Is logit adjustment a free lunch for heterogeneous federated learning?

Li Ju Prashant Singh Andreas Hellander

Department of Information Technology, Uppsala University

Motivations

- Federated learning (FL) is a collaborative machine learning technique without the need of sharing private data of participants [1].
- When data distribution across participants are different, performance of federated models deteriorate.
- By applying logit adjustment (LA) [2] to local training, it is reported to improve the performance [3].

The standard federated formulation is asymptomatically equivalent to the centralised counterpart. However, the behaviour of locally logit adjustment federated learning is understudied:

- Under what conditions does it recover the Bayes classifier $\hat{y} = \arg_{y\in[C]} \max \Pr(Y = y \mid x)$?
- Why does it alleviate the heterogeneity problem, and at what cost?
- When should (not) we use logit adjusted FL?

But what is logit adjustment?

Let $X \in \mathbb{R}^D$ and $Y \in \{0, 1, \cdots, C\}$ denote data and labels. We are interested in the classifier $\hat{y} = \arg_{y\in[C]} \max \Pr(Y = y \mid x)$. To approximate $P(Y \mid x)$, we generally use model $\mathcal{P} \coloneqq \{\mathrm{softmax}(f(x;\theta)) \mid \theta \in \Theta\}$ where $f(x;\theta) : \mathbb{R}^D \to \mathbb{R}^C$. With a loss function $\ell(\cdot,\cdot)$, the optimal approximation is given by $\mathrm{softmax}(f(x;\theta^\star))$ with

$$\theta^\star = \arg\min_{\theta\in\Theta} \mathbb{E}_{p(x)} \left[\ell(p(y \mid x), \mathrm{softmax}(f(x;\theta)))\right].$$

- Standard formulation: $f(x;\theta) = \mathrm{NN}(x;\theta)$, where $\mathrm{NN}(\cdot;\theta)$ denotes a neural network parameterised by θ.
- Logit adjustmented: $f(x;\theta) = \mathrm{NN}(x;\theta) + \pi$, where $\pi \in \mathbb{R}^C$ is a constant characterising a predefined class prior.

By applying logit adjustment in local training with local class priors $\pi_k,\ \forall k \in [K]$, we obtain logit adjusted federated learning (LA-FL).

Analysis of LA-FL

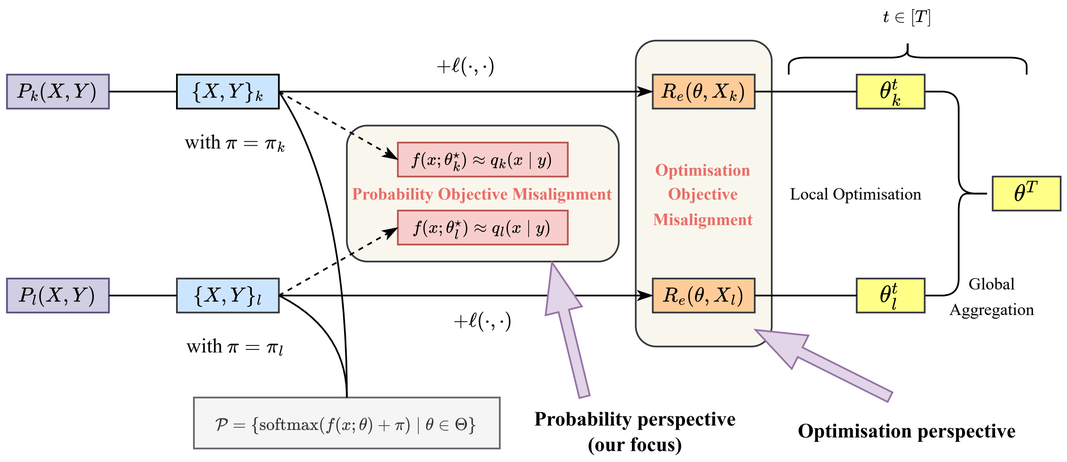

Figure 1. Objective misalignment from two perspectives.

Most previous theoretical works for heterogeneous FL are done from the optimisation perspective (i.e. focusing on local empirical risk functions), but they are not applicable for LA-FL:

- Local objective functions are not empirical risk functions anymore.
- Local objective functions are different by design, due to the introduction of different local class priors $\pi_k,\ \forall k \in [K]$.

We propose an *unified analysis framework* from the probability perspective for both the standard and logit adjusted federated formulation.

Main theoretical takeaways

- The logit adjusted federated models does *not* recover the Bayes classifier, *unless* the mixture distribution of local class priors is explicitly applied to the federated model (*!! privacy compromise*).
- By choosing appropriate class priors, the probability misalignment of local objectives are reduced, resulting in faster convergence compared to the standard formulation.
- However, the faster convergence is at the cost of less informative asymptotic federated models.

Note: All results are asymptotic behaviours of both formulation and may reflect the predictive performances of federated models with limited communication and computational budgets.

To adjust, or not to adjust?

Although in practice with limited communication and computation budgets, the theoretical results may not be reflected on the predictive performance of models directly, we can derive principles for in what conditions federated learning may benefit from logit adjustment:

- The less computation and communication budgets are given, the more benefits logit adjustment brings.
- The more complex the problem (model) is, the more benefits federated learning obtain from LA.
- Partial participation benefits from LA due to the alignment of objectives brought by LA.
- Momentum in local training is more compatible with LA-FL.

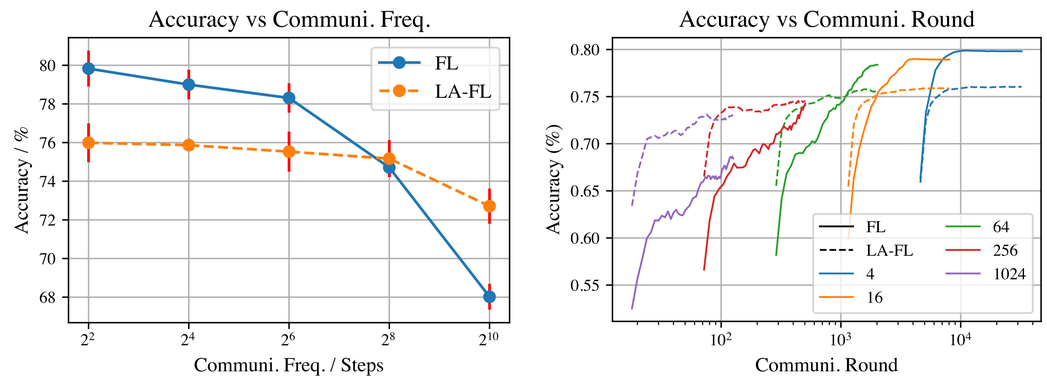

Figure 2. Given a fixed overall computation budget with different communication frequencies, the performances of standard and logit adjusted federated learning are compared with FedAvg. **Left**: With more aligned objectives, LA-FL suffers less from the reduction of communication compared with FL. **Right**: LA-FL converges much faster than FL but the speedup is at the cost of sub-optimality.

Acknowledge & Contact

I acknowledge The Centre for Interdisciplinary Mathematics (CIM) for funding my PhD study and National Academic Infrastructure for Supercomputing in Sweden (NAISS) for providing computing resources. We also thank support from eSSENCE project, a strategic collaborative research programme in e-science.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[3] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022.

li.ju@it.uu.se