Single-shot General Hyperparameter Optimization for Federated Learning

—

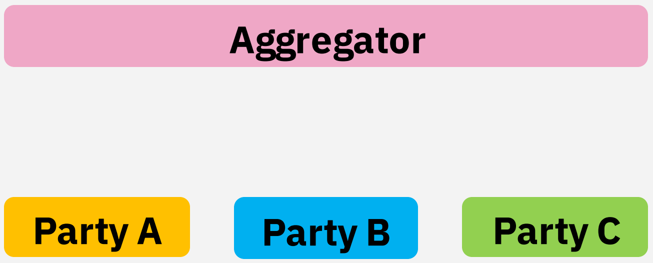
Yi Zhou, Parikshit Ram, Theodoros Salonidis, Nathalie Baracaldo,
Horst Samulowitz, Heiko Ludwig
IBM Research

Federated Model Training

- Multiple clients
- Orchestrating aggregator
- Multiple rounds of communication
- Fixed hyperparameter $\theta \in \Theta$

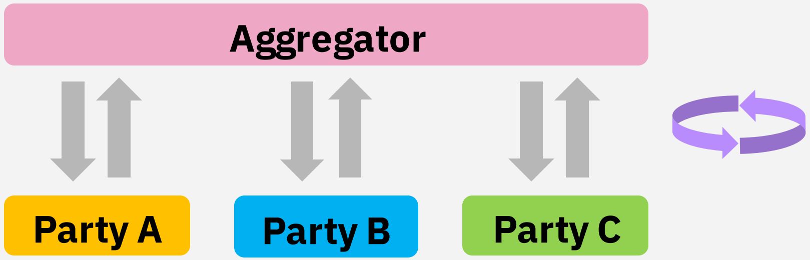
Federated Model Training

- Multiple clients
- Orchestrating aggregator
- Multiple rounds of communication
- Fixed hyperparameter $\theta \in \Theta$



Federated Model Training

- Multiple clients
- Orchestrating aggregator
- Multiple rounds of communication
- Fixed hyperparameter $\theta \in \Theta$



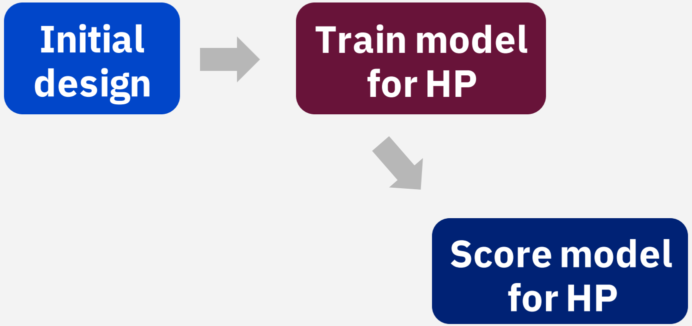
Hyperparameter Optimization

- Generate initial design
- Train & score models for each HP
- Iteratively (until budget consumed)
  - Create loss surface from model scores
  - Select next HP minimizing loss surface
  - Train & score model for HP

Initial design

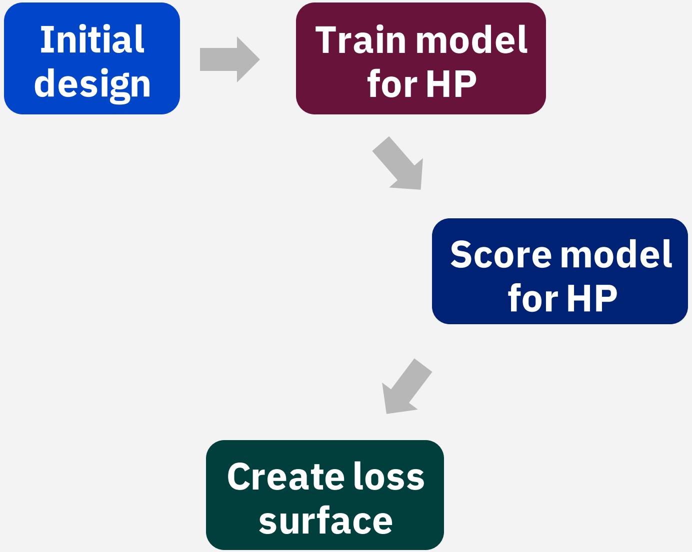
Hyperparameter Optimization

- Generate initial design
- Train & score models for each HP
- Iteratively (until budget consumed)
  - Create loss surface from model scores
  - Select next HP minimizing loss surface
  - Train & score model for HP



