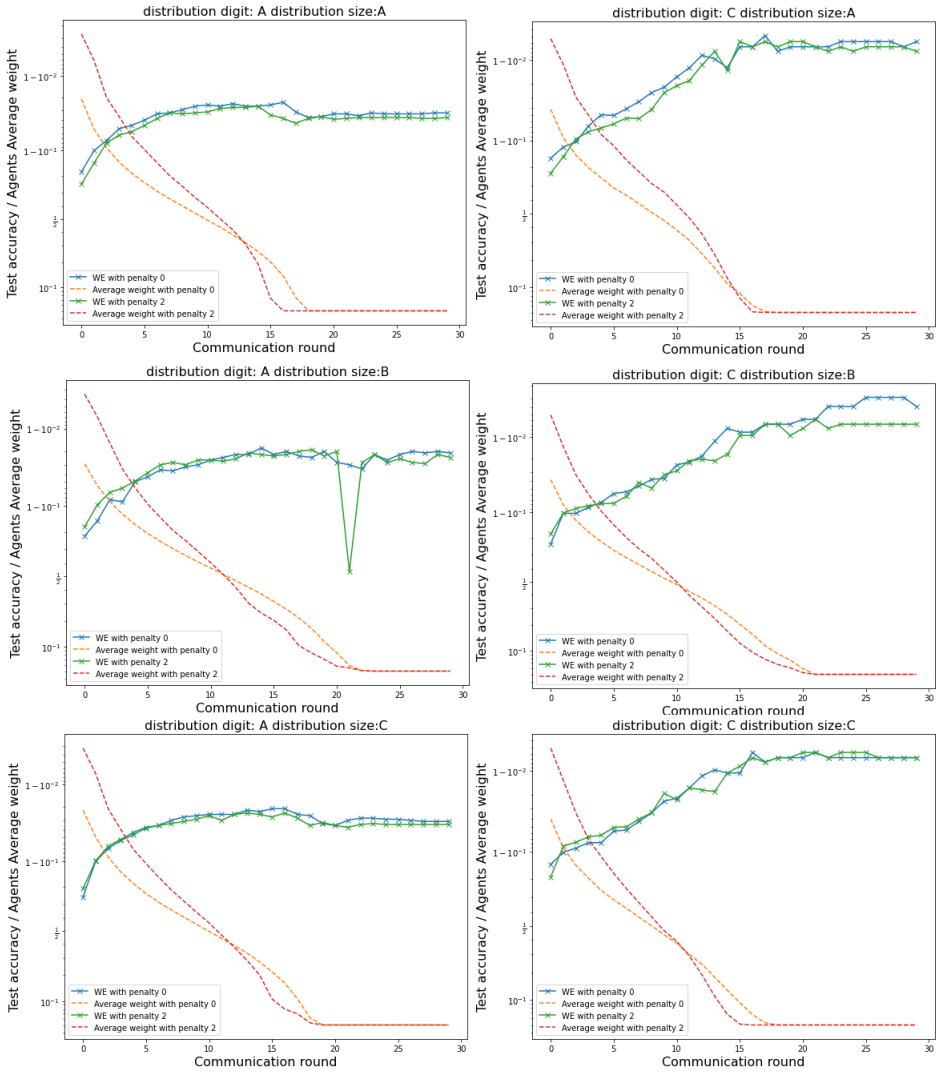

Figure 5: Comparison of the accuracy using MNIST between `Weight Erosion` with or without a distance penalty factor following different distribution of labels and size of samples. The definition of all distributions used can be seen in Table 4.

Table 4: Label and size sample distribution for distance penalty experiment, example Label distribution A with Size distribution C mean that agent 4 will have two times more data than agents 0,1,2,3,7,8,9 and data is IID.

