Supplementary Material: Bayesian Coreset Optimization for Personalized Federated Learning

In this supplementary material we discuss extensively on the proofs. involved for the theoretical analysis for CORESET-PFEDBAYES along with more fine-grained experimental details and corresponding baselines.

9 PROOFS

Here we discuss the proofs involved with particular propositions and theorems specified in the Theoretical Contributions of this paper. Utilising the assumptions taken in Zhang et al. (2022b), Polson & Ročková (2018) we consider the analysis for equal-width Bayesian Neural network.

Assumption 1: The widths of the neural network are equal width i.e. $s_i = M$.

Assumption 2: Each individual client $i \in [N]$ has equal coreset size of samples $n_k < n$.

Assumption 3: Parameters s_0, n (*total client dataset size*) n_k (*coreset client dataset size*), M, L (number of DNN layers as per Section 3) are large enough such that the sequence σ_n^2 is bounded as follows

$$\sigma_n^2 = \frac{T}{8n} A \leq \Omega^2,$$

where $\tau = \Omega M$ and

$$A = \log^{-1}(3s_0 M) \cdot (2\tau)^{-2(L+1)} \left[\left(s_0 + 1 + \frac{1}{\tau - 1} \right)^2 + \frac{1}{(2\tau)^2 - 1} + \frac{2}{(2\tau - 1)^2} \right]^{-1}.$$

Here T indicates the total number of parameters as defined in Section 3

Similarly, utilising the coreset regime, we have the following:

$$\sigma_{n_k}^2 = \frac{T}{8n_k} A \leq \Omega^2,$$

Since $n_k << n$, hence $\sigma_{n_k}^2 >> \sigma_n^2$

Assumption 4: We consider 1-Lipschitz continuous activation function $\sigma(\bullet)$

We also define here a few terms as defined in (Zhang et al., 2022b) which would be useful for our following proof proposals as well.

Definition 2. *Preliminaries and Definitions required for theoretical proofs under* PFEDBAYES

$$d^2(\mathcal{P}_{\boldsymbol{\theta}}^i, \mathcal{P}^i) = \mathbb{E}_{X^i}(1 - e^{-\frac{\left[f_{\boldsymbol{\theta}}^i(X^i) - f^i(X^i)\right]^2}{8\sigma_\epsilon^2}})$$

$$r_n = ((L+1)T/n)\log M + (T/n)\log\left(s_0\sqrt{n/T}\right)$$

$$\xi_n^i = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}(L, \boldsymbol{\mathcal{S}}), \|\boldsymbol{\theta}\|_\infty \leq \Omega} \left\| f_{\boldsymbol{\theta}}^i - f^i \right\|_\infty^2,$$

$$\varepsilon_n = n^{-\frac{1}{2}} \sqrt{(L+1)T\log M + T\log\left(s_0\sqrt{n/T}\right)} \log^\delta(n) = \sqrt{r_n}\log^\delta(n),$$

where $\delta > 1$

Here r_n indicates the variational error incurred due to the Bayesian approximation to the true posterior distribution in Equation 1 and ξ_n^i indicates the approximation error incurred during regression w.r.t the actual function to be learnt.

Similarly for the coreset size n_k we define the following:

Definition 3. *Preliminaries and Definitions required for theoretical proofs under* CORESET-PFEDBAYES

$$d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}}, \mathcal{P}^i) = \mathbb{E}_{X^i}(1 - e^{-\frac{\left[f^i_{\boldsymbol{\theta},\boldsymbol{w}}(X^i) - f^i(X^i)\right]^2}{8\sigma_\epsilon^2}})$$

$$\xi^i_{n_k} = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(L,\boldsymbol{\mathcal{S}}),\|\boldsymbol{\theta}\|_\infty\leq\Omega} \left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}} - f^i\right\|^2_\infty$$

$$r_{n_k} = ((L+1)T/n_k)\log M + (T/n_k)\log\left(s_0\sqrt{n_k/T}\right)$$

$$\varepsilon_{n_k} = {n_k}^{-\frac{1}{2}}\sqrt{(L+1)T\log M + T\log\left(s_0\sqrt{n_k/T}\right)}\log^\delta(n_k) = \sqrt{r_{n_k}}\log^\delta(n_k)$$

Lemma 1. *The Hellinger Distance from Definition 1 is symmetrical in its arguments $\mathcal{P}^i_{\boldsymbol{\theta}}$ and $\mathcal{P}^i$.*

Proof. It is easy to show that,

$$d^2(\mathcal{P}^i_{\boldsymbol{\theta}}, \mathcal{P}^i) = \mathbb{E}_{X^i}(1 - e^{-\frac{[f^i_{\boldsymbol{\theta}}(X^i) - f^i(X^i)]^2}{8\sigma_\epsilon^2}}) \tag{10}$$

$$= \mathbb{E}_{X^i}(1 - e^{-\frac{[f^i(X^i) - f^i_{\boldsymbol{\theta}}(X^i)]^2}{8\sigma_\epsilon^2}}) \tag{11}$$

$$= d^2(\mathcal{P}^i, \mathcal{P}^i_{\boldsymbol{\theta}}) \tag{12}$$

□

PROOF. OF THEOREM 1

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- $\mathfrak{F}(\boldsymbol{w}, n_k)$. generalization error in the original full data setup*

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

Proof. Let us define $\log\eta(P^i_{\boldsymbol{\theta}}, P^i) = l_n(P^i_{\boldsymbol{\theta}}, P^i)/\zeta + nd^2(P^i_{\boldsymbol{\theta}}, P^i)$.