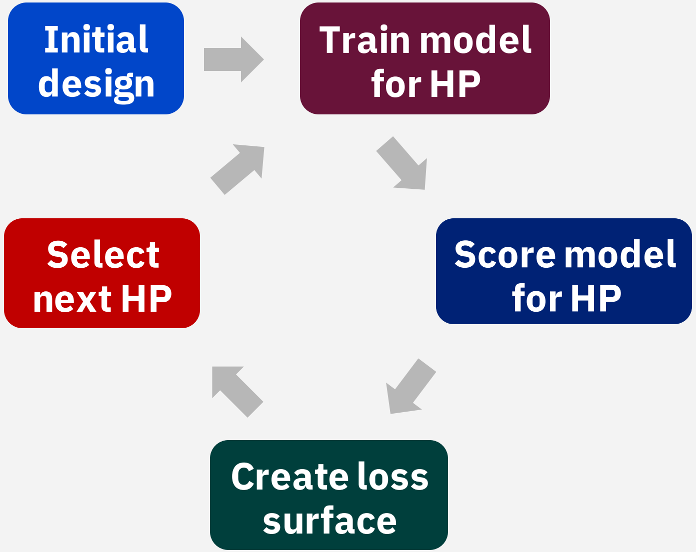
Hyperparameter Optimization

- Generate initial design
- Train & score models for each HP
- Iteratively (until budget consumed)
  - Create loss surface from model scores
  - Select next HP minimizing loss surface
  - Train & score model for HP



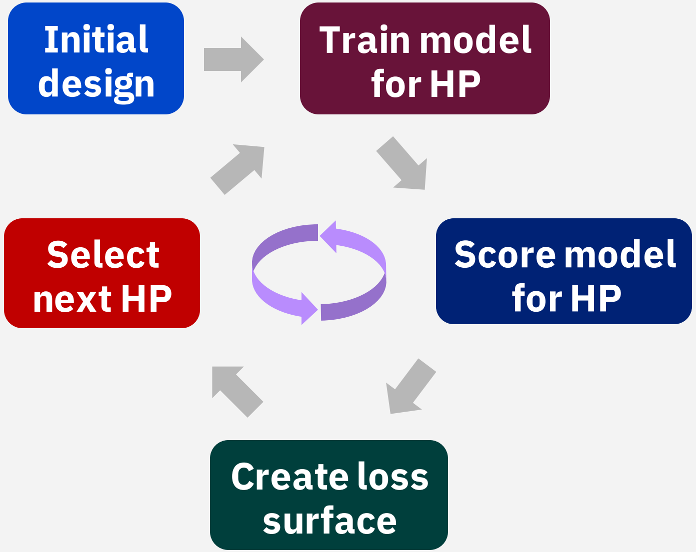
Hyperparameter Optimization

- Generate initial design
- Train & score models for each HP
- Iteratively (until budget consumed)
  - Create loss surface from model scores
  - Select next HP minimizing loss surface
  - Train & score model for HP

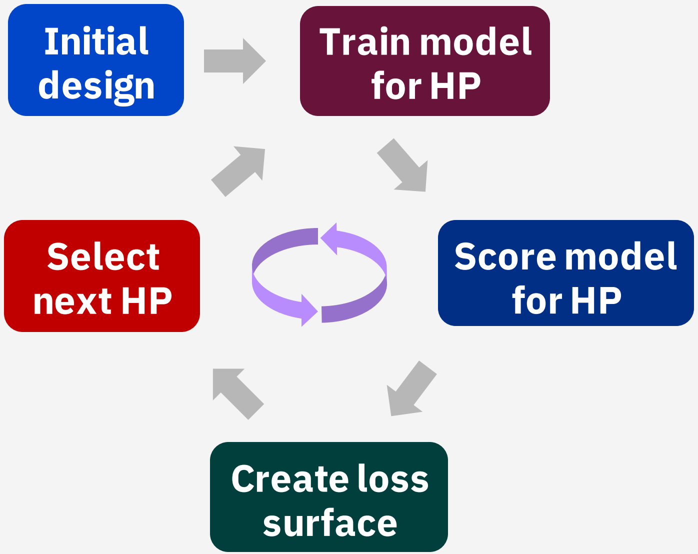


Hyperparameter Optimization

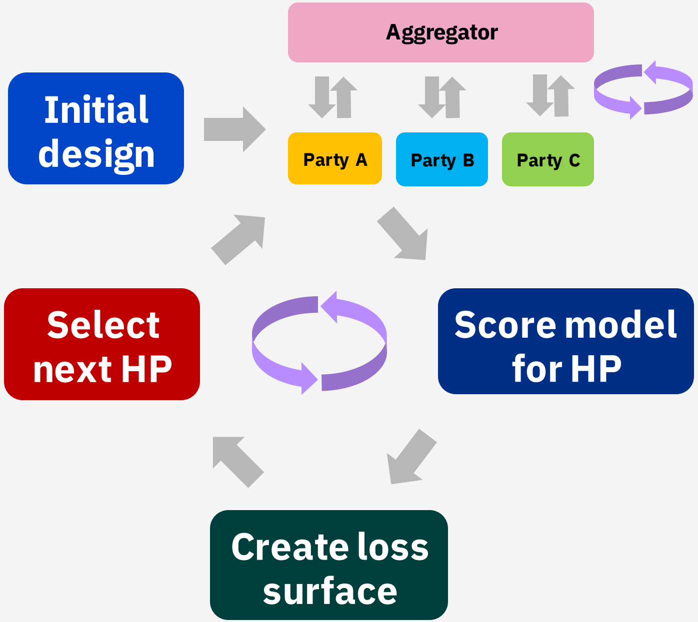
- Generate initial design
- Train & score models for each HP
- Iteratively (until budget consumed)
  - Create loss surface from model scores
  - Select next HP minimizing loss surface
  - Train & score model for HP



Hyperparameter Optimization



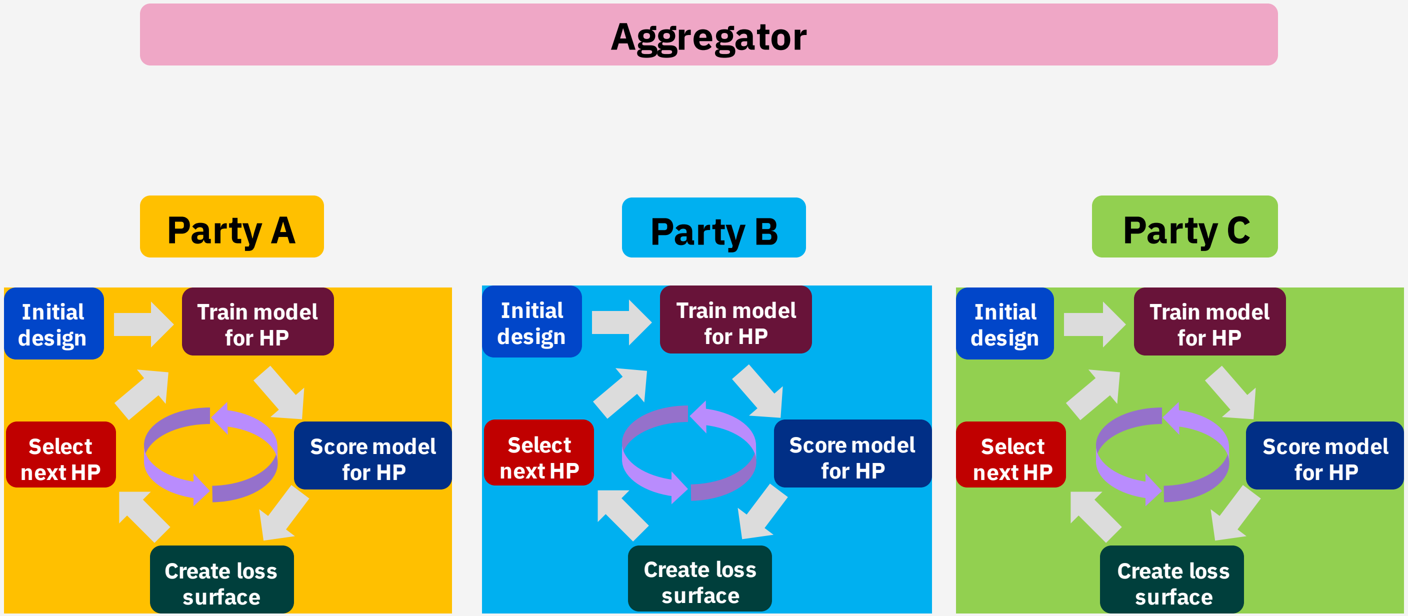
Multi-shot Federated Hyperparameter Optimization



- Significant communication overhead
- Computationally infeasible

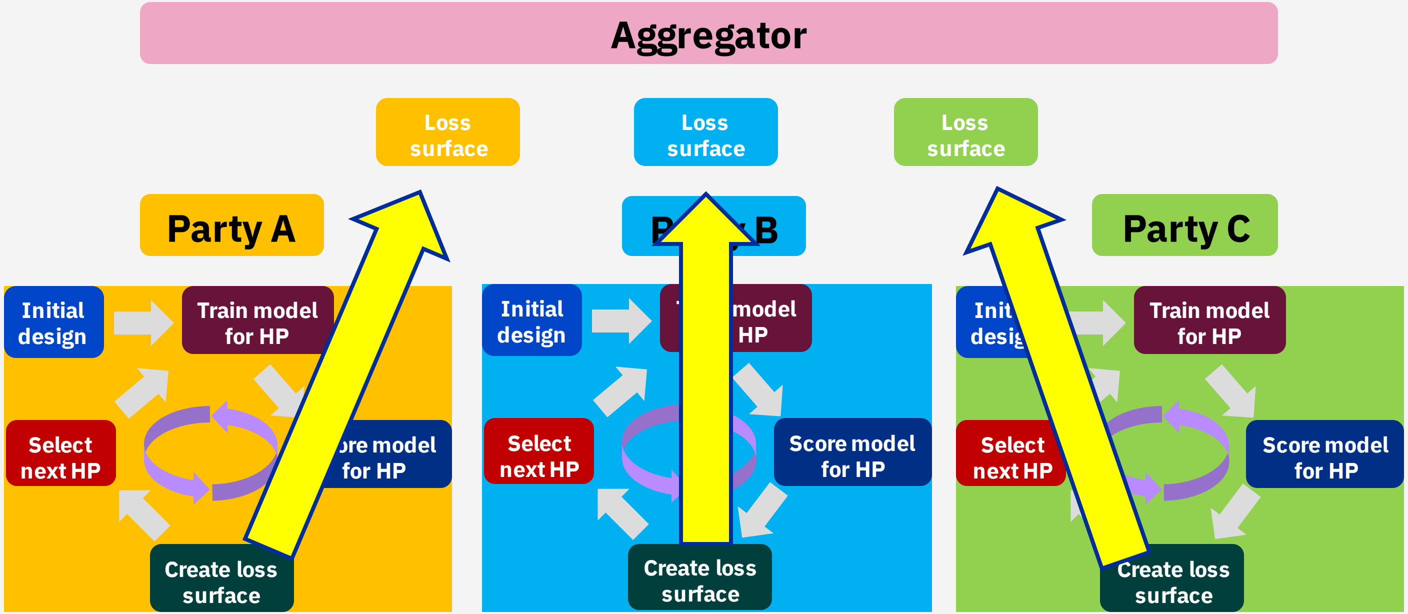
FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation



Per-client independent local HPO

FLoRA: Federated Loss Surface Aggregation



Collect loss surfaces at the aggregator

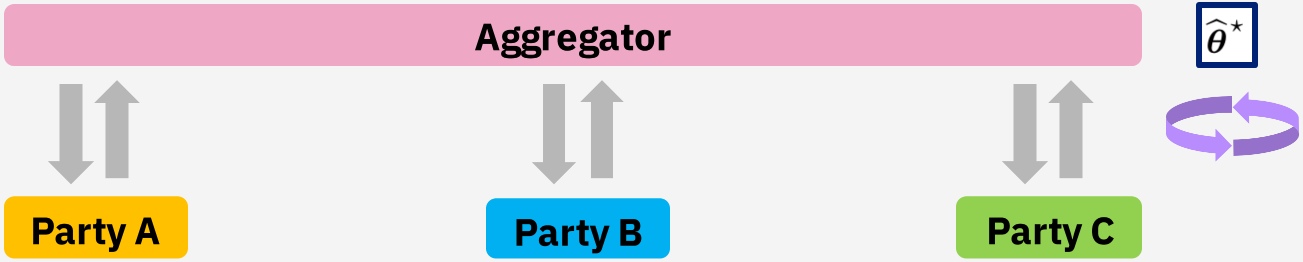
FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Surface	$\hat{\ell}(\theta) :=$
SGM	$f(\theta)$
SGM+U	$f(\theta) + \alpha \cdot u(\theta)$
MPLM	$\max_{i \in [p]} f_i(\theta)$
APLM	$1/p \sum_{i \in [p]} f_i(\theta)$

$$\widehat{\boldsymbol{\theta}}^{\star} \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{\ell}(\boldsymbol{\theta})$$

Aggregate loss surfaces & select most promising HP

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation



Single federated model training with selected HP

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Advantages

- Single-shot: Single federated training needed
- Agnostic to machine learning model type
- No "weight-sharing" requirement
- Low additional communication overhead

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D})$$

FLoRA loss

FL**o**RA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D})$

FLoRA loss

$\tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})$

Optimal loss

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D}) - \tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})$$

FLoRA loss **Optimal loss**

Optimality Gap

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D}) - \tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})$$

FLoRA loss **Optimal loss**

Optimality Gap

$$\leq \max_{\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}} \sum\nolimits_{i \in [p]} C_{\boldsymbol{\alpha}} \left\{ C_{\beta} \sum\nolimits_{j \in [p], j \neq i} w_j \mathcal{W}_1 (\mathcal{D}_j, \mathcal{D}_i) + C_{\tilde{L}, \hat{L}_i} \min\nolimits_{t \in [T]} d(\boldsymbol{\theta}, \boldsymbol{\theta}_t^{(i)}) + \delta_i \right\}$$

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\underbrace{\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D})}_{\text{FLoRA loss}} - \underbrace{\tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})}_{\text{Optimal loss}}$$

Optimality Gap

$$\leq \max_{\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}} \sum_{i \in [p]} C_{\boldsymbol{\alpha}} \left\{ C_{\beta} \boxed{\sum_{j \in [p], j \neq i} w_j \mathcal{W}_1(\mathcal{D}_j, \mathcal{D}_i)} + C_{\tilde{L}, \hat{L}_i} \min_{t \in [T]} d(\boldsymbol{\theta}, \boldsymbol{\theta}_t^{(i)}) + \delta_i \right\}$$

Wasserstein distance between per-party distributions – measures data heterogeneity

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D}) - \tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})$$

FLoRA loss

Optimal loss

Optimality Gap

$$\leq \max_{\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}} \sum_{i \in [p]} C_{\boldsymbol{\alpha}} \Big\{ C_\beta \sum_{j \in [p], j \neq i} w_j \mathcal{W}_1(\mathcal{D}_j, \mathcal{D}_i)$$

$$+ C_{\tilde{L}, \hat{L}_i} \boxed{\min_{t \in [T]} d(\boldsymbol{\theta}, \boldsymbol{\theta}_t^{(i)})} + \delta_i \Big\}$$

Proximity to HPs seen during local per-client HPO

FLoRA: Federated Loss Surface Aggregation

Theoretical Guarantee

$$\underbrace{\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D})}_{\text{FLoRA loss}} - \underbrace{\tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})}_{\text{Optimal loss}}$$

Optimality Gap

$$\leq \max_{\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}} \sum_{i \in [p]} C_{\boldsymbol{\alpha}} \Big\{ C_{\beta} \sum_{j \in [p], j \neq i} w_j \mathcal{W}_1(\mathcal{D}_j, \mathcal{D}_i)$$

$$+ C_{\tilde{L}, \hat{L}_i} \min_{t \in [T]} d(\boldsymbol{\theta}, \boldsymbol{\theta}_t^{(i)}) + \boxed{\delta_i}$$

Loss surface approximation

FLoRA: **F**ederated **Lo**ss Su**r**face **A**ggregation

Theoretical Guarantee

$$\tilde{\ell}(\widehat{\boldsymbol{\theta}}^{\star}, \mathcal{D}) - \tilde{\ell}(\boldsymbol{\theta}^{\star}, \mathcal{D})$$

FLoRA loss

Optimal loss

Optimality Gap

With IID data

$$\leq \max_{\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}} \sum_{i \in [p]} C_{\boldsymbol{\alpha}} \left\{ C_{\beta} \cancel{\sum_{j \in [p], j \neq i} w_j \mathcal{W}_1(\mathcal{D}_j, \mathcal{D}_i)} \right.$$

$$\left. + C_{\tilde{L}, \hat{L}_i} \min_{t \in [T]} d(\boldsymbol{\theta}, \boldsymbol{\theta}_t^{(i)}) \cancel{+ \delta_i} \right\}$$

Non-parametric loss surface

FLoRA: Federated Loss Surface Aggregation

Empirical Performance

- Gradient boosted trees, support vector machines, and neural networks
- 7 OpenML datasets and all loss surface aggregation schemes
- Improved performance over single-shot baseline, APLM performs best

Aggregate	ML Method	SGM	SGM+U	MPLM	APLM
FLoRA	HGB	6/0/1	6/0/1	7/0/0	7/0/0
Wins/Ties/Losses	SVM	4/0/2	4/0/2	3/0/3	5/0/1
	MLP	6/0/1	4/1/2	5/1/1	6/0/1
	Overall	16/0/4	14/1/5	15/1/4	**18/0/2**

FLoRA: Federated Loss Surface Aggregation

Empirical Performance

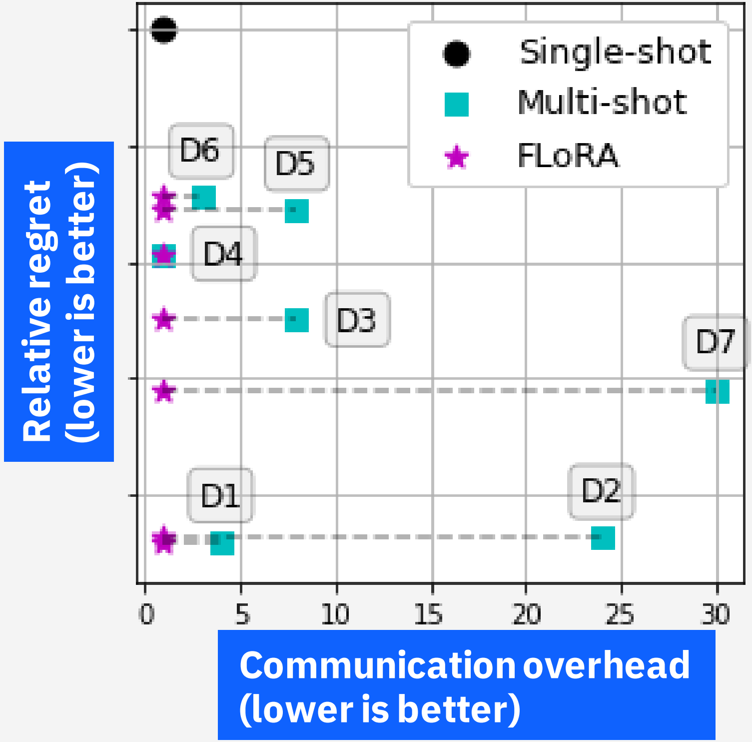
- Gradient boosted trees with 3 OpenML datasets
- Number of parties and data heterogeneity increased
- Performance drops as heterogeneity increases
- MPLM and APLM show most robust performance and graceful degradation

Data	p	γ_p	SGM	SGM+U	MPLM	APLM
EEG 14980 rows	3	1.01	0.14	0.12	0.11	0.12
	10	1.03	0.08	0.00	0.16	0.01
	25	1.08	0.35	0.92	0.17	0.04
	50	1.20	0.20	0.23	0.67	0.12
Electricity 45312 rows	3	1.01	0.17	0.14	0.09	0.12
	10	1.02	0.03	0.06	0.32	0.14
	25	1.04	0.40	0.42	1.42	0.89
	50	1.07	1.57	1.57	0.89	1.13
	100	1.14	1.45	1.47	0.48	1.11
Pollen 3848 rows	3	1.02	0.43	0.54	0.43	0.69
	6	1.10	1.02	0.91	0.54	0.56
	10	1.16	1.05	0.73	0.75	1.12

FLoRA: Federated Loss Surface Aggregation

Empirical Performance

- Gradient boosted trees with 7 OpenML datasets and APLM
- Comparison against single-shot and multi-shot baseline
- Improved performance over single-shot baselines
- Lower communication overhead compared to multi-shot for same performance



Conclusion

Novel capabilities of FLoRA

- Single-shot
- ML model agnostic
- Rigorous theoretical guarantees
- Strong empirical performance

Limitations

- Doesn’t apply to HPs absent in local HPO
- Aggregator HPs not handled

Thank you

Parikshit Ram
IBM Research

—

Parikshit.Ram@ibm.com