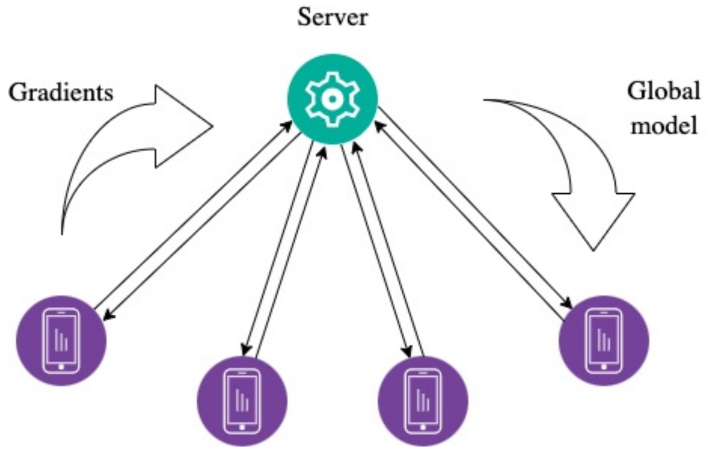
# Fine-Tuning Personalization in Federated Learning to Mitigate Adversarial Clients

Youssef Allouah - **Abdellah El Mrini** - Rachid Guerraoui
Nirupam Gupta - Rafael Pinot

# Heterogeneity and Byzantine Attacks in FL




Heterogeneity: Different data distributions

Byzantine Robustness: Learning despite possible malicious updates

- n participants, among which $f < n/2$ are adversarial
- Each honest participant has a local data distribution
- Personalization Goal: each participants learns a model that works best for their local data

# Personalization through interpolation

**ERM setting**

$$\mathcal{R}_i(\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\ell(h_\theta(x), y)\right]$$

True error

Empirical error

$$\mathcal{L}_i(\theta) \coloneqq \frac{1}{m}\sum_{(x,y)\in S_i} \ell(h_\theta(x), y)$$

**Problem formulation**

For $\lambda \in [0,1]$: $\min_{\theta_i\in\Theta} \mathcal{L}_i^\lambda(\theta_i) \coloneqq (1-\lambda)\mathcal{L}_i(\theta_i) + \lambda\mathcal{L}_\mathcal{C}(\theta_i)$

where $\mathcal{L}_\mathcal{C}(\theta) \coloneqq \frac{1}{|\mathcal{C}|}\sum_{j\in\mathcal{C}} \mathcal{L}_j(\theta)$

$\mathcal{C}$ the set of correct clients

Smooth and strongly convex

**Algorithm 1** Interpolated Personalized Gradient Descent for client $i \in \mathcal{C}$

**Require:** Initialization $\theta_i^0$, aggregation rule $F$, learning rate $\eta$, number of iterations $T$, and collaboration parameter $\lambda$.

1: **for** $t = 1 \ldots T$ **do**
2: Broadcast $\theta_i^{t-1}$ to all clients
3: **for** $j = 1 \ldots n$, $j \neq i$ **do**
4: Receive $g_{i,j}^t = \nabla\mathcal{L}_j(\theta_i^{t-1})$ from client $j$ ▷ adversarial clients send corrupted gradients
5: **end for**
6: Compute local gradient $g_{i,i}^t = \nabla\mathcal{L}_i(\theta_i^{t-1})$
7: Robustly aggregate $R_i^t = F(g_{i,1}^t, \ldots, g_{i,n}^t)$
8: Update and project local parameters

$$\theta_i^t = \Pi_\Theta\left(\theta_i^{t-1} - \eta\left((1-\lambda)g_{i,i}^t + \lambda R_i^t\right)\right)$$

9: **end for**

# Effect on Optimization

At iteration T:

$$\mathcal{L}_i^\lambda(\theta_i^T) - \mathcal{L}_{i,*}^\lambda \le \left(1 - \frac{\mu}{2L}\right)^T \frac{L}{\mu}\mathcal{L}_0 + \mathcal{O}\left(\lambda^2 \kappa G^2\right)$$

