FedPALS: A method for mitigating label shift in federated learning

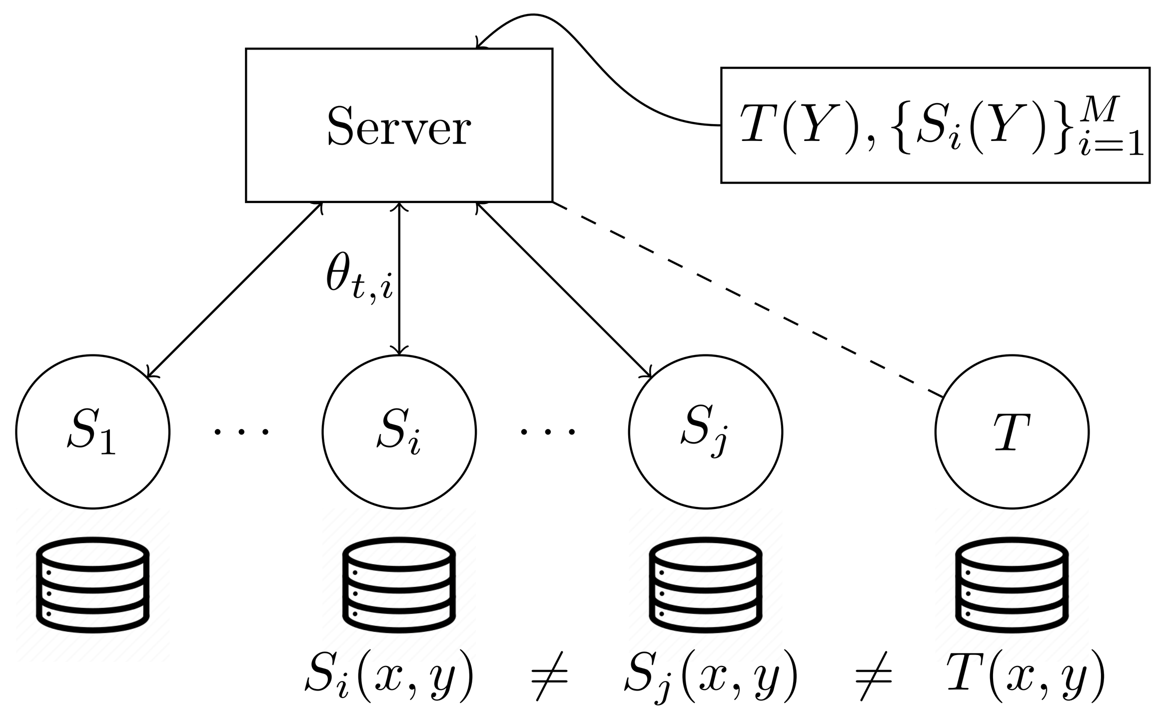
Adam Breitholtz, Chalmers university of Technology
Joint work with Edvin Listo Zec, RISE
Computer Science and Engineering, DSAI
Supervisors: Ass.Prof. Fredrik D. Johansson (CTH), Prof. Devdatt Dubhashi (CTH)

CHALMERS
UNIVERSITY OF TECHNOLOGY

Motivation & Research Goals

Federated learning enables multiple actors to collaboratively train models without sharing private data. Existing algorithms like Federated Averaging are well-justified when clients and the target domain share the same distribution of features and labels.
One common violation of this is label shift, where the label distributions differ across clients or between clients and the target domain, which can significantly degrade model performance. To address this problem, we propose FedPALS, a novel model aggregation scheme that adapts to label shifts by leveraging knowledge of the target label distribution at the central server. Our approach ensures unbiased updates under stochastic gradient descent and outperforms baselines on several tasks.

Methods



Label shift

$\forall i,j \in [M]$
$S_i(X \mid Y) = S_j(X \mid Y) = T(X \mid Y)$

- We consider a setting which assumes server has access to label distributions from clients and the target.
- Possible to mirror target label distribution using clients
- Leads to unbiased estimate of updates in SDG, but could lead to high variance estimates

