Federated PAC-Bayesian Learning on Non-IID Data

Zihao Zhao[1], Yang Liu[2], Wenbo Ding[1], Xiao-Ping Zhang[1]
[1]Tsinghua-Berkeley Shenzhen Institute, [2]Institute for AI Industry Research

TBSI

1. Central Research Question

- Data across different clients is non-identically and independently distributed (non-i.i.d.) in FL.
- How to deal with clients that have different prior distributions and posterior distributions?
- What is the generalization performance when clients have non-i.i.d. data and different prior distributions?

2. Problem Setup

- Total K clients, each equipped with its own dataset $S_k = (x_i, y_i)_{i=1}^n \subseteq (\mathcal{X}, \mathcal{Y})^n$.
- Let $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ be a given loss function and $h_k \in \mathcal{H}$ is a stochastic estimator on client k where $\mathcal{H}$ is the hypothesis class.
- In the PAC-Bayesian framework, each client holds a tailored prior distribution P_k. The goal is to optimize the posterior distribution $Q_k \in \mathcal{H}$.
- Define the *population risk*:

$$L(Q_1, \ldots, Q_K) \triangleq \frac{1}{K} \sum_{k=1}^{K} \mathop{\mathbb{E}}_{h_k \sim Q_k} \mathop{\mathbb{E}}_{(x_k, y_k) \sim D_k} [\ell(h_k(x_k), y_k)],$$

and the *empirical risk*:

$$\hat{L}(Q_1, \ldots, Q_K) \triangleq \frac{1}{nK} \sum_{k=1}^{K} \mathop{\mathbb{E}}_{h_k \sim Q_k} \sum_{i=1}^{n} \ell(h_k(x_{k,i}), y_{k,i}).$$

- Federated learning procedure: each client maintains its prior P_k locally, while sharing the posterior; the posteriors will be aggregated as $\bar{Q} = \prod_{k=1}^{K} Q_k^{p(k)}$.
  - Intuition of this aggregation: $\min_h L(h) = \min_h \sum_{k=1}^{K} p(k) L_k(h) = \max_h \ln \prod_{k=1}^{K} p(h \mid \mathcal{D}_k)^{p(k)}$.

3. Main Theorem

Theorem 1 (Federated PAC-Bayesian bound) *For any $\delta \in (0,1]$, assume the loss function $\ell(\cdot,\cdot)$ is bounded in $[0, C]$, the following inequality holds uniformly for all posterior distributions Q and for any $\delta \in (0,1)$,*

$$\mathop{\mathbb{P}}_{S_1,\ldots,S_K} \left\{ \forall Q_1, \ldots, Q_K, L(Q_1, \ldots, Q_K) \le \hat{L}(Q_1, \ldots, Q_K) + \frac{\sum_{k=1}^{K} p(k) D_{KL}(Q_k \| P_k) + \log \frac{1}{\delta}}{\lambda} + \frac{\lambda C^2}{8Kn} \right\} > 1 - \delta.$$

4. FedPB: Iteratively Optimizing Upper Bound

Local objective function: $\mathcal{J}(Q_k) = \lambda \mathcal{L}_k + p(k) D_{KL}(Q_k \| P_k)$, where $\mathcal{L}_k = \mathop{\mathbb{E}}_{h_k \sim Q_k} \frac{1}{n} \sum_{i=1}^{n} \ell(h_k(x_{k,i}), y_{k,i})$.