Standard error for the strongly convex case

Additional error due to Heterogeneity and Byzantine adversaries

=> The optimization error is controllable: local learning is best

=> To see the full picture, we need to look at generalization

# Optimization/Generalization Tradeoff

## True Error Analysis

$$\mathcal{R}_i(\theta_i^T) - \mathcal{R}_i(\theta_i^*) \leq \left(1 - \frac{\mu}{2L}\right)^T \frac{L}{\mu}\mathcal{L}_0 + \frac{5L\lambda^2\kappa G^2}{\mu^2} + 2\lambda\Phi(\mathcal{D}_i, \mathcal{D}_{\mathcal{C}}) + 4\beta\sqrt{\frac{\left(1 - \lambda + \frac{\lambda}{n-f}\right)^2}{m} + \frac{\lambda^2}{m(n-f)}}$$

with $\beta \approx \sqrt{\mathrm{Pdim}\mathcal{H}}$

Task complexity

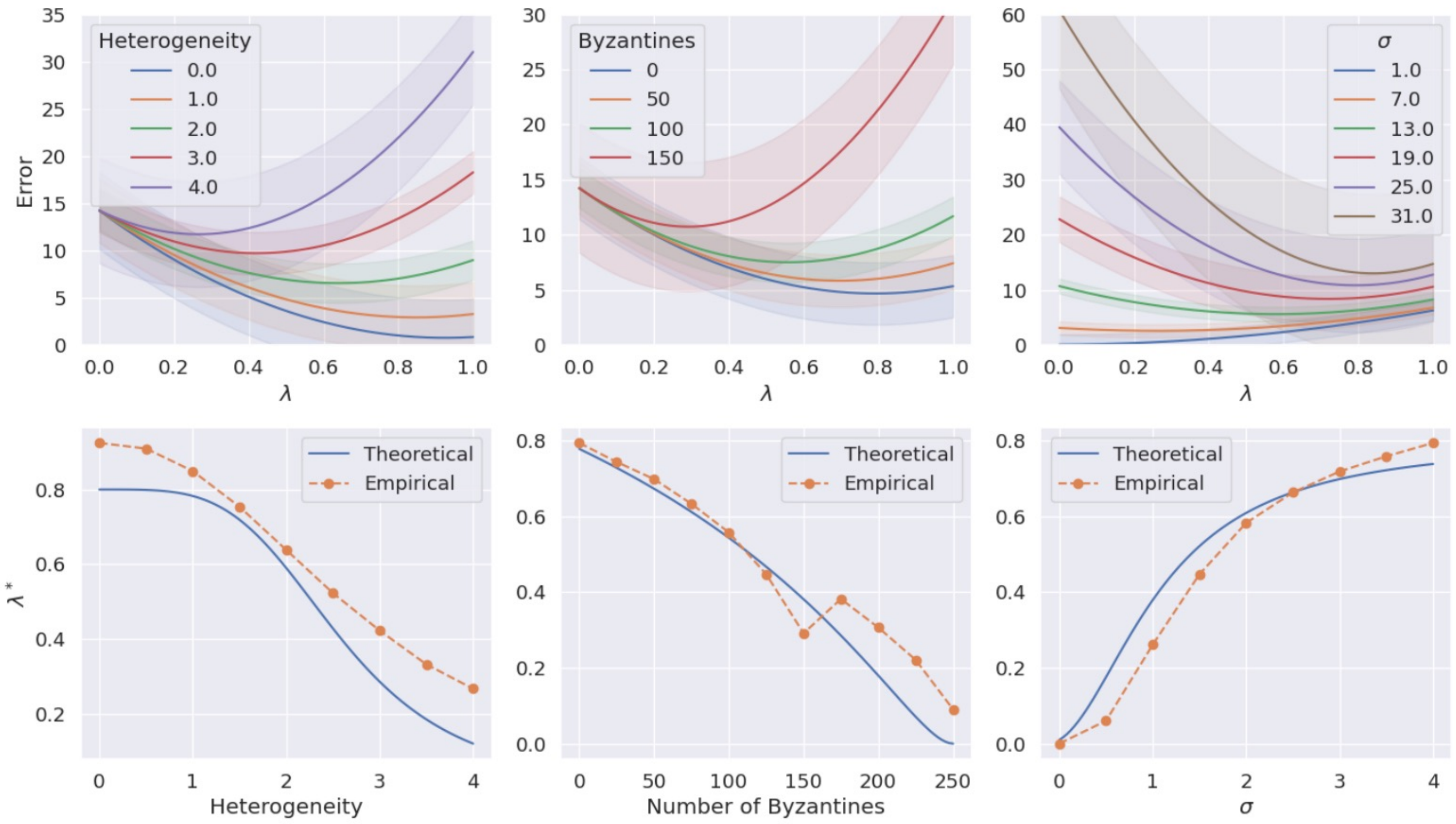
Heterogeneity

## Consequence

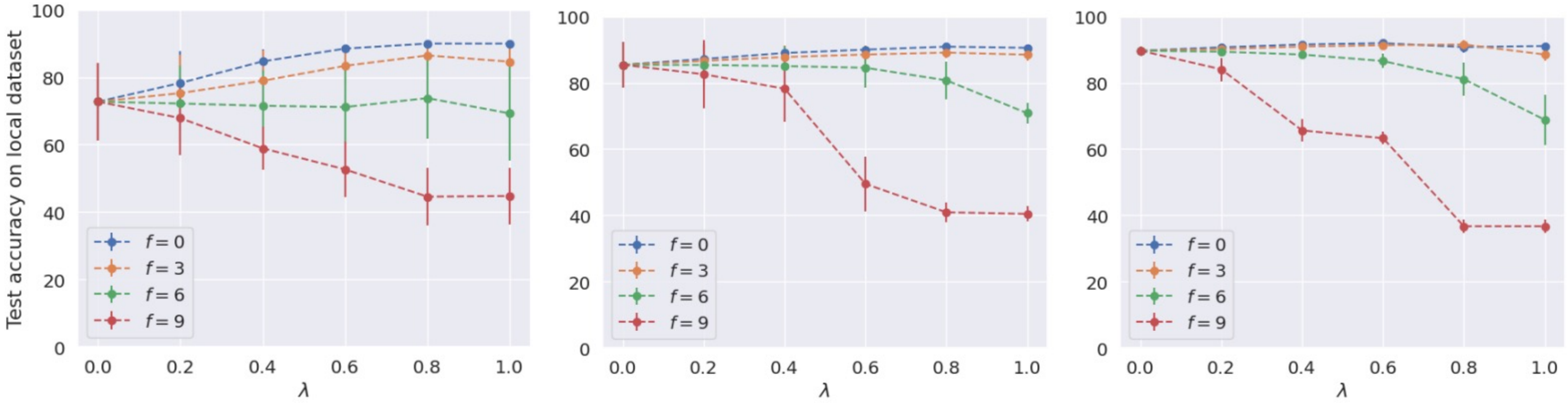
$$\lambda^* \approx \Pi_{[0,1]}\left(\frac{\sqrt{\frac{\mathrm{Pdim}(\mathcal{H})}{m}} - \Phi(\mathcal{D}_i, \mathcal{D}_{\mathcal{C}})}{\frac{f}{n}G^2}\right)$$

Byzantine effect

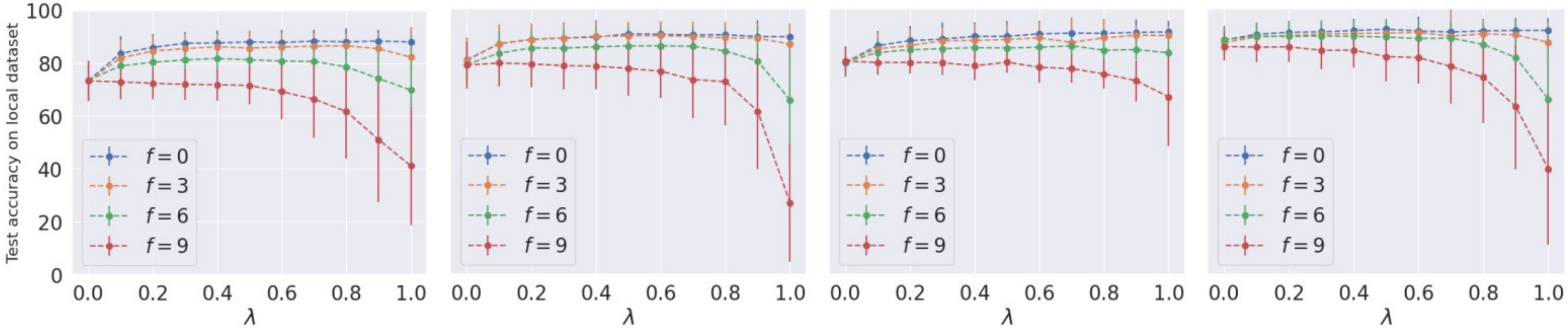
# Experimental Results – Mean Estimation



# Experimental Results – Classification




(a) $m = 16$ (b) $m = 48$ (c) $m = 128$




(d) $\alpha = \infty, m = 32$ (e) $\alpha = 0.5, m = 32$ (f) $\alpha = \infty, m = 64$ (g) $\alpha = 0.5, m = 64$