	Label									
Label distribution	0	1	2	3	4	5	6	7	8	9
A	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
C	0.25	0.25	0.25	0.25	0	0	0	0	0	0
Size distribution										
A	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
B	0.21	0.06	0.11	0.06	0.11	0.06	0.16	0.06	0.11	0.06
C	0.1	0.1	0.1	0.1	0.2	0.05	0.05	0.1	0.1	0.1

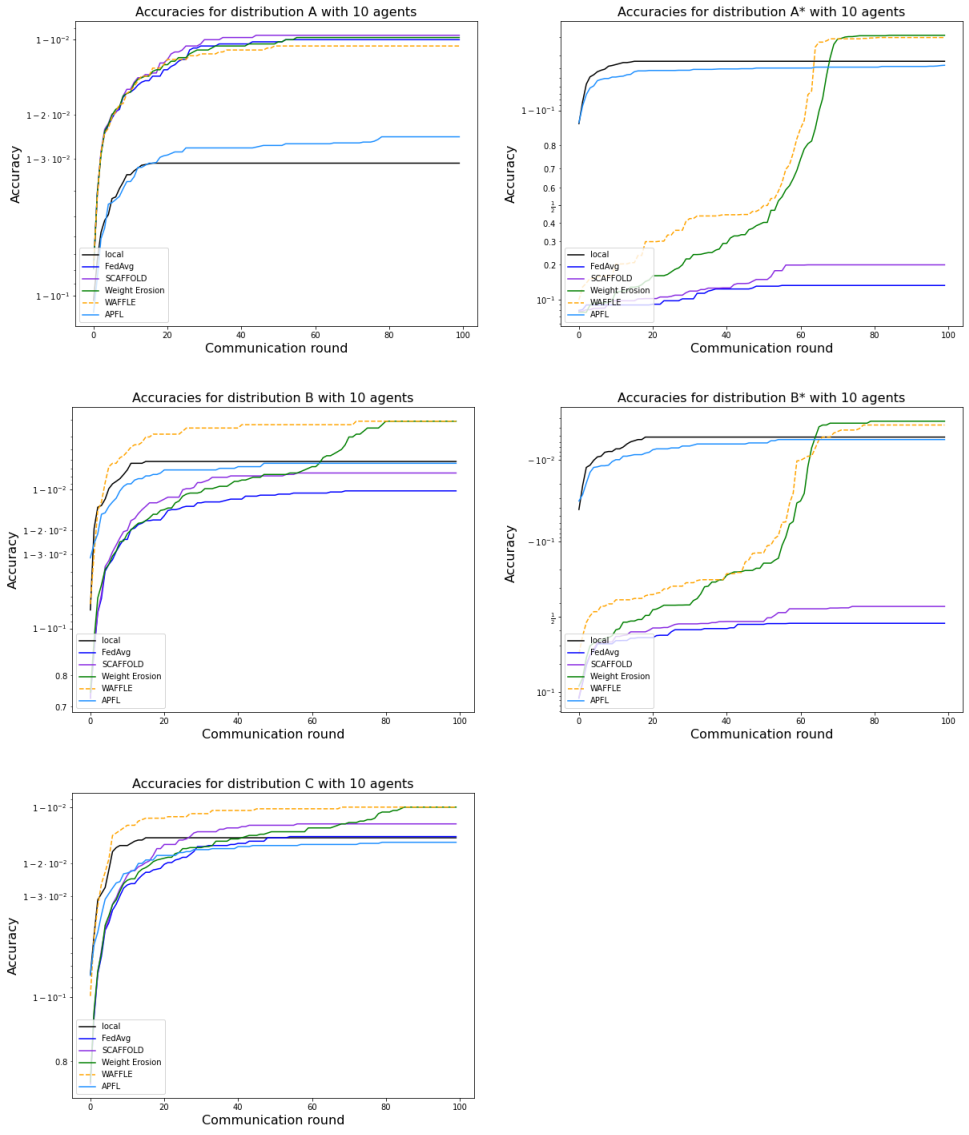

Figure 6: Evolution of the accuracy on MNIST, average over five seeds of the best accuracy obtained up to each turn.

