Tackling Distribution Shifts in Federated Learning with Superquantile Aggregation

Krishna Pillutla*, Yassine Laguel*, Jérôme Malick, Zaid Harchaoui

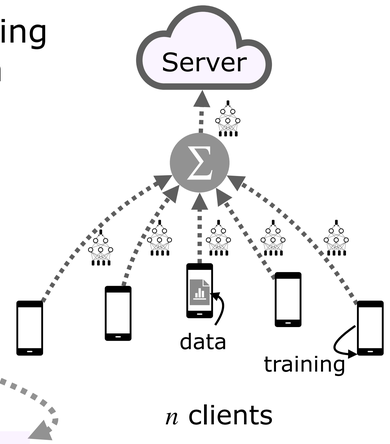
Federated learning

FL = Collaborative learning on decentralized data

Objective on client i:

$$F_i(w) = \mathbb{E}_{z\sim p_i}\left[f(w;z)\right]$$

p_i : data distribution on client i



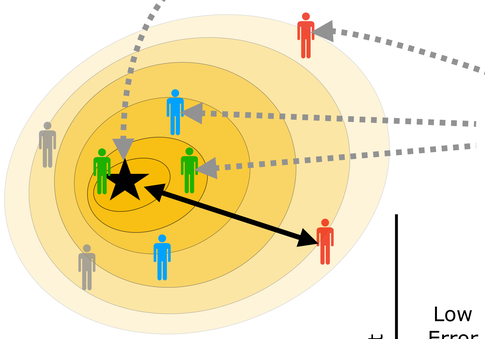
Characteristics of FL

- Data heterogeneity
- Communication cost
- Privacy of user data

Distribution shifts in FL

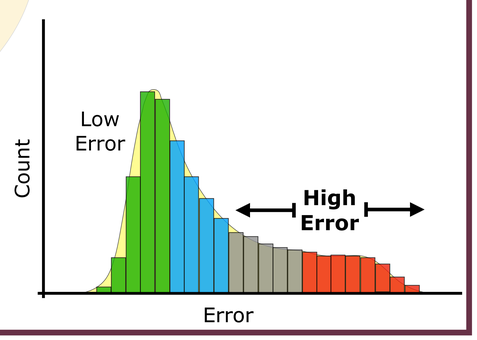
Usual objective (ERM): $\min_w \frac{1}{n}\sum_{i=1}^n F_i(w)$

Model trained to fit the average distribution $\frac{1}{n}\sum_{i=1}^n p_i$



Model is deployed on individual clients

Train-test mismatch $\Longrightarrow$



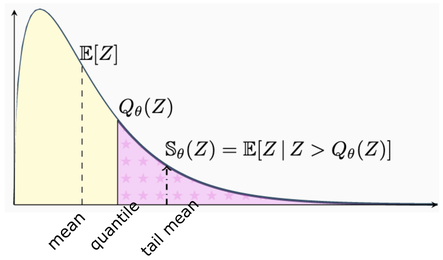
Our Approach: Simplicial-FL

Approach: Minimize the tail error directly

$$\min_w \left[F_\theta(w) := \mathbb{S}_\theta\Big(\left(F_1(w),\cdots,F_n(w)\right)\Big)\right]$$

Superquantile | Conditional Value at Risk

[Rockafellar & Uryasev (2002)]



θ = tail fraction

$Q_\theta(Z) = (1-\theta)$-quantile of Z

$\mathbb{S}_\theta(Z) = (1-\theta)$-superquantile

Distributional robustness: for a new client with distribution $p_\pi = \sum_{i=1}^n \pi_i p_i$, our objective is equivalent to

$$F_\theta(w) = \max_{\pi:\, \pi_i \le (\theta n)^{-1}} \mathbb{E}_{z\sim p_\pi}\left[f(w;z)\right]$$

Optimization Algorithm

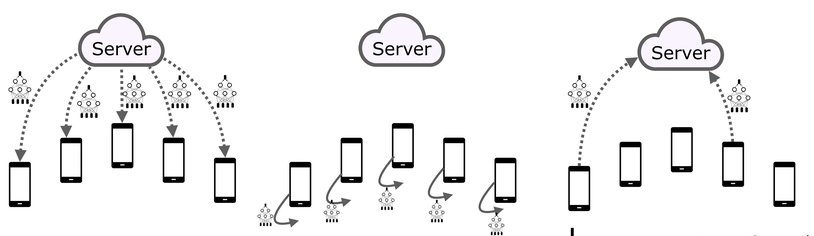
Subgradient expression: if θn is an integer then

$$\partial F_\theta(w) \ni \sum_{i=1}^n \pi_i^\star \nabla F_i(w) \quad \text{where} \quad \pi_i^\star \propto \mathbb{I}(F_i(w) > q)$$

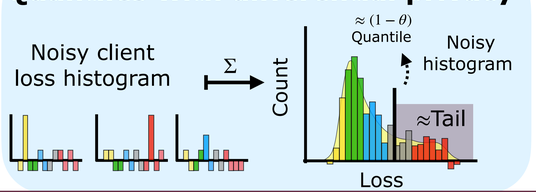
$$q = Q_\theta\big(F_1(w),\ldots,F_n(w)\big)$$

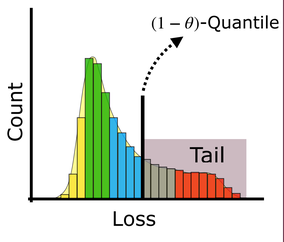
Algorithm:

1. Client sampling + model broadcast
2. Local updates
3. Aggregate updates from tail clients only



Quantiles with differential privacy





Theory

Challenge: unbiased gradient estimator not possible

Optimize mini-batch surrogate which is $(\theta m)^{-1/2}$-close:

$$\tilde{F}_{\theta,m}(w) = \mathbb{E}_{i_1,\ldots,i_m}\left[\mathbb{S}_\theta\big(F_{i_1}(w),\ldots,F_{i_m}(w)\big)\right]$$

Theorem

If client losses are L-smooth & G-Lipschitz, we have the following rates on $\tilde{F}_{\theta,m}$

Nonconvex case: $\sqrt{\frac{LG^2}{t}}$

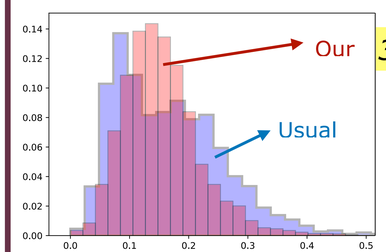
λ-strongly convex case: $\exp(-t/\kappa^{3/2}) + G^2\kappa/t$

where $\kappa = L/\lambda$ is the local condition number

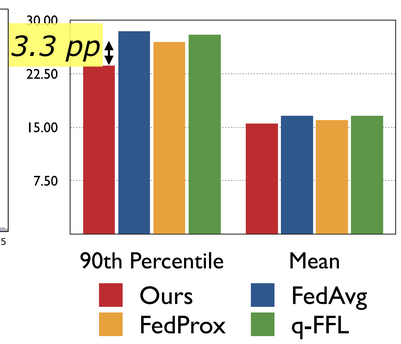
Experiments

Simplicial-FL leads to improvement in the tail performance

Histogram of per-client errors



Misclassif. Error



Pillutla*, Laguel*, Malick, Harchaoui. *Federated Learning with Superquantile Aggregation for Heterogenous Data*. Mach. Learn. (To appear, 2022)

Code

SCAN ME

krishnap25.github.io

KrishnaPillutla