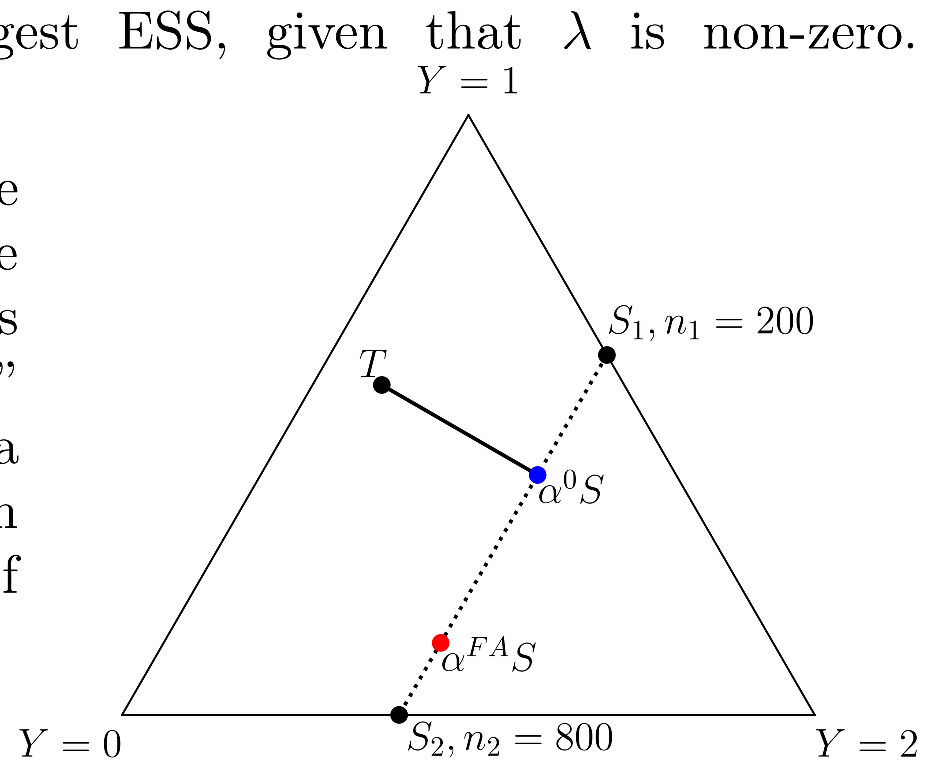
Instead we leverage the notion of Effective sample size (ESS) and combine this with the above to obtain the following aggregation:

$$\textbf{FedPALS}: \quad \alpha^\lambda = \underset{\alpha \in \Delta^{M-1}}{\arg\min} \|T(Y) - \sum_{i=1}^{M} \alpha_i S_i(Y)\|_2^2 + \lambda \sum_i \frac{\alpha_i^2}{n_i}.$$

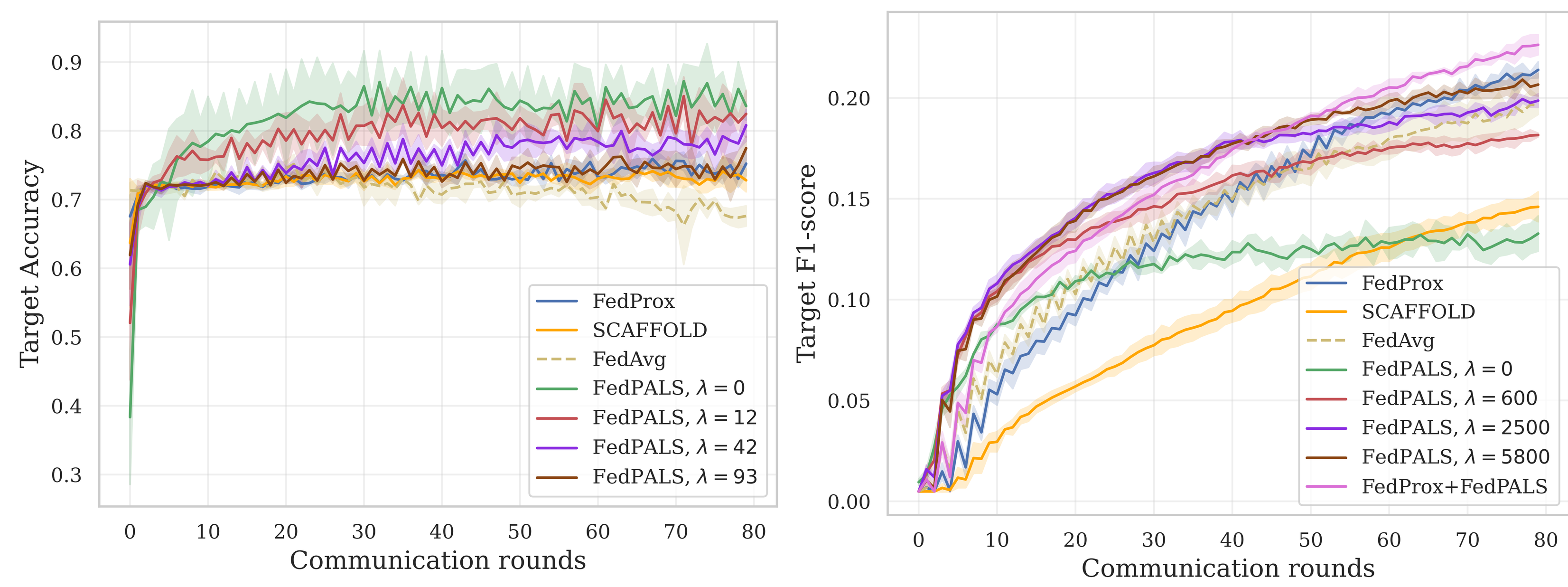
Case 1: $\lambda \to \infty \Rightarrow$ Federated averaging In the limit $\lambda \to \infty$, as the regularization parameter λ grows large, FedPALS aggregation approaches FedAvg aggregation. As a rare special case, whenever $T(Y) = \bar{S} = \sum_{i=1}^M \frac{n_i}{N} S_i(Y)$, FedAvg weights $\alpha^{FA} = \alpha^\lambda$ for any value of λ, since both terms attain their mimima at this point.