- Phase 1 (Optimize the posterior):

$$\hat{Q}_k^{t+1} = \arg\min_{Q_k} \mathcal{J}(Q_k),$$

yielding the solution

$$\frac{d\hat{Q}_k^{t+1}}{dP_k^t}(h) = \frac{\exp(-\lambda \ell(h, z_i))}{\mathbb{E}_{h \sim P_k^t}[\exp(-\lambda \ell(h, z_i))]}.$$

- Phase 2 (Optimize the prior):

$$\hat{P}_k^{t+1} = Q_k^t.$$

Corollary 1 (The choice of λ) *Suppose $\lambda \in \Xi := \{0, \ldots, \xi\}$ and $|\cdot|$ denotes the cardinality of a set. For any $\delta \in (0,1)$ and a properly chosen λ, with probability at least $1-\delta$,*

$$L(Q_1, \ldots, Q_K) \le \hat{L}(Q_1, \ldots, Q_K) + C \sqrt{\frac{\sum_{k=1}^{K} p(k) D_{KL}(Q_k \| P_k) + \log \frac{|\Xi|}{\delta}}{2Kn}}.$$

- Optimal value of the hyper-parameter λ:

$$\lambda^* = \sqrt{8Kn \left(\sum_{k=1}^{K} p(k) D_{KL}(Q_k \| P_k) + \log \frac{|\Xi|}{\delta} \right) / C}.$$