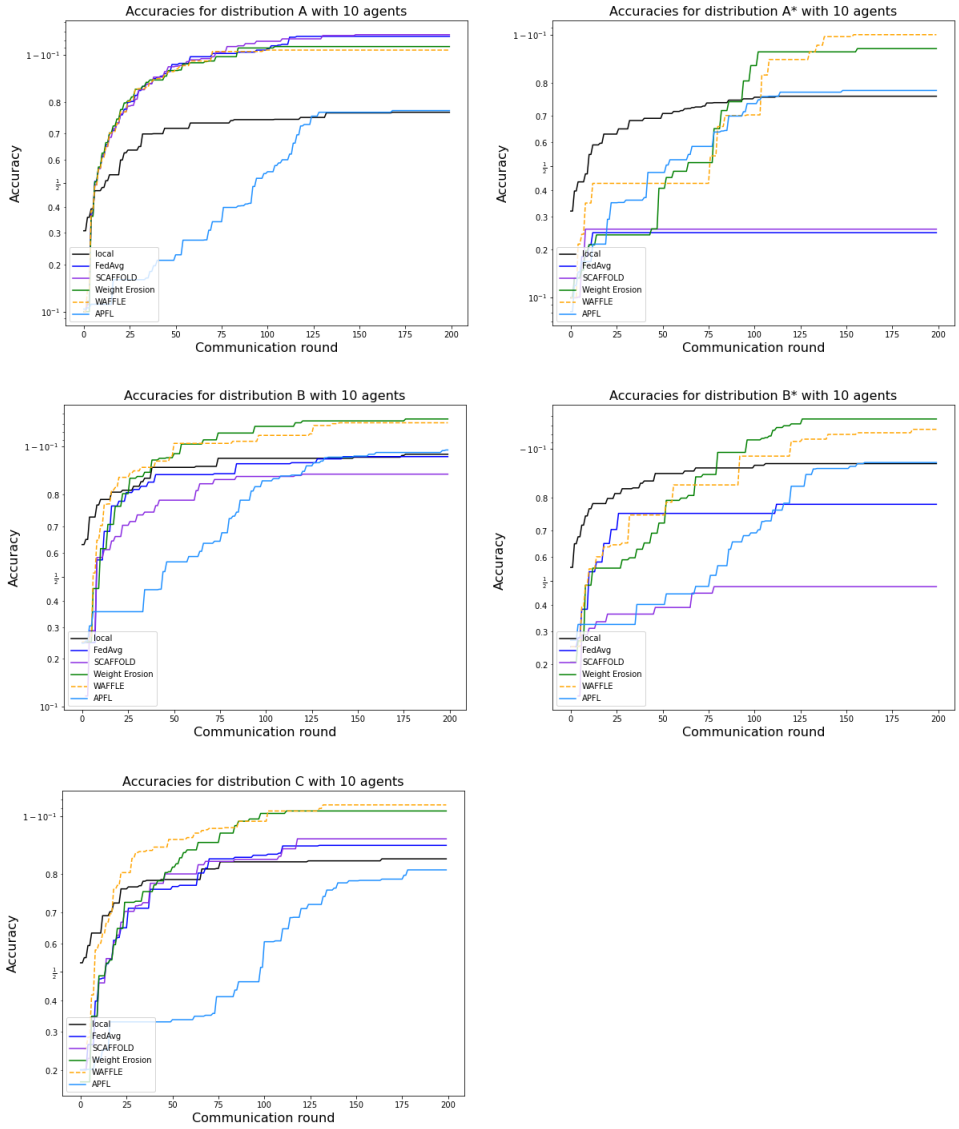

Figure 7: Evolution of the accuracy on CIFAR10, best accuracy obtained up to each turn.

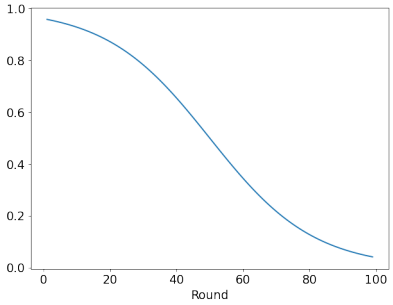

Figure 8: Evolution value of omega with $\Delta\Omega = 3.2$