Using Theorem 3.1 of Pati et al. (2018) with probability at most $e^{-Cn_k\varepsilon^2_{n_k}}$, where C is a constant, with high probability for CORESET-PFEDBAYES we have

$$\int_\Theta \eta(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}}, \mathcal{P}^i)z^*(\boldsymbol{\theta})d\boldsymbol{\theta} \leq e^{Cn_k\varepsilon^2_{n_k}} \tag{13}$$

Similarly with high probability at most $e^{-Cn\varepsilon^2_n}$ for the vanilla PFEDBAYES

$$\int_\Theta \eta(\mathcal{P}^i_{\boldsymbol{\theta}}, \mathcal{P}^i)z^*(\boldsymbol{\theta})d\boldsymbol{\theta} \leq e^{Cn\varepsilon^2_n} \tag{14}$$

Using Lemma A.1 from Zhang et al. (2022b) we know that for any probability measure μ and any measurable function h with $e^h \in L_1(\mu)$,

$$\log\int e^{h(\eta)}\mu(d\eta) = \sup_\rho\left[\int h(\eta)\rho(d\eta) - \mathbb{D}_{KL}(\rho\|\mu)\right]$$

Further, we let $l_n\left(P^i, P^i_{\boldsymbol{\theta}}\right)$ is the log-likelihood ratio of P^i and $P^i_{\boldsymbol{\theta}}$

$$l_n\left(P^i, P^i_{\theta}\right) = \log \frac{\mathcal{P}^i\left(\boldsymbol{D}^i\right)}{\mathcal{P}^i_{\theta}\left(\boldsymbol{D}^i\right)}.$$

Hence,

$$\begin{aligned} nd^2(P^i_{\boldsymbol{\theta}}, P^i) &= l_n(P^i_{\boldsymbol{\theta}}, P^i)/\zeta - \log \eta(P^i_{\boldsymbol{\theta}}, P^i) \\ &= l_n(P^i, P^i_{\boldsymbol{\theta}})/\zeta - \log \eta(P^i, P^i_{\theta}) \quad \text{since } d^2(P^i_{\boldsymbol{\theta}}, P^i) = d^2(P^i, P^i_{\boldsymbol{\theta}}) \text{ from Lemma 1} \end{aligned}$$

This follows from 9

Similarly, for the weighted likelihood based Hellinger Distance,

$$n_k d^2(P^i_{\boldsymbol{\theta},\boldsymbol{w}}, P^i) = l_n(P^i_{\boldsymbol{\theta},\boldsymbol{w}}, P^i)/\zeta - \log \eta(P^i_{\boldsymbol{\theta},\boldsymbol{w}}, P^i) \tag{15}$$

By using Lemma A.1 with $h(\eta) = \log \eta\left(P^i_{\boldsymbol{\theta}}, P^i\right), \mu = z^\star(\boldsymbol{\theta})$ and $\rho = \hat{q}^i(\boldsymbol{\theta})$, we obtain

$$\begin{aligned} \int_\Theta d^2\left(P^i_{\boldsymbol{\theta}}, P^i\right) \hat{q}^i(\boldsymbol{\theta}) d\boldsymbol{\theta} &\leq \frac{1}{n}\left[\frac{1}{\zeta}\int_\Theta l_n\left(P^i, P^i_{\boldsymbol{\theta}}\right) \hat{q}^i(\boldsymbol{\theta}) d\boldsymbol{\theta} + \mathbb{D}_{KL}\left(\hat{q}^i(\boldsymbol{\theta}) \| z^\star(\boldsymbol{\theta})\right) + \log \int_\Theta \eta\left(P^i_{\boldsymbol{\theta}}, P^i\right) z^\star(\boldsymbol{\theta}) d\boldsymbol{\theta}\right] \\ &\leq \frac{1}{n}\left[\frac{1}{\zeta}\int_\Theta l_n\left(P^i, P^i_{\boldsymbol{\theta}}\right) \hat{q}^i(\boldsymbol{\theta}) d\boldsymbol{\theta} + \mathbb{D}_{KL}\left(\hat{q}^i(\boldsymbol{\theta}) \| z^\star(\boldsymbol{\theta})\right)\right] + C\varepsilon_n^2 \end{aligned}$$

$$\int_\Theta d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}}, \mathcal{P}^i) q^i(\hat{\boldsymbol{\theta}}, \boldsymbol{w}) d\boldsymbol{\theta} \leq \frac{1}{n_k}\left[\frac{1}{\zeta}\int_\Theta l_n\left(P^i, P^i_{\boldsymbol{\theta},\boldsymbol{w}}\right) q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w}) d\boldsymbol{\theta} + \mathbb{D}_{KL}\left(q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w}) \| z^\star(\boldsymbol{\theta})\right)\right] + C\varepsilon^2_{n_k}$$

Utilising analysis under Supplementary in (Bai et al., 2020), there exists an upper bound for the term

$$\int_\Theta l_n\left(P^i, P^i_{\boldsymbol{\theta}}\right) \hat{q}^i(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq C^{''}(nr_n + n\xi^i_n) \tag{16}$$

Lemma 2 from (Zhang et al., 2022b) provides the upper bound for the KL divergence term

$$\mathbb{D}_{KL}\left(\hat{q}^i(\boldsymbol{\theta}) \| z^\star(\boldsymbol{\theta})\right) \leq C^{'}(nr_n) \tag{17}$$

Therefore we can write the following expression that captures the weighted Hellinger distance displacement given in our coreset framework CORESET-PFEDBAYES as compared to PFEDBAYES

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta}},\mathcal{P}^i)q^i(\hat{\boldsymbol{\theta}})d\boldsymbol{\theta} - \frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}},\mathcal{P}^i)q^i(\hat{\boldsymbol{\theta}},\boldsymbol{w})d\boldsymbol{\theta}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\left[\frac{1}{\zeta}\int_{\Theta} l_n\left(P^i,P^i_{\boldsymbol{\theta}}\right)\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta} + \mathbb{D}_{KL}\left(\hat{q}^i(\boldsymbol{\theta})\|z^\star(\boldsymbol{\theta})\right)\right] + C\varepsilon_n^2 -$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_k}\left[\frac{1}{\zeta}\int_{\Theta} l_n\left(P^i,P^i_{\boldsymbol{\theta},w}\right)q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w})d\boldsymbol{\theta} + \mathbb{D}_{KL}\left(q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w})\|z^\star(\boldsymbol{\theta})\right)\right] - C\varepsilon_{n_k}^2$$

Using Eq:(17) and Eq:(16)

$$\leq C\varepsilon_n^2 - C\varepsilon_{n_k}^2 + n\left(C^{'}\zeta r_n + \frac{C^{''}}{N}\sum_{i=1}^{N}\xi_n^i\right) - n_k\left(C^{'}\zeta r_{n_k} + \frac{C^{''}}{N}\sum_{i=1}^{N}\xi_{n_k}^i\right)$$

$$\leq C(\varepsilon_n^2 - \varepsilon_{n_k}^2) + \zeta C^{'}(nr_n - n_k r_{n_k}) + \frac{C^{''}}{N}\sum_{i=1}^{N}(n\xi_n^i - n_k\xi_{n_k}^i)$$

$$= C\underbrace{(\varepsilon_n^2 - \varepsilon_{n_k}^2)}_{\substack{\text{Estimation error}\\ \text{Type I } \textbf{Drift}}} + \zeta C^{'}\underbrace{(nr_n - n_k r_{n_k})}_{\substack{\text{Estimation error}\\ \text{Type II } \textbf{Drift}}} + \frac{C^{''}}{N}\underbrace{\sum_{i=1}^{N}(n\xi_n^i - n_k\xi_{n_k}^i)}_{\text{Approximation Error } \textbf{Drift}}$$

$$= \underbrace{\mathfrak{S}(\boldsymbol{w},n_k)}_{\geq 0}$$

Where $\mathfrak{S}(\boldsymbol{w},n_k) = C(\varepsilon_n^2 - \varepsilon_{n_k}^2) + \zeta C^{'}(nr_n - n_k r_{n_k}) + \frac{C^{''}}{N}\sum_{i=1}^{N}(n\xi_n^i - n_k\xi_{n_k}^i)$ where each of the coefficients of the closed form function are constants related to $s_0, \beta, \boldsymbol{\Lambda}, L, M, \zeta$ and n_k

Using Lemma 2, 3 and with suitable assumptions on the Approximation drift error such that we see that each of the individual error terms are positive, there by indicating $\mathfrak{S}(\boldsymbol{w},n_k) \geq 0$ □

Lemma 2. *The Estimation error Type II* ***Drift*** *is a positive quantity i.e.* $nr_n > n_k r_{n_k}$.

Proof. By Definition,

$$r_n = ((L+1)T/n)\log M + (T/n)\log\left(s_0\sqrt{n/T}\right)$$

and

$$r_{n_k} = ((L+1)T/n_k)\log M + (T/n_k)\log\left(s_0\sqrt{n_k/T}\right)$$

Hence

$$\frac{r_n}{r_{n_k}} = \frac{n_k}{n} \times \frac{((L+1)T)\log M + (T)\log\left(s_0\sqrt{n/T}\right)}{((L+1)T)\log M + (T)\log\left(s_0\sqrt{n_k/T}\right)}$$

$$\frac{r_n}{r_{n_k}} = \frac{n_k}{n} \times \frac{(L+1)\log M + \log\left(s_0/\sqrt{T}\right) + \log\left(\sqrt{n}\right)}{(L+1)\log M + \log\left(s_0/\sqrt{T}\right) + \log\left(\sqrt{n_k}\right)}$$

Considering $(L+1)\log M + \log\left(s_0/\sqrt{T}\right)$ as a constant $\mathfrak{G}$ we have

$$\frac{r_n}{r_{n_k}} = \frac{n_k}{n} \times \frac{\mathfrak{G} + \log(\sqrt{n})}{\mathfrak{G} + \log(\sqrt{n_k})}$$

Thus,

$$\frac{nr_n}{n_k r_{n_k}} = \frac{\mathfrak{G} + \log(\sqrt{n})}{\mathfrak{G} + \log(\sqrt{n_k})}$$

It is clear since $\log(\bullet)$ is an increasing function for $n > n_k$ we have $nr_n > n_k r_{n_k}$. □

Lemma 3. *The Estimation error Type I* ***Drift*** *is a positive quantity i.e.* $\varepsilon_n^2 > \varepsilon_{n_k}^2$.

Proof. From the definition under Assumption 3

$$\varepsilon_{n_k} = {n_k}^{-\frac{1}{2}} \sqrt{(L+1)T \log M + T \log\left(s_0 \sqrt{n_k/T}\right)} \log^{\delta}(n_k) = \sqrt{r_{n_k}} \log^{\delta}(n_k)$$

Hence $\varepsilon_{n_k}^2 = r_{n_k} \log^{2\delta}(n_k)$. Similarly, $\varepsilon_n^2 = r_n \log^{2\delta}(n)$

$$\frac{\varepsilon_n^2}{\varepsilon_{n_k}^2} = \frac{r_n \log^{2\delta}(n)}{r_{n_k} \log^{2\delta}(n_k)} = \frac{nr_n \frac{\log^{2\delta}(n)}{n}}{n_k r_{n_k} \frac{\log^{2\delta}(n_k)}{n_k}}$$

From Lemma 2 we know that $nr_n > n_k r_{n_k}$, hence

$$\frac{\varepsilon_n^2}{\varepsilon_{n_k}^2} > \frac{\frac{\log^{2\delta}(n)}{n}}{\frac{\log^{2\delta}(n_k)}{n_k}} > 1$$

This follows due to the increasing nature of the function.

□

PROOF. OF THEOREM 2

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:*

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

and

$$\inf_{\left\{\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\right\|_\infty \leq F\right\}_{i=1}^N \left\{\|f^i\|_\infty \leq F\right\}_{i=1}^N} \frac{C_F}{N} \sum_{i=1}^{N} \int_{\boldsymbol{\theta}} \left\| f^i_{\boldsymbol{\theta},\boldsymbol{w}}\left(X^i\right) - f^i\left(X^i\right) \right\|_{L^2}^2 \hat{q}^i(\boldsymbol{\theta};\boldsymbol{w}) d\boldsymbol{\theta} \geq C_3 {n_k}^{-\frac{2\beta}{2\beta+\boldsymbol{\Lambda}}}$$

where n_k *denotes the coreset size per client dataset and* n *denotes the original per client dataset size and* $\frac{d^2\left(P^i_{\boldsymbol{\theta},\boldsymbol{w}}, P^i\right)}{\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}(X^i) - f^i(X^i)\right\|_{L^2}^2} \geq \frac{1-\exp\left(-\frac{4F^2}{8\sigma_\epsilon^2}\right)}{4F^2} \triangleq C_F$.

We present the choice of T for a typical class of functions. We already assumed that $\{f^i\}$ are β-Hölder-smooth functions (Definition 4. (Nakada & Imaizumi, 2020)) and the intrinsic dimension of data is $\mathbf{\Lambda}$.

From our above theorem result from Theorem: 1 we say the following:

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}},\mathcal{P}^i)\hat{q}^i(\boldsymbol{\theta};\mathbf{w})d\boldsymbol{\theta} \leq C\varepsilon^2_{n_k} + C' r_{n_k} + \frac{C''}{N\zeta}\sum_{i=1}^{N}\xi^i_{n_k} \tag{18}$$

Utilising Corollary 6 in (Nakada & Imaizumi, 2020) , the approximation error is upper-bounded as follows

$$\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}} - f^i\right\|_{\infty} \leq C_0 T^{-\frac{\beta}{\mathbf{\Lambda}}}$$

where $C_0 > 0$ is a constant related to s_0, β and $\mathbf{\Lambda}$

Thus from the above definitions 2 and 3, we have the following

$$\xi^i_n, \xi^i_{n_k} \leq C_0 T^{-\frac{2\beta}{\mathbf{\Lambda}}}, i = 1, \ldots, N$$

Utilising the above upper bound in 18 and substituting $T = C_1 n^{\frac{\mathbf{\Lambda}}{2\beta+\mathbf{\Lambda}}}$, we get

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta} d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}},\mathcal{P}^i)\hat{q}^i(\boldsymbol{\theta};\mathbf{w})d\boldsymbol{\theta} &\leq C\varepsilon^2_{n_k} + C' r_{n_k} + \frac{C''}{N\zeta}\sum_{i=1}^{N} C_0 T^{-\frac{2\beta}{\mathbf{\Lambda}}} \\
&\leq C r_{n_k}\log^{2\delta}(n_k) + C' r_{n_k} + \frac{C''}{N\zeta}\sum_{i=1}^{N} C_0 T^{-\frac{2\beta}{\mathbf{\Lambda}}} \;\because \varepsilon^2_{n_k} = r_{n_k}\log^{2\delta}(n_k) \\
&\leq C_2 n_k^{-\frac{2\beta}{2\beta+\mathbf{\Lambda}}}\log^{2\delta'}(n_k)\big[\text{ substituting } T \text{ in } r_{n_k}\big]
\end{aligned}$$

where $\delta' > \delta > 1$, and $C_1, C_2 > 0$ are constants related to $s_0, \beta, \mathbf{\Lambda}, L, M, \zeta$ and n_k.

Similar to Theorem 1.1 from (Bai et al., 2020) and Theorem 1 from (Zhang et al., 2022b) norm, we can write the following

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

Now, using the minimax lower bound under L^2 norm in Theorem 8 of (Nakada & Imaizumi, 2020), we see that for coreset regime the same formulation holds similar to our original setting as shown in (Zhang et al., 2022b)

$$\inf_{\left\{\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\right\|_{\infty}\leq F\right\}_{i=1}^{N}\left\{\|f^i\|_{\infty}\leq F\right\}_{i=1}^{N}} \frac{C_F}{N}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\left(X^i\right) - f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta};\boldsymbol{w})d\boldsymbol{\theta} \geq C_3 n_k^{-\frac{2\beta}{2\beta+\mathbf{\Lambda}}}$$

where $C_3 > 0$ is a constant.

Combining the above two equations, the convergence rate of the generalization error of the coreset weighted objective is minimax optimal upto a logarithmic term for bounded functions $\{f^i_{\boldsymbol{\theta},\boldsymbol{w}}\}_{i=1}^{N}$ and $\{f^i_{\boldsymbol{\theta}}\}_{i=1}^{N}$.

PROOF. OF THEOREM 3

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* $\mathcal{O}(n_k^{-\frac{2\beta}{2\beta+\Lambda}})$

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

Proof. As we know $n_k < n$ hence $C_3 n_k^{-\frac{2\beta}{2\beta+\Lambda}} > C_3 n^{-\frac{2\beta}{2\beta+\Lambda}}$ ($\because C_3$ is a constant independent of n or n_k), which therefore means that inequality holds in the lower bound (l.b.) of the two expressions (shown by the previous proposition 2).

$$\left[\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\left(X^i\right)-f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta},\boldsymbol{w})d\boldsymbol{\theta}\right]_{l.b.} > \left[\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta}}\left(X^i\right)-f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta}\right]_{l.b.}$$

Let us denote $\boldsymbol{\Delta}^{l.b}_{deviation}$ as follows

$$\boldsymbol{\Delta}^{l.b}_{deviation} = \left[\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta},\boldsymbol{w}}\left(X^i\right)-f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta},\boldsymbol{w})d\boldsymbol{\theta}\right]_{l.b.} - \left[\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left\|f^i_{\boldsymbol{\theta}}\left(X^i\right)-f^i\left(X^i\right)\right\|^2_{L^2}\hat{q}^i(\boldsymbol{\theta})d\boldsymbol{\theta}\right]_{l.b.}$$

And the $\boldsymbol{\Delta}^{l.b.}_{deviation}$ term is given by $\left(C_3 n_k^{-\frac{2\beta}{2\beta+\Lambda}} - C_3 n^{-\frac{2\beta}{2\beta+\Lambda}}\right) \approx \mathcal{O}(n_k^{-\frac{2\beta}{2\beta+\Lambda}})$. □

PROOF. OF THEOREM 4

Theorem 4. *The lower bound incurred in the overall generalization error across all N clients of* CORESET-PFEDBAYES *is always higher compared to that of the generalization error in the original full data setup*

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

Proof. It is easy to show since from Theorem 3, we know the lower bounds for the individual terms and also since $n > n_k$ holds, hence we can rewrite as follows:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\int_{\Theta}d^2(\mathcal{P}^i_{\boldsymbol{\theta},\boldsymbol{w}},\mathcal{P}^i)q^i(\hat{\boldsymbol{\theta}},\boldsymbol{w})d\boldsymbol{\theta} - \frac{1}{N}\sum_{i=1}^{N}\int_{\Theta}d^2(\mathcal{P}^i_{\boldsymbol{\theta}},\mathcal{P}^i)q^i(\hat{\boldsymbol{\theta}})d\boldsymbol{\theta}& \\ \geq C_3 n_k^{-\frac{2\beta}{2\beta+\Lambda}} - C_3 n^{-\frac{2\beta}{2\beta+\Lambda}}& \\ \geq 0&\end{aligned}$$

The implication of this proof states that the overall error incurred due to coreset weighted deviation is always more than that of the original deviation which can be measured approximately in order of n_k, the coreset sample size. □

Proposition 1. *The gradient of the first term in Equation 7 i.e.*

$$\nabla_w \mathbb{D}_{KL}(\hat{q^i}(\boldsymbol{\theta};\mathbf{w})||\hat{q^i}(\boldsymbol{\theta}))$$

is given by the following expression

$$\int_{\Theta} \nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \Big[\log \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) + 1 - \log \hat{q^i}(\boldsymbol{\theta}) \Big] d\boldsymbol{\theta}$$

where

$\nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) = \frac{\hat{q^i}(\boldsymbol{\theta}; \boldsymbol{w})}{\varrho^i(\boldsymbol{\theta_{i,m}}; \boldsymbol{w})} g_m^{'}(\boldsymbol{w}) + g_m(\boldsymbol{w}) \nabla_w \prod_{k \neq m}^{T} \varrho^i(\boldsymbol{\theta_{i,k}}; \boldsymbol{w})$ *and* $q^i(\boldsymbol{\theta}; \boldsymbol{w}) = \prod_{m=1}^{T} \varrho^i(\boldsymbol{\theta}_{i,m}; \boldsymbol{w})$

Proof.

$$\begin{aligned}
&\nabla_w \mathbb{D}_{KL}(\hat{q^i}(\theta; \mathbf{w}) || \hat{q^i}(\theta)) \\
&= \nabla_w \mathrm{E}_{\hat{q^i}(\theta; \mathbf{w})} \Big[\log \hat{q^i}(\theta; \mathbf{w}) - \log \hat{q^i}(\theta) \Big] \\
&= \nabla_w \Big[\int_{\Theta} \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \log \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) d\boldsymbol{\theta} - \int_{\Theta} \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \log \hat{q^i}(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big] \\
&= \Big[\int_{\Theta} \nabla_w \Big(\hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \log \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \Big) d\boldsymbol{\theta} - \int_{\Theta} \nabla_w \Big(\hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \log \hat{q^i}(\boldsymbol{\theta}) \Big) d\boldsymbol{\theta} \Big] \\
&= \Big[\int_{\Theta} \Big(\log q^i(\hat{\boldsymbol{\theta}}; \mathbf{w}) \nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) + \nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \Big) d\boldsymbol{\theta} - \int_{\Theta} \log \hat{q^i}(\boldsymbol{\theta}) \nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) d\boldsymbol{\theta} \Big] \\
&= \int_{\Theta} \nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) \Big[\log \hat{q^i}(\boldsymbol{\theta}; \mathbf{w}) + 1 - \log \hat{q^i}(\boldsymbol{\theta}) \Big] d\boldsymbol{\theta}
\end{aligned} \tag{19}$$

In order to compute the gradient $\nabla_w \hat{q^i}(\boldsymbol{\theta}; \mathbf{w})$, the following objective can be utilized.
Let $z^*(\boldsymbol{\theta})$ be the optimal variable solution to Equation (5).

$$\begin{aligned}
&\nabla_{q^i(\theta)} F_i^w(z^*) \Big|_{q^i(\hat{\theta}; \mathbf{w})} = 0 \\
\implies &\underbrace{\nabla_{q^i(\boldsymbol{\theta})} \int_{\Theta} - \log \mathcal{P}_{\theta, w}(\boldsymbol{\mathcal{D}}^i) q^i(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big|_{q^i(\hat{\theta}; \mathbf{w})}}_{\textbf{First Part}} + \underbrace{\zeta \nabla_{q^i(\boldsymbol{\theta})} \mathbb{D}_{KL}(q^i(\boldsymbol{\theta}) || z^*(\boldsymbol{\theta})) \Big|_{q^i(\hat{\theta}; \mathbf{w})}}_{\textbf{Second Part}} = 0
\end{aligned}$$

For the **first part**,

$$\begin{aligned}
&\nabla_{q^i(\theta)} \int_{\Theta} - \log \mathcal{P}_{\theta, w}(\boldsymbol{\mathcal{D}}^i) q^i(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big|_{q^i(\hat{\theta}; \mathbf{w})} \\
&= \int_{\Theta} \nabla_{q^i(\theta)} \Big[- \log \mathcal{P}_{\theta, w}(\boldsymbol{\mathcal{D}}^i) q^i(\boldsymbol{\theta}) \Big] d\boldsymbol{\theta} \Big|_{q^i(\hat{\theta}; \mathbf{w})} \\
&= \underbrace{\int_{\Theta} \Big[-q^i(\boldsymbol{\theta}) \nabla_{q^i(\theta)} \log \mathcal{P}_{\theta, w}(\boldsymbol{\mathcal{D}}^i) + \log \mathcal{P}_{\theta, w}(\boldsymbol{\mathcal{D}}^i) \Big] d\boldsymbol{\theta}}_{\textbf{Modified First part}} \Big|_{q^i(\hat{\theta}; \mathbf{w})}
\end{aligned} \tag{20}$$

By the assumption that the distribution $q^i(\boldsymbol{\theta})$ satisfies mean-field decomposition i.e.

$$\begin{aligned}
q^i(\boldsymbol{\theta}) &= \prod_{m=1}^{T} \mathcal{N}(\theta_{i,m}, \sigma_n^2) \\
&= \prod_{m=1}^{T} \varrho^i(\boldsymbol{\theta}_{i,m})
\end{aligned} \tag{21}$$

Let us denote $\mathcal{M}_w = \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)$.

Therefore, we extract out the following portion from (20): $\nabla_{q^i(\theta)} \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)$

$$\nabla_{q^i(\theta)} \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) = \nabla_{q^i(\theta)} \log \mathcal{M}_w \tag{22}$$

We now consider the individual partial differentials here

$$\frac{\partial}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} \log \mathcal{M}_w = \frac{1}{\mathcal{M}_w} \frac{\partial \mathcal{M}_w}{\partial w} \frac{\partial w}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} \tag{23}$$

Thus, we can rewrite (20) from the perspective of individual components of $q^i(\boldsymbol{\theta})$ as follows:

$$\begin{aligned} &\int_\Theta \left[-q^i(\boldsymbol{\theta}) \frac{1}{\mathcal{M}_w} \frac{\partial \mathcal{M}_w}{\partial w} \frac{\partial w}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} + \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \right] d\boldsymbol{\theta} \Big|_{q^i(\hat{\theta};\mathbf{w})} \\ = &\underbrace{\int_\Theta \left[-q^i(\boldsymbol{\theta}) \frac{1}{\mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)} \frac{\partial \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)}{\partial w} \frac{\partial w}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} + \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \right] d\boldsymbol{\theta}}_{\textbf{Modified First part}} \Big|_{q^i(\hat{\theta};\mathbf{w})} \end{aligned} \tag{24}$$

Now, we can rewrite the **second part** as follows:

$$\begin{aligned} &\zeta \nabla_{q^i(\theta)} \mathbb{D}_{KL}(q^i(\boldsymbol{\theta}) || z^*(\theta)) \Big|_{q^i(\hat{\boldsymbol{\theta}};\mathbf{w})} \\ = &\zeta \nabla_{q^i(\theta)} \left[\int_\Theta q^i(\boldsymbol{\theta}) \log q^i(\boldsymbol{\theta}) - q^i(\boldsymbol{\theta}) \log(z^*(\boldsymbol{\theta})) d\boldsymbol{\theta} \right] \Big|_{q^i(\hat{\theta};\mathbf{w})} \\ = &\zeta \int_\Theta \nabla_{q^i(\theta)} \left[q^i(\boldsymbol{\theta}) \log q^i(\boldsymbol{\theta}) - q^i(\boldsymbol{\theta}) \log(z^*(\boldsymbol{\theta})) \right] d\boldsymbol{\theta} \Big|_{q^i(\hat{\theta};\mathbf{w})} \\ = &\underbrace{\zeta \int_\Theta \Big(\log q^i(\boldsymbol{\theta}) + 1 - \log(z^*(\boldsymbol{\theta})) \Big) d\boldsymbol{\theta}}_{\textbf{Modified Second Part}} \Big|_{q^i(\hat{\theta};\mathbf{w})} \end{aligned} \tag{25}$$

□

Combining both the first and second part we get

$$\begin{aligned} &\int_\Theta \left[-q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w}) \frac{1}{\mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)} \frac{\partial \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)}{\partial w} \frac{\partial w}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} + \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \right] d\boldsymbol{\theta} \\ &\qquad + \zeta \int_\Theta \Big(\log q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w}) + 1 - \log(z^*(\boldsymbol{\theta})) \Big) d\boldsymbol{\theta} = 0 \\ \implies &\zeta \int_\Theta \Big(\log q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w}) + 1 - \log(z^*(\boldsymbol{\theta})) + \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \Big) d\boldsymbol{\theta} \\ &= \int_\Theta \left(q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w}) \frac{1}{\mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)} \frac{\partial \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)}{\partial w} \frac{\partial w}{\partial \varrho^i(\boldsymbol{\theta_{i,m}})} \right) d\boldsymbol{\theta} \end{aligned} \tag{26}$$

Let us assume without loss of generality that each of the individual components of the optimal coreset weighted client distribution $q^i(\hat{\boldsymbol{\theta}};\boldsymbol{w})$ can be denoted as some function $g(\boldsymbol{w})$. More, specifically,

$$\begin{aligned} \varrho^i(\boldsymbol{\theta_{i,j}}; \boldsymbol{w}) &= g_j(\boldsymbol{w}) \\ \nabla_w \varrho^i(\boldsymbol{\theta_{i,j}}; \boldsymbol{w}) &= g_j^{'}(\boldsymbol{w}) \end{aligned} \tag{27}$$

Thus we can reuse the above expression to simplify (26)

$$g_m^{'}(\boldsymbol{w}) = \frac{\int_\Theta \left(q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w}) \frac{1}{\mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)} \frac{\partial \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i)}{\partial w} \right) d\boldsymbol{\theta}}{\zeta \int_\Theta \left(\log q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w}) + 1 - \log(z^*(\boldsymbol{\theta})) + \log \mathcal{P}_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \right) d\boldsymbol{\theta}} \tag{28}$$

We now go back to utilizing the above derived expression in our main Eq. (19) to replace $\nabla_w q^i(\hat{\boldsymbol{\theta}}; \mathbf{w})$

$$\begin{aligned} \nabla_w q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w}) &= \nabla_w \prod_{k=1}^{T} \varrho^i(\boldsymbol{\theta_{i,k}}; \boldsymbol{w}) \\ &= \nabla_w \prod_{k=1}^{T} g_k(\boldsymbol{w}) \\ &= \prod_{k \neq m}^{T} g_k(\boldsymbol{w}) \nabla_w g_m(\boldsymbol{w}) + g_m(\boldsymbol{w}) \nabla_w \prod_{k \neq m}^{T} g_k(\boldsymbol{w}) \\ &= \prod_{k \neq m}^{T} g_k(\boldsymbol{w}) g_m^{'}(\boldsymbol{w}) + g_m(\boldsymbol{w}) \nabla_w \prod_{k \neq m}^{T} g_k(\boldsymbol{w}) \\ &= \frac{q^i(\hat{\boldsymbol{\theta}}; \boldsymbol{w})}{\varrho^i(\boldsymbol{\theta_{i,m}}; \boldsymbol{w})} g_m^{'}(\boldsymbol{w}) + g_m(\boldsymbol{w}) \nabla_w \prod_{k \neq m}^{T} \varrho^i(\boldsymbol{\theta_{i,k}}; \boldsymbol{w}) \end{aligned} \tag{29}$$

Thus, we now have a closed form solution to computing the gradient of the KL divergence $\mathbb{D}(q^i(\hat{\boldsymbol{\theta}}; \mathbf{w}) || q^i(\hat{\boldsymbol{\theta}}))$ w.r.t the coreset weight parameters $\boldsymbol{w}$.

Proposition 2. *The gradient of the second term in Equation 8 w.r.t $\boldsymbol{w}$ i.e.*

$$\nabla_w || P_\theta(\boldsymbol{\mathcal{D}^i}) - P_{\theta,w}(\boldsymbol{\mathcal{D}^i}) ||^2_{\hat{\pi},2}$$

is given by the following expression

$$-2\mathcal{P}_{\boldsymbol{\Phi}}^T \left(\mathcal{P} - \mathcal{P}_{\boldsymbol{\Phi}} \boldsymbol{w} \right)$$

where $\mathcal{P} = \sum_{j=1}^{n} \hat{g_j}$ and $\mathcal{P}_{\boldsymbol{\Phi}} = [\hat{g_1}, \hat{g_2}, \ldots, \hat{g_n}]$

Proof. First, we reformulate the given expression in terms

$$\begin{aligned} &\left\| P_\theta(\boldsymbol{\mathcal{D}}^i) - P_{\theta,w}(\boldsymbol{\mathcal{D}}^i) \right\|^2_{\hat{\pi},2} \\ &= \mathrm{E}_{\theta \sim \hat{\pi}} [(P_\theta(\boldsymbol{\mathcal{D}}^i) - P_{\theta,w}(\boldsymbol{\mathcal{D}}^i))^2] \end{aligned}$$

We define $g_j = \mathcal{P}_\theta(\boldsymbol{\mathcal{D}}_j^i) - \mathrm{E}_{\theta\sim\hat{\pi}} P_\theta(\boldsymbol{\mathcal{D}}_j^i)$

As a result the equivalent optimization problem becomes minimizing $\left\|\sum_{j=1}^n g_j - \sum_{j=1}^n w_j g_j\right\|_{\hat{\pi},2}^2$

Further, using Monte Carlo approximation, given S samples $\{\theta_j\}_{j=1}^S$, $\theta_j \sim \hat{\pi}$, the $L^2(\hat{\pi})$-norm can be approximated as follows

$$\left\|\sum_{j=1}^n \hat{g_j} - \sum_{j=1}^n w_j \hat{g_j}\right\|_2^2$$

where

$\hat{g_j} = \frac{1}{\sqrt{S}}\left[\mathcal{P}_{\theta_1}(\boldsymbol{\mathcal{D}}_j^i) - \mathcal{P}(\bar{\boldsymbol{\mathcal{D}}}_j^i), \mathcal{P}_{\theta_2}(\boldsymbol{\mathcal{D}}_j^i) - \mathcal{P}(\bar{\boldsymbol{\mathcal{D}}}_j^i), \ldots, \mathcal{P}_{\theta_S}(\boldsymbol{\mathcal{D}}_j^i) - \mathcal{P}(\bar{\boldsymbol{\mathcal{D}}}_j^i)\right]$ and $\mathcal{P}(\bar{\boldsymbol{\mathcal{D}}}_j^i) = \frac{1}{S}\sum_{k=1}^S \mathcal{P}_{\theta_k}(\boldsymbol{\mathcal{D}}_j^i)$

We can write the above problem in matrix notation as follows

$$f(\boldsymbol{w}) := \|\mathcal{P} - \mathcal{P}_{\boldsymbol{\Phi}}\boldsymbol{w}\|_2^2$$

where $\mathcal{P} = \sum_{j=1}^n \hat{g_j}$ and $\mathcal{P}_{\boldsymbol{\Phi}} = [\hat{g_1}, \hat{g_2}, \ldots, \hat{g_n}]$

Thus we have the gradient w.r.t $\boldsymbol{w}$ as follows:

$$\nabla_w f(\boldsymbol{w}) = -2\mathcal{P}_{\boldsymbol{\Phi}}^T\left(\mathcal{P} - \mathcal{P}_{\boldsymbol{\Phi}}\boldsymbol{w}\right) \tag{30}$$

□

10 Experiments

All the experiments have been done using the following configuration: Nvidia RTX A4000(16GB) and Apple M2 Pro 10 cores and 16GB memory.

10.1 Proposal for a modified objective in Equation 8

$$\{\boldsymbol{w_i}^*\} \triangleq \arg\min_{\boldsymbol{w}} \mathbb{D}_{KL}(\hat{q^i}(\boldsymbol{\theta}, \boldsymbol{w})||\hat{q^i}(\boldsymbol{\theta})) + \left\|P_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{D}^i}) - P_{\boldsymbol{\theta},\boldsymbol{w_i}}(\boldsymbol{\mathcal{D}^i})\right\|_{\hat{\pi},2}^2 \quad ||\boldsymbol{w_i}||_0 \leq k \tag{31}$$

We discuss here the utility of our proposed modified client side objective function via an ablation study where we want to gauge the inclusion of the first term in our objective function as just inlcuding the coreset loss.

Through experimental analysis, we find that just including the coreset loss optimization results in early saturation, possibly hinting towards getting stuck in local minima, but however inclusion the KL Divergence loss and forcing the coreset weighted local distribution of the client and the normal local distribution of the client to be similar leads to better stability in the training loss and better convergence.

10.2 Communication Complexity Analysis For Different Coreset Sizes

Here we showcase an analysis for different coreset sample size for different datasets and how it affects on the final accuracy and the total number of communication rounds in the Federated Learning setting. This showcases cost-effectiveness of our approach where by using only a small number of communication rounds our proposed approach is able to attain near-optimal performance as per the table below. In addition Fig: 5 substantiates the cost-effectiveness of our approach.

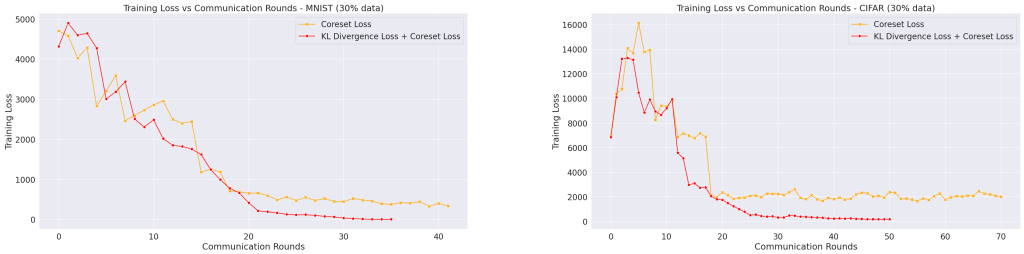

Figure 4: Ablation Study on using KL divergence between two local distribution w.r.t just using coreset weights

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.**

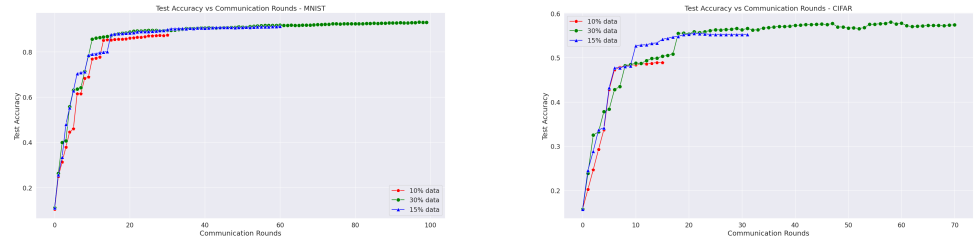

Figure 5: Communication Rounds across Different Sample Size - Convergence analysis

10.3 COMPUTING LIKELIHOOD OBJECTIVE USING AIHT

Here, we showcase how we utilised the Accelerated Iterative Hard Thresholding algorithm (A-IHT) for computing the likelihood.