5. Numerical Experiments

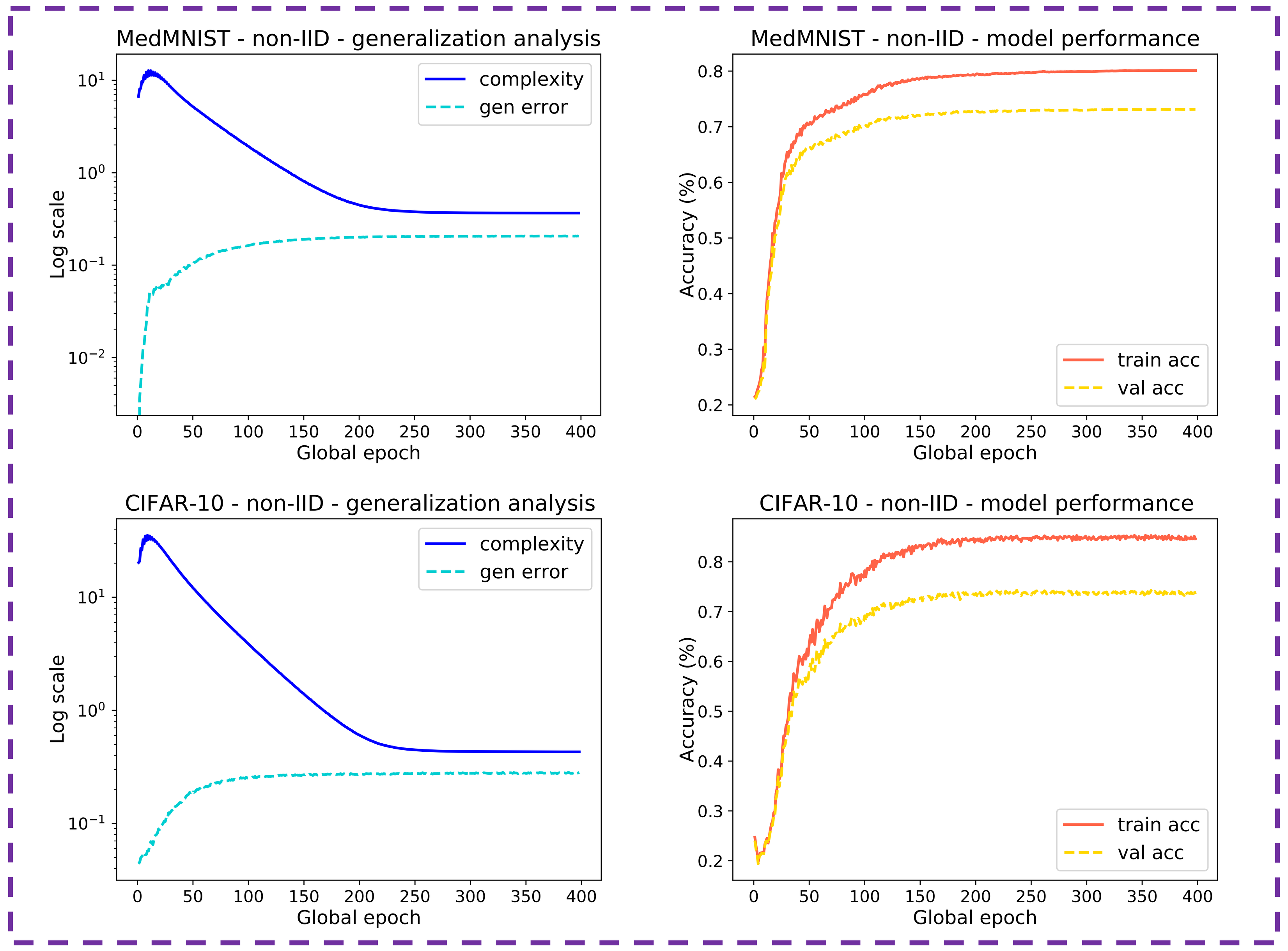
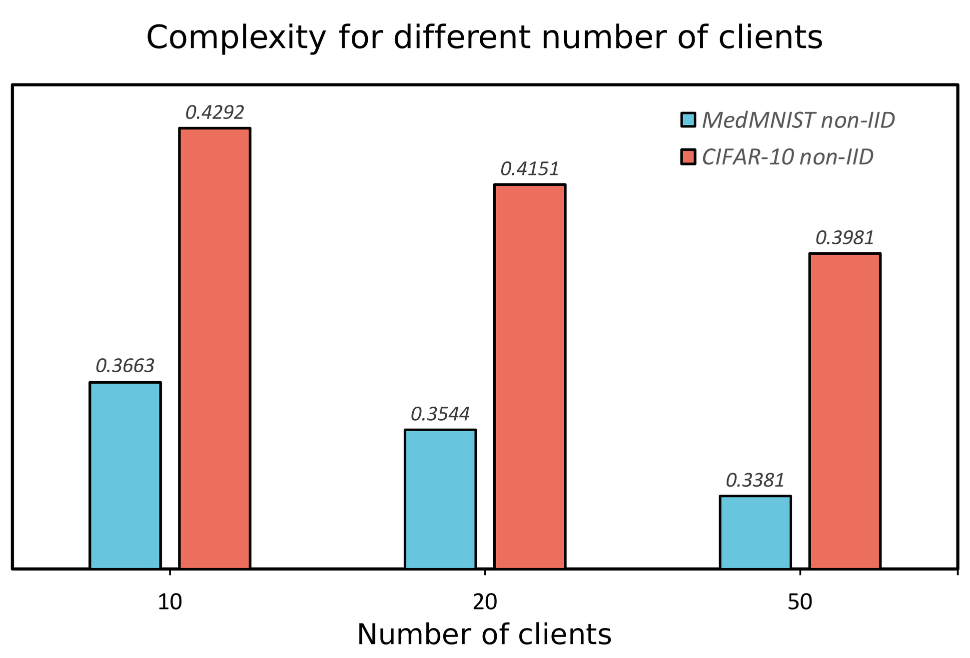


Table 1: Model accuracy (%) for the data-independent prior and data-dependent prior in three data-generating scenarios.

Method	MedMNIST			CIFAR-10		
	Balanced	Unbalanced	Dirichlet	Balanced	Unbalanced	Dirichlet
Data-independent	53.47 ± 1.12	49.44 ± 1.10	55.24 ± 6.92	50.89 ± 0.62	47.19 ± 0.92	57.93 ± 0.55
Data-dependent	$\mathbf{77.10 \pm 4.25}$	$\mathbf{77.34 \pm 3.42}$	$\mathbf{77.48 \pm 4.75}$	$\mathbf{84.41 \pm 0.94}$	$\mathbf{79.39 \pm 0.56}$	$\mathbf{86.11 \pm 0.53}$



Our code