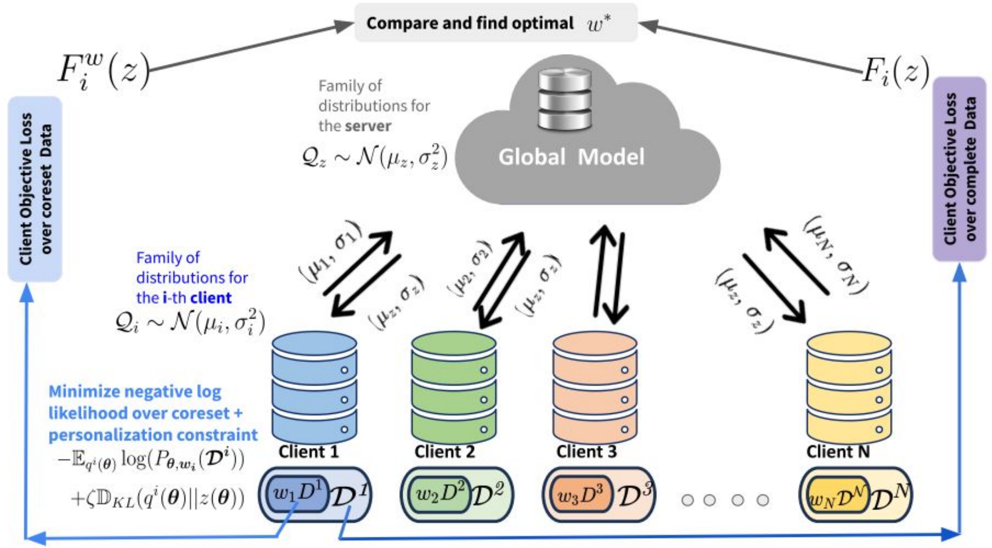
Bayesian Coresets for Personalized Federated Learning

ICLR 2024

Prateek Chanda, Shrey Modi, Ganesh Ramakrishnan
Dept. of Computer Science, IIT Bombay

Problem Setting: Personalized Federated Learning (PFL)



1 Server, **N** clients. Each client has individual data $\mathcal{D}_j^i = (\boldsymbol{x}_j^i, \boldsymbol{y}_j^i)$

Further ith client satisfy a regression model

$$\boldsymbol{y}_j^i = f^i(\boldsymbol{x}_j^i) + \epsilon_j^i$$

$f^i(\bullet) \Rightarrow \beta$-Hölder-smooth functions

Personalized Federated Learning Objectives

Client Side Objective

find the closest distribution from the family of distributions $\mathcal{Q}_i$ to match the posterior distribution via minimizing the KL-divergence as follows

$$\mathcal{F}_i(z) := \min_{q^i(\boldsymbol{\theta})\in\mathcal{Q}_i} \mathbb{D}_{KL}(q^i(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}^i)) \Leftrightarrow \min_{q^i(\boldsymbol{\theta})\in\mathcal{Q}_i} \overbrace{-\mathbb{E}_{q^i(\boldsymbol{\theta})}[\log P_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}}^i)]}^{\text{reconstruction error over } \boldsymbol{\mathcal{D}}} + \overbrace{\zeta\mathbb{D}_{KL}(q^i(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))}^{\text{regularization term}}$$

Server Side Objective

the global model tries to find the closest distribution in $\mathcal{Q}_z$ to the client's distribution by minimizing the aggregate KL divergence from all the clients as follows

$$\min_{z(\boldsymbol{\theta})\sim\mathcal{Q}_z} \mathcal{F}(z) := \frac{1}{N}\sum_{i=1} \mathcal{F}_i(z)$$

Bayesian Coreset Objectives

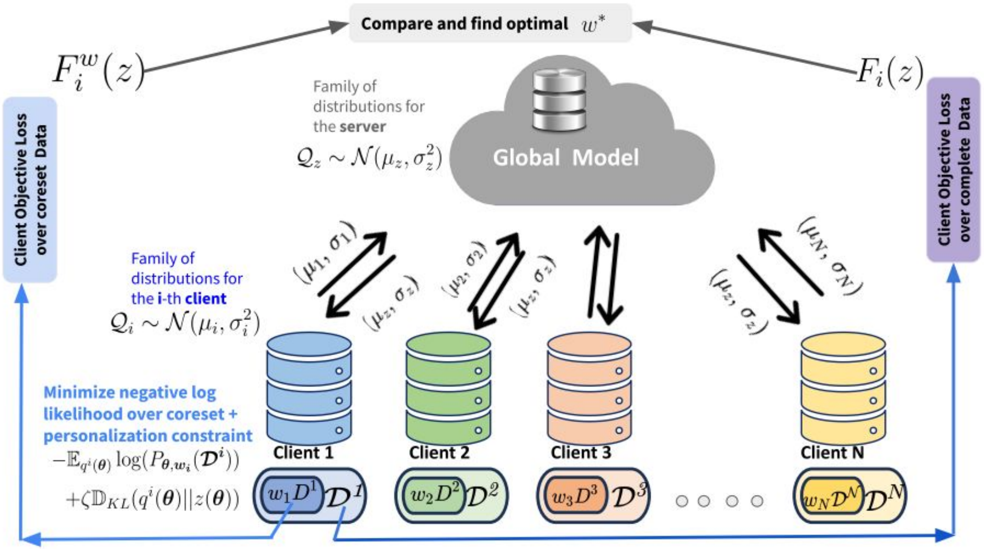
Assign to each client's data a *weight vector* that will act as the corresponding **coreset weight** for the i-th client.

Goal : to control the **deviation of coreset log-likelihood from the true log-likelihood via sparsity**

$$\arg\min_{\boldsymbol{w}_i \in \mathbb{R}^n} \mathcal{G}^i(\boldsymbol{w}_i) := \left\| \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}}^{\boldsymbol{i}}) - \mathcal{P}_{\boldsymbol{\theta},\boldsymbol{w}_i}(\boldsymbol{\mathcal{D}}^{\boldsymbol{i}}) \right\|_{\hat{\pi},2}^{2} \quad s.t. \ \ ||\boldsymbol{w}_i||_0 \leq k, \quad \forall i \in [N] \tag{3}$$

where the coreset weights $\boldsymbol{w}_i$ are considered over the data points for client i and $L^2(\hat{\pi})$-norm as the distance metric is considered in the embedding Hilbert Space. Specifically, $\hat{\pi}$ is the weighting distribution that has the same support as true posterior π. The above equation can be further approximated

Combining Coreset Optimization with PFL



Personalized Federated Learning Objectives

$$\arg\min_{\boldsymbol{w}_i \in \mathbb{R}^n} \mathcal{G}^i(\boldsymbol{w}_i) := \left\| \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}}^i) - \mathcal{P}_{\boldsymbol{\theta},\boldsymbol{w}_i}(\boldsymbol{\mathcal{D}}^i) \right\|_{\hat{\pi},2}^2 \quad s.t. \quad ||\boldsymbol{w}_i||_0 \leq k$$

Bayesian Coreset Optimization

- Connecting two optimization frameworks
- New Theoretical Bounds on Convergence Error Rate
- Better Efficient Performance w.r.t Random and Submodular based Subset selection strategies

Client Side Modifications

Client Side Objective

$$F_i(z) \triangleq \min_{q^i(\boldsymbol{\theta})\in\mathcal{Q}_i} \mathbb{D}_{KL}(q^i(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}^i})) \Leftrightarrow \min_{q^i(\boldsymbol{\theta})\in\mathcal{Q}_i} \overbrace{-\mathbb{E}_{q^i(\boldsymbol{\theta})}[\log P_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}^i})]}^{\text{reconstruction error over } \boldsymbol{\mathcal{D}}} + \zeta \overbrace{\mathbb{D}_{KL}(q^i(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))}^{\text{regularization term}} \quad (1)$$

Here $\pi(\boldsymbol{\theta})$ denotes the prior distribution and $P_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}^i})$ denotes the likelihood and ζ is a personalization constant which we include in our objective (Zhang et al., 2022b).

Modified Client Side Objective We now aim towards incorporating the coreset formulation in our federated learning setting from Eq: 1. Assuming the personalized bayesian coreset weights setup for each client i, we introduce a new modified client objective function

$$F_i^w(z) \triangleq \min_{q^i(\boldsymbol{\theta})\sim\mathcal{Q}_i} [-\mathbb{E}_{q^i(\boldsymbol{\theta})} \log(P_{\boldsymbol{\theta},\boldsymbol{w_i}}(\boldsymbol{\mathcal{D}^i})) + \zeta \mathbb{D}_{KL}(q^i(\boldsymbol{\theta})||z(\boldsymbol{\theta}))] \quad (5)$$

where $z(\boldsymbol{\theta})$ and $q^i(\boldsymbol{\theta})$ denote the global distribution and the local distribution for the i-the client that is to be optimized respectively.

Optimal client distribution based on optimal coreset weights

$$\min_{z(\boldsymbol{\theta})\sim\mathcal{Q}_z} \mathcal{F}(z) := \frac{1}{N}\sum_{i=1} \mathcal{F}_i(z)$$

Optimal Solution →

$$\boldsymbol{z}^*(\boldsymbol{\theta})$$

$$\boldsymbol{z}^*(\boldsymbol{\theta})$$

$$\hat{q}^i(\boldsymbol{\theta})$$

i^th Client's vanilla objective Solution

$$\hat{q}^i(\boldsymbol{\theta}; \boldsymbol{w})$$

i^th Client's coreset weighted objective Solution

Optimal client distribution based on optimal coreset weights

$$\{\boldsymbol{w}_i^*\} := \arg\min_{\boldsymbol{w}} \mathbb{D}_{KL}(\mathcal{F}_i^{\boldsymbol{w}}(z)_{\text{arg}} || \mathcal{F}_i(z)_{\text{arg}}) \quad \Leftrightarrow \quad \arg\min_{\boldsymbol{w}} \mathbb{D}_{KL}(\hat{q^i}(\boldsymbol{\theta}, \boldsymbol{w}) || \hat{q^i}(\boldsymbol{\theta})) \quad \|\boldsymbol{w_i}\|_0 \leq n_k$$

So we want to formulate a new objective function such that for each client we minimize the divergence between the two optimal distributions resulting from the coreset and normal objective functions.

Theoretical Results

$\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta},w}, \mathcal{P}^i)\hat{q^i}(\boldsymbol{\theta}; \boldsymbol{w})d\boldsymbol{\theta}$

Generalization Error Term

$f^i_{\boldsymbol{\theta}}$ Denotes the DNN Model for the i'th client

$f^i_{\boldsymbol{\theta},w}$ Denotes the DNN Model for the coreset weighted i'th client

we define the Hellinger Distance as follows: $d^2(\mathcal{P}^i_{\boldsymbol{\theta}}, \mathcal{P}^i) = \mathbb{E}_{X^i}\left(1 - e^{-\frac{[f^i_{\boldsymbol{\theta}}(X^i) - f^i(X^i)]^2}{8\sigma^2_\epsilon}}\right)$

Theoretical Results

Theorem 1. *The difference in the upper bound incurred in the overall generalization error of* CORESET-PFEDBAYES *as compared w.r.t that of* PFEDBAYES *is always upper bounded by a closed form positive function that depends on the coreset weights and coreset size-* $\Im(\boldsymbol{w}, n_k)$. *generalization error in the original full data setup*

$$\left[\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta}}, \mathcal{P}^i)\hat{q^i}(\boldsymbol{\theta})d\boldsymbol{\theta}\right]_{u.b.} - \left[\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta},w}, \mathcal{P}^i)\hat{q^i}(\boldsymbol{\theta};\boldsymbol{w})d\boldsymbol{\theta}\right]_{u.b.} \leq \Im(\boldsymbol{w}, n_k)$$

Implies: that the overall generalization error is in closed form of the coreset weights and coreset size and can be measured directly