Case 2: Covered target, $T \in \mathrm{Conv}(S)$ When the target label distribution is in the convex hull of the source label distributions we can find a convex combination α^c of source distributions $S_i(Y)$ that recreate $T(Y)$, that is, $T(Y) = \sum_{i=1}^M \alpha_i^c S_i(Y)$. However, when there are more clients than labels, $M > K$, such a satisfying combination α^c need not be unique. FedPALS returns the one with the largest ESS, given that λ is non-zero.

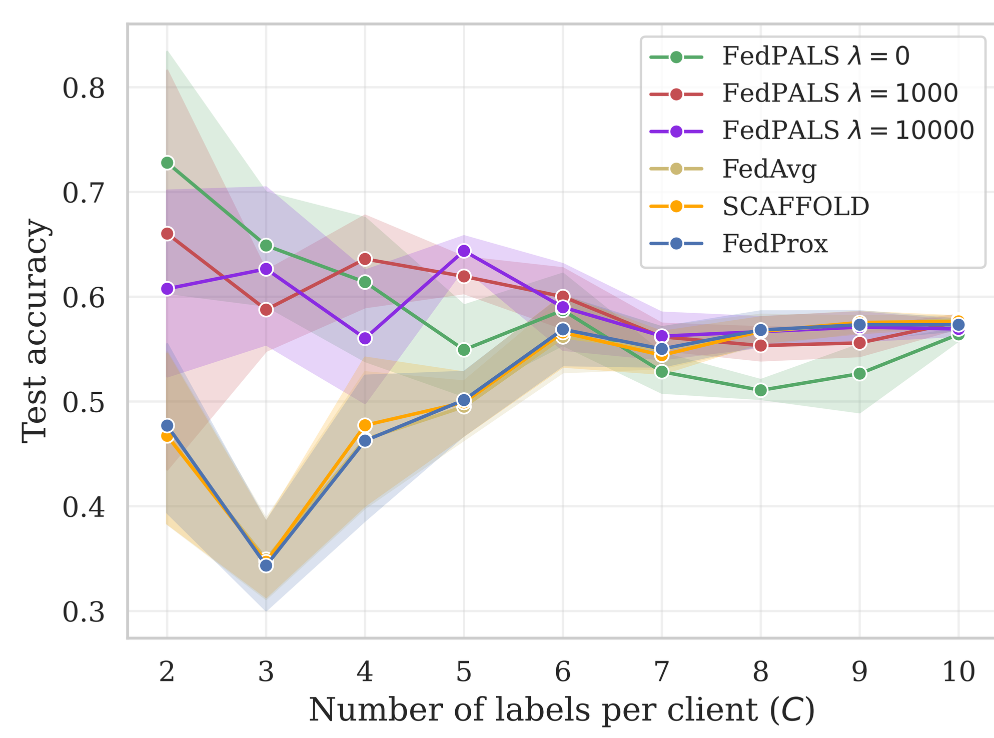
Case 3: $T \notin \mathrm{Conv}(S)$ If the target distribution does not lie in $\mathrm{Conv}(S)$, FedPALS projects the target to the "closest point" in $\mathrm{Conv}(S)$ if $\lambda = 0$, and to a tradeoff between this projection and the FedAvg aggregation if $\lambda > 0$.



Selected Results



In experiments on PACS (left) we observe that our method outperforms baselines significantly. For the larger iWildCam experiment (right) we see that we initially get good performance, but the performance plateaus at different points depending on λ.
A combination of our method with the FedProx baseline, which are not mutually exclusive, results in the strongest model. We do not find a clear policy of how to choose λ from our experiments, the optimal choice seems problem-dependent.



Introducing sparsity in the clients and target we see that our method performs similarly to baselines when only a few labels are missing. With more extreme sparsity we find that our model outperforms baselines substantially.

Conclusion

- We find that our aggregation method outperforms or matches baselines in all tasks considered, especially when there is more substantial sparsity.
- FedPALS provides a possibility to trade-off the mirroring of the target with a higher ESS
- A limitation of our method is the necessity of choosing an appropriate value for λ, the trade-off parameter.
- Mitigating this by introducing adaptive tuning strategies for the parameter is a promising direction for future work.

WASP
WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM