Personalized Federated Learning via Variational Bayesian Inference

Xu Zhang *[1], Yinchuan Li *[2], Wenpeng Li [2], Kaiyang Guo [2] and Yunfeng Shao [2]

[1] LSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences,

[2] Noah's Ark Lab, Huawei

*Equal contribution

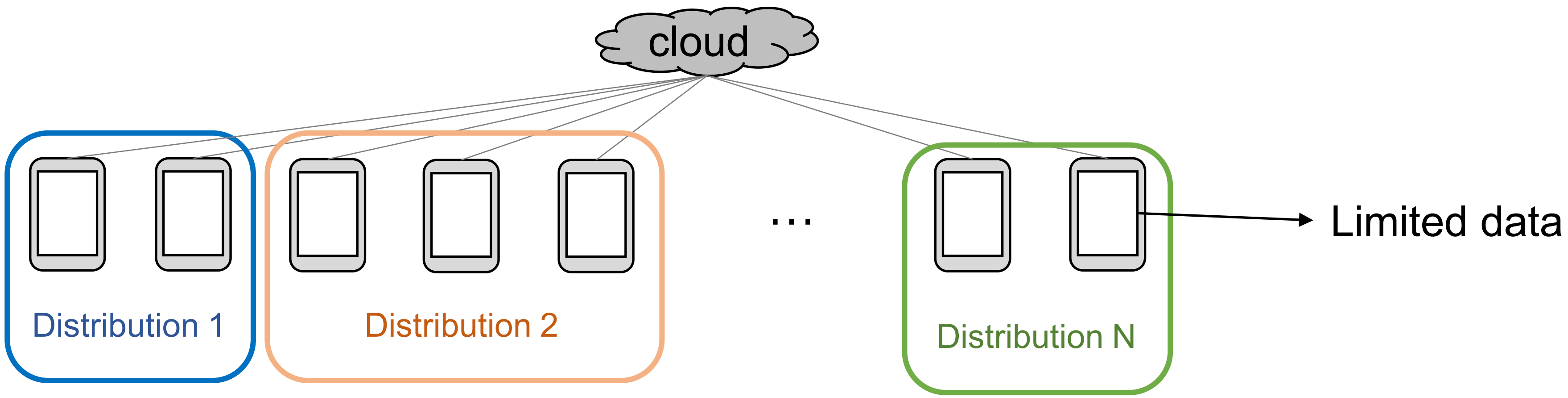
Background

Federated Learning (FL)

- model machine learning for distributed end devices while preserving their privacy

Challenges

- Non-i.i.d. data: due to the differences in user preferences, locations and living habits…
- Limited data: data from clients are usually limited



Goal

- design a Bayesian federated learning algorithm to address these two challenges together

Bayesian Neural Network

Considering a distributed system that contains one server and N clients. Let the i-th client satisfy the model

$$\boldsymbol{y}_j^i = f^i(\boldsymbol{x}_j^i) + \varepsilon_j^i,\ j = 1, \ldots, n,\ \varepsilon_j^i \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

where $\boldsymbol{x}_j^i \in \mathbb{R}^{s_0}$, $\boldsymbol{y}_j^i \in \mathbb{R}^{s_{L+1}}$ for $j = 1, \ldots, n$, $i = 1, \ldots, N$, $f^i(\cdot) : \mathbb{R}^{s_0} \to \mathbb{R}^{s_{L+1}}$ denotes a nonlinear function, n denotes the sample size and σ_ε denotes the variance of noise.

BNN aims to find the closest distribution to the posterior distribution in the variational family of distributions $\mathcal{Q}$

$$\min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} \mathsf{KL}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta} | \boldsymbol{D}))$$

posterior distribution   collected data

Using Bayes theorem gives the equivalent form

$$\min_{q(\boldsymbol{\theta}) \in \mathcal{Q}} -\mathbb{E}_{q(\boldsymbol{\theta})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{D})] + \mathsf{KL}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta}))$$

likelihood   prior distribution

Personalized Federated Bayesian Learning

Optimization Problem

$$\text{Server:} \quad \min_{w(\boldsymbol{\theta}) \in \mathcal{Q}_w} \left\{ F(w) \triangleq \frac{1}{N} \sum_{i=1}^{N} F_i(w) \right\}$$

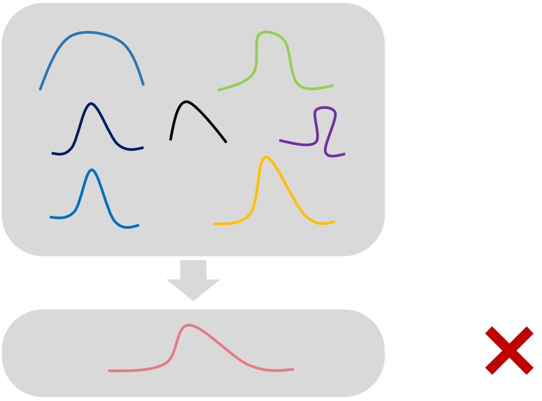
global distribution   distributions for global parameters

datasets $\boldsymbol{D}^i = (\boldsymbol{D}_1^i, \ldots, \boldsymbol{D}_n^i) \longleftarrow \boldsymbol{D}_j^i = (\boldsymbol{x}_j^i, \boldsymbol{y}_j^i)$

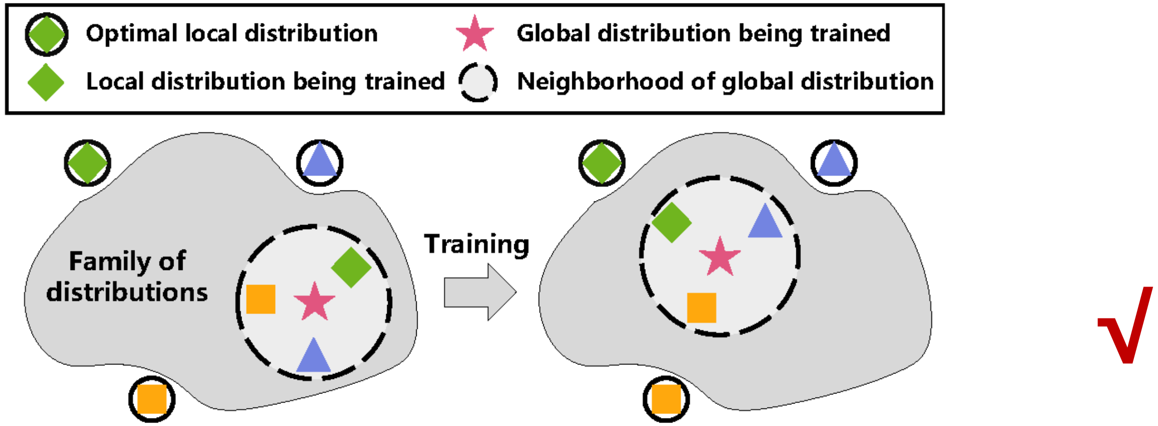
$$\text{Clients:} \quad F_i(w) \triangleq \min_{q^i(\boldsymbol{\theta}) \in \mathcal{Q}^i} \left\{ -\mathbb{E}_{q^i(\boldsymbol{\theta})}[\log p_{\boldsymbol{\theta}}^i(\boldsymbol{D}^i)] + \zeta \, \mathrm{KL}(q^i(\boldsymbol{\theta}) || w(\boldsymbol{\theta})) \right\}$$

The trained global distribution servers as the prior distribution

local distribution   distributions for local parameters   tradeoff parameter



Distributions cannot be aggregated directly



Find the distribution aligned with the client from the cloud distribution family

Personalized Federated Bayesian Learning

Theoretical Analysis

Define the Hellinger distance as follows

$$d^2(P^i_{\boldsymbol{\theta}}, P^i) = \mathbb{E}_{X^i}\Big(1 - \exp\{-[f^i_{\boldsymbol{\theta}}(X^i) - f^i(X^i)]^2/(8\sigma^2_\epsilon)\}\Big)$$

Theorem 1. Assume that $\{f^i\}$ are β-HÃűlder-smooth functions and the intrinsic dimension of data is d. With dominating probability, the following upper bound holds

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(P^i_{\boldsymbol{\theta}}, P^i)\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta} \le C_1 n^{-\frac{2\beta}{2\beta+d}}\log^{2\delta}(n),$$

where $\delta > 1$ and C_1 is a constant.

For bounded functions $\|f^i\|_\infty \le F$ and $\|f^i_{\boldsymbol{\theta}}\|_\infty \le F$, $i = 1, \ldots, N$,

$$\inf_{\{\|f^i_{\boldsymbol{\theta}}\| \le F\}_{i=1}^N} \sup_{\{\|f^i\|_\infty \le F\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(P^i_{\boldsymbol{\theta}}, P^i)\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta} \ge C_2 n^{-\frac{2\beta}{2\beta+d}}$$

The convergence rate of the generalization error is minimax optimal.

Personalized Federated Bayesian Learning

Algorithm

Server: $\min_{w(\boldsymbol{\theta})\in\mathcal{Q}_w}\left\{F(w)\triangleq\frac{1}{N}\sum_{i=1}^{N}F_i(w)\right\}$

Clients: $F_i(w)\triangleq\min_{q^i(\boldsymbol{\theta})\in\mathcal{Q}^i}\left\{-\mathbb{E}_{q^i(\boldsymbol{\theta})}[\log p^i_{\boldsymbol{\theta}}(\boldsymbol{D}^i)]+\zeta\,\mathrm{KL}(q^i(\boldsymbol{\theta})\|w(\boldsymbol{\theta}))\right\}$

Network is reparameterized by

$$\boldsymbol{v}=(\boldsymbol{\mu},\boldsymbol{\rho})\qquad\boldsymbol{\theta}=h(\boldsymbol{v},\boldsymbol{g})$$
$$\theta_m=h(v_m,g_m)=\mu_m+\log(1+\exp(\rho_m))\cdot g_m,\;\; g_m\sim\mathcal{N}(0,1)$$

Loss functions:

sample size batch size Monte Carlo sample size

$$\Omega^i(\boldsymbol{v})\approx-\frac{n}{b}\frac{1}{K}\sum_{j=1}^{b}\sum_{k=1}^{K}\log p^i_{h(\boldsymbol{v},\boldsymbol{g}_k)}(\boldsymbol{D}^i_j)+\zeta\,\mathrm{KL}(q^i_{\boldsymbol{v}}(\boldsymbol{\theta})\|w_{\boldsymbol{v}}(\boldsymbol{\theta}))$$
$$\Omega^i_w(\boldsymbol{v})=\mathrm{KL}(q^i_{\boldsymbol{v}}(\boldsymbol{\theta})\|w_{\boldsymbol{v}}(\boldsymbol{\theta}))$$

Algorithm 1 `pFedBayes`: Personalized Federated Learning via Bayesian Inference Algorithm

Cloud server executes:
- **Input** $T,R,S,\lambda,\eta,\beta,b,\boldsymbol{v}^0=(\boldsymbol{\mu}^0,\boldsymbol{\sigma}^0)$
- **for** $t=0,1,...,T-1$ **do**
  - **for** $i=1,2,...,N$ **in parallel do**
    - $\boldsymbol{v}_i^{t+1}\leftarrow$ ClientUpdate$(i,\boldsymbol{v}^t)$
  - $\mathbb{S}^t\leftarrow$ Random subset of clients with size S
  - $\boldsymbol{v}^{t+1}=(1-\beta)\boldsymbol{v}^t+\frac{\beta}{S}\sum_{i\in S^t}\boldsymbol{v}_i^{t+1}$

ClientUpdate$(i,\boldsymbol{v}^t)$**:**
- $\boldsymbol{v}^t_{w,0}=\boldsymbol{v}^t$
- **for** $r=0,1,...,R-1$ **do**
  - $\boldsymbol{D}^i_\Lambda\leftarrow$ sample a minibatch Λ with size b from $\boldsymbol{D}^i$
  - $\boldsymbol{g}_{i,r}\leftarrow$ Randomly draw K samples from $\mathcal{N}(0,1)$
  - $\Omega^i(\boldsymbol{v}^t_r)\leftarrow$ Use (26) and (27) with $\boldsymbol{g}_{i,r}$, $\boldsymbol{D}^i_\Lambda$ and $\boldsymbol{v}^t_r$
  - $\nabla_{\boldsymbol{v}}\Omega^i(\boldsymbol{v}^t_r)\leftarrow$ Back propagation w.r.t $\boldsymbol{v}^t_r$
  - $\boldsymbol{v}^t_r\leftarrow$ Update with $\nabla_{\boldsymbol{v}}\Omega^i(\boldsymbol{v}^t_r)$ using GD algorithms
  - $\Omega^i_w(\boldsymbol{v}^t_{w,r})\leftarrow$ Forward propagation w.r.t $\boldsymbol{v}$
  - $\nabla\Omega^i_w(\boldsymbol{v}^t_{w,r})\leftarrow$ Back propagation w.r.t $\boldsymbol{v}$
  - Update $\boldsymbol{v}^t_{w,r+1}$ with $\nabla\Omega^i_w(\boldsymbol{v}^t_{w,r})$ using GD algorithms
- return $\boldsymbol{v}^t_{w,R}$ to the cloud server

Personalized Federated Bayesian Learning

Experimental Results

- For small, medium and large datasets of MNIST/FMNIST, there were 50, 200, 900 training samples for each class, respectively.
- MNIST: PM outperforms other SOTA by 1.25%, 1.78% and 0.52%; GM outperforms other SOTA by 2.79%, 1.67% and 1.97%
- FMNIST: PM outperforms other SOTA by 0.42%, 0.63% and 0.79%

- For the small, medium and large datasets of CIFAR-10, there were 25, 100, 450 training samples for each class, respectively.
- CIFAR: PM outperforms other SOTA by 11.71%, 7.19% and 6.33%, GM outpersorms other SOTA by 3.47% and 3.49%.

Table 1: Results on MNIST, FMNIST and CIFAR-10. Best results are bolded.

Dataset	Method	Small (Acc. (%))		Medium (Acc. (%))		Large (Acc. (%))	
		PM	GM	PM	GM	PM	GM
MNIST	FedAvg	-	87.38±0.27	-	90.60±0.19	-	92.39±0.24
	Fedprox	-	87.65±0.30	-	90.66±0.17	-	92.42±0.23
	BNFed	-	78.70±0.69	-	80.02±0.60	-	82.95±0.22
	Per-FedAvg	89.29±0.59	-	95.19±0.33	-	98.27±0.08	-
	pFedMe	92.88±0.04	87.35±0.08	95.31±0.17	89.67±0.34	96.42±0.08	91.25±0.14
	HeurFedAMP	90.89±0.17	-	94.74±0.07	-	96.90±0.12	-
	pFedGP	85.96±2.30	-	91.96±0.97	-	95.66±0.43	-
	Ours	**94.13±0.27**	**90.44±0.45**	**97.09±0.13**	**92.33±0.76**	**98.79±0.13**	**94.39±0.32**
FMNIST	FedAvg	-	81.51±0.19	-	83.90±0.13	-	**85.42±0.14**
	Fedprox	-	**81.53±0.08**	-	**83.92±0.21**	-	85.32±0.14
	BNFed	-	66.54±0.64	-	69.68±0.39	-	70.10±0.24
	Per-FedAvg	79.79±0.83	-	84.90±0.47	-	88.51±0.28	-
	pFedMe	88.63±0.07	81.06±0.14	91.32±0.08	83.45±0.21	92.02±0.07	84.41±0.08
	HeurFedAMP	86.38±0.24	-	89.82±0.16	-	92.17±0.12	-
	pFedGP	86.99±0.41	-	90.53±0.35	-	92.22±0.13	-
	Ours	**89.05±0.17**	80.17±0.19	**91.95±0.02**	82.33±0.37	**93.01±0.10**	83.30±0.28
CIFAR-10	FedAvg	-	44.24±3.01	-	56.73±1.81	-	79.05±0.44
	Fedprox	-	43.70±1.38	-	57.35±3.11	-	**77.65±1.62**
	BNFed	-	34.00±0.16	-	39.52±0.56	-	44.37±0.19
	Per-FedAvg	33.96±1.12	-	52.98±1.21	-	69.61±1.21	-
	pFedMe	49.66±1.53	43.67±2.14	66.75±1.87	51.18±2.57	77.13±1.06	70.86±1.04
	HeurFedAMP	46.72±0.39	-	59.94±1.42	-	73.24±0.80	-
	pFedGP	43.66±0.32	-	58.54±0.40	-	72.45±0.19	-
	Ours	**61.37±1.40**	**47.71±1.19**	**73.94±0.97**	**60.84±1.26**	**83.46±0.13**	64.40±1.22

Thank you.