Theoretical Results: Continued

Theorem 2. *The convergence rate of the generalization error under L^2 norm of* CoreSet-PFedBayes *is minimax optimal up to a logarithmic term (in order n_k) for bounded functions (β-Hölder-smooth functions) $\{f^i\}_{i=1}^N$, $\{f^i_{\boldsymbol{\theta}}\}_{i=1}^N$ and $\{f^i_{\boldsymbol{\theta},\boldsymbol{w}}\}_{i=1}^N$ where C_2, C_3 and δ' are constants and $\boldsymbol{\Lambda}$ being the intrinsic dimension of each client's data:*

$$\frac{C_F}{N}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\left(X^i\right)-f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta};\boldsymbol{w})d\boldsymbol{\theta}\leq C_2 n_k^{-\frac{2\beta}{2\beta+\boldsymbol{\Lambda}}}\log^{2\delta'}(n_k).$$

Implies: that the convergence rate of the generalization error is in logarithmic bounds of coreset size

Theoretical Results: Continued

Theorem 3. *The lower bound (l.b.) incurred for the deviation for the weighted coreset* CoreSet-PFedBayes (5) *generalization error is always higher than the lower bound of that for the original* PFedBayes *objective* (1) *with a delta difference* (**Error I** - **Error II**) *as* $\mathcal{O}(n_k^{-\frac{2\beta}{2\beta+\Lambda}})$

$$\underbrace{\left[\sum_{i=1}^{N}\int_{\Theta}\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}(X^i)-f^i(X^i)\|^2_{L^2}\,\hat{q}^i(\boldsymbol{\theta},\boldsymbol{w})d\boldsymbol{\theta}\right]_{l.b.}}_{\textit{Coreset weighted objective Generalization Error } (\textbf{Error I})} > \underbrace{\left[\sum_{i=1}^{N}\int_{\Theta}\|f^i_{\boldsymbol{\theta}}(X^i)-f^i(X^i)\|^2_{L^2}\,\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta}\right]_{l.b.}}_{\textit{Vanilla objective Generalization Error } (\textbf{Error II})}$$

Implies: that the generalization error suffers in the case due to limited coreset samples but that is bounded in closed form w.r.t. the coreset sample size.

Results

Method (Percentage = sampling fraction)	**MNIST**		**FashionMNIST**		**CIFAR**	
	Personal Model	Global Model	Personal Model	Global Model	Personal Model	Global Model
FedAvg (Full/ 50%)	-	**92.39**(90.60)	-	85.42(83.90)	-	**79.05**(56.73)
BNFed (Full / 50%)	-	82.95(80.02)	-	70.1(69.68)	-	44.37(39.52)
pFedMe (Full / 50%)	-	91.25(89.67)	**92.02**(84.71)	84.41(83.45)	**77.13**(66.75)	**70.86**(51.18)
perFedAvg (Full / 50%)	**98.27**	-	88.51(84.90)	-	69.61(52.98)	-
PFEDBAYES (Full / 50%)	**98.79**(90.88)	**97.21**(92.33)	**93.01**(85.95)	**93.30**(82.33)	**83.46**(73.94)	64.40(60.84)
RANDOMSUBSET (50%)	80.2	88.4	87.12	**90.75**	48.31	61.35
CORESET-PFEDBAYES (k = 50%)	**92.48**	**96.3**	**89.55**	**92.7**	69.66	**71.5**

(a) We report accuracies on both global and personal model for the current set of proposed methods across major datasets like **MNIST, CIFAR, FashionMNIST**. **Red** indicates the highest accuracy column-wise. Similarly **Orange** and **Magenta** indicates the 2nd and 3rd best modelwise accuracy. (-) indicates no accuracy reported due to very slow convergence of the corresponding algorithm. **Full indicates training on full dataset and 50% is on using half the data size after randomly sampling 50% of the training set.**

Results : Continued

Table 3: Comparative results of test accuracies across different coreset sample complexity

Method (Percentage = sampling fraction)	MNIST		FashionMNIST		CIFAR	
	Test Accuracy	Communication Rounds	Test Accuracy	Communication Rounds	Test Accuracy	Communication Rounds
PFedBayes (Full)	98.79	194	93.01	215	83.46	266
RandomSubset (50%)	80.2	135	87.12	172	48.31	183
CoreSet-PFedBayes (k = 50%)	92.48	98	89.55	93	69.66	112
CoreSet-PFedBayes (k = 30%)	90.17	84	88.16	72	59.12	70
CoreSet-PFedBayes (k = 15%)	88.75	62	85.15	38	55.66	32
CoreSet-PFedBayes (k = 10%)	85.43	32	82.64	24	48.25	16

(a) We report test accuracies across different sample complexity for datasets like **MNIST, CIFAR, Fashion-MNIST**. **Full indicates training on full dataset and 50% is on using half the data size after randomly sampling 50% of the training set.**

Results : Continued

Table 2: Comparative results of classwise global accuracies of all 9 methods on **3 different medical datasets** and **2 clients**

Method (Percentage = sampling fraction)	**COVID-19 Radiography Database**			**APTOS 2019 Blindness Detection**			**OCTMNIST Dataset**		
	Normal X-ray	COVID X-ray	Lung Opacity X-ray	Normal Retina	Mild Diabetic Retinopathy	Severe Diabetic Retinopathy	Normal Retina	DME	Drusen
Vanilla FedAvg (Full)	**0.914 ± 0.007**	**0.924 ± 0.005**	**0.898 ± 0.007**	**0.968 ± 0.023**	**0.927 ± 0.019**	**0.853 ± 0.004**	**0.908 ± 0.026**	0.837 ± 0.103	**0.855 ± 0.092**
PFEDBAYES(Full)	**0.953 ± 0.006**	**0.938 ± 0.004**	**0.902 ± 0.011**	**0.951 ± 0.057**	**0.941 ± 0.052**	**0.911 ± 0.028**	**0.926 ± 0.013**	0.851 ± 0.021	**0.874 ± 0.012**
Independent Learning (Full)	**0.898 ± 0.001**	**0.869 ± 0.002**	**0.884 ± 0.003**	**0.945 ± 0.025**	0.877 ± 0.049	0.830 ± 0.053	**0.890 ± 0.073**	0.798 ± 0.076	**0.890 ± 0.041**
RandomSub FedAvg (50%)	0.892 ± 0.024	0.670 ± 0.059	0.583 ± 0.033	0.918 ± 0.047	0.835 ± 0.091	0.832 ± 0.021	0.811 ± 0.070	0.753 ± 0.089	0.805 ± 0.068
LogDet FedAvg (50%)	0.887 ± 0.046	0.838 ± 0.086	0.810 ± 0.062	0.918 ± 0.027	**0.885 ± 0.082**	**0.850 ± 0.057**	**0.842 ± 0.046**	**0.897 ± 0.039**	0.845 ± 0.068
DispSum FedAvg (50%)	**0.907 ± 0.015**	**0.925 ± 0.049**	**0.812 ± 0.086**	**0.945 ± 0.043**	**0.890 ± 0.095**	**0.852 ± 0.061**	0.834 ± 0.044	**0.887 ± 0.082**	**0.863 ± 0.094**
DispMin FedAvg (50%)	0.866 ± 0.018	0.780 ± 0.045	0.751 ± 0.069	**0.963 ± 0.021**	0.851 ± 0.067	0.765 ± 0.033	0.831 ± 0.011	**0.892 ± 0.066**	0.835 ± 0.085
CORESET-PFEDBAYES (50%)	**0.932 ± 0.003**	**0.919 ± 0.013**	**0.871 ± 0.025**	0.921 ± 0.016	**0.894 ± 0.029**	**0.886 ± 0.017**	**0.916 ± 0.042**	0.805 ± 0.008	**0.816 ± 0.011**

(a) We report classwise accuracies for the current set of proposed methods for all 3 medical datasets. **Red** indicates the highest value in accuracy column-wise (i.e. for a particular class for a dataset across all 9 baselines). Similarly **Orange** and **Magenta** indicates the 2nd and 3rd best classwise accuracy. **Colors for Vanilla FedAvg, PFEDBAYES , CORESET-PFEDBAYES are grouped together** to primarily compare against subset selection strategies

Please visit our paper at : https://openreview.net/forum?id=uz7d2N2zul

Project Page: https://coresetfederatedlearning.github.io/

THANK YOU