10.4 MEDICAL DATASET EXPERIMENT DETAILS

Owing to the rise of Federated Learning based approaches in the medical setting due to privacy-preserving features, we chose to perform our experiments on 3 medical datasets in addition to our main experiments.

For our Federated Learning setup, we considered the setting where we have only 2 clients and one global server.

For each of the 3 datasets in the medical dataset setting, we consider each client has X-ray images of symptomatic type **A/** type **B** and Normal images . We perform a classification task at each client.

10.5 BASLINE COMPARISONS: DIVERSITY BASED SUBMODULAR OPTIMIZATION FUNCTIONS

For our second set of experiments, we chose different diversity based submodular optimization functions, specifically the following functions whose definition have been provided here

Definition 4. ***Log-determinant Function*** *is a diversity-based submodular function. It is non-monotone in nature. Let* $\mathbf{L}$ *denote a positive semidefinite kernel matrix and* $\mathbf{L_S}$ *denote the subset of rows and columns indexed by set* $\mathbf{S}$*. Log-determinant function* f *is specified as:*

$$f(\mathbf{S}) = logdet(\mathbf{L_S}) \tag{32}$$

The log-det function models diversity and is closely related to a determinantal point process.

Definition 5. ***Disparity Sum Function*** *characterizes diversity by considering the sum of distances between every pair of points in a subset* $\mathbf{S}$*. For any two points* $i, j \in \mathbf{S}$*, let* d_{ij} *denote the distance between them.*

$$f(\mathbf{S}) = \sum_{i,j \in \mathbf{S}} d_{ij} \tag{33}$$

The aim is to select a subset $\mathbf{S}$ *such that* $f(\mathbf{S})$ *is maximized. Disparity sum is not a submodular function.*

Definition 6. ***Disparity Min Function*** *characterizes diversity by considering the minimum distance between any two non-similar points in a subset* $\mathbf{S}$*.*

$$f(\mathbf{S}) = \min_{i,j \in \mathbf{S}, i \neq j} d_{ij} \tag{34}$$

The aim is to select a subset $\mathbf{S}$ *such that* $f(\mathbf{S})$ *is maximized. Disparity min is not a submodular function.*

For the above experiments we utilise the *Submodlib library* [3] for our implementation Kaushal et al. (2022).

10.6 Experiment Configuration

10.6.1 MNIST experiment configuration

For both CoreSet-PFedBayes and corresponding baseline PFedBayes, we use a fully connected DNN model with 3 layers [784,100,10] on MNIST dataset.

Learning rate hyperparameters: As per Zhang et al. (2022b)'s proposal i.e. PFedBayes the learning rates for personalized (client model) and global model (η_1, η_2) are set to 0.001 since these choices result in the best setting for PFedBayes. To compare against the stable best hyperparameters of PFedBayes, we also fix the same for our proposal CoreSet-PFedBayes.

Personalization Hyperparameter: The ζ parameter adjusts the degree of personalization in the case of clients. Again for a fair comparison against our baseline PFedBayes, we fix the ζ parameter for our proposal CoreSet-PFedBayes to the best setting given by the baseline. In Zhang et al. (2022b) the authors tune $\zeta \in \{0.5, 1, 5, 10, 20\}$ and find that $\zeta = 10$ results in the best setting. We, therefore, fix the personalization parameter $\zeta = 10$.

10.6.2 Medical Datasets experiment configuration

We discuss here the detailed configuration and models used for our further experiments.

Here we specifically consider the setting where we only have *2* clients and a single global server. Each of the 2 clients are assigned with data from only 2 classes along with a shared class for classification purpose.

For example, client 1 has class A and *Normal* (shared class) images while client 2 has class B and the remaining *Normal* images.

COVID-19 Radiography Database: Client 1 has COVID-19 x-ray images while client 2 has lung opacity x-ray images. Normal X-ray images are shared across both clients. Fig. 6 depicts the dataset distribution. For random subset selection, we randomly choose $\lambda = 0.1$ fraction of samples on the client side. For diversity-based subset selection, we first convert each of the 299×299 images into a [512×1] vector embeddings using a ResNet architecture. Diversity functions are then applied to

[3] Submodlib decile library

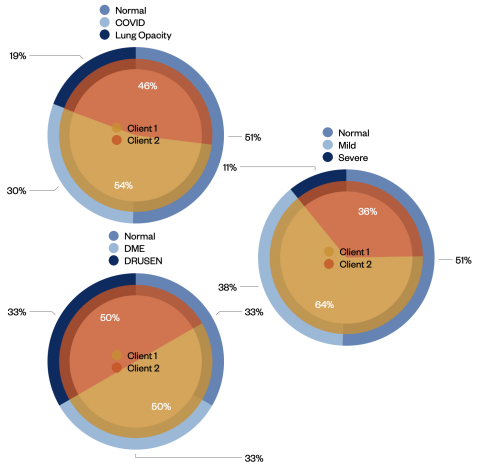

Figure 6: Data distribution for Medical Datasets

these embeddings to retrieve a final subset of diverse and representative embeddings. Eventually, we decode back to the original space using the chosen representative indices.

APTOS 2019 Blindness Detection: Unlike the COVID-19 radiography dataset, the APTOS dataset has 3 RGB channels and a higher resolution. We rescale the dimension of images to 299x299 for maintaining uniformity across all datasets. The same model configuration is followed as in the COVID-19 radiography dataset.

OCTMNIST: The OCTMnist dataset is a large dataset with single-channel images of a higher resolution. We have resized the images to 299×299 resolution for our experiments. The Normal class has above 50,000 train images itself, with the other two classes having close to 10,000 train images. Due to this class imbalance, we have randomly selected 8,000 images from each class for our experiments. Post which we again use a ResNet architecture to reduce the feature dimensions, which we then feed into the CORESET-PFEDBAYES pipeline.

Baseline : Independent Learning In this scenario, each of the 2 clients solve the classification problem independently without any involvement of a server as opposed to federated learning. Thus there is no sharing of model weights to a common server as compared to the federated setting.

Baseline : Independent Learning on other client's test data In this scenario, similar to the independent learning setup, we report the metrics for a particular client not only on its own test data but also on the other client's test data by training on the individual client's own training data.

For all the experiments for the medical dataset analysis across all the baselines, we report the class-wise accuracy in Table 2.

Definition 7. ***Submodular Functions*** *are set functions which exhibit diminishing returns. Let* $\mathbf{V}$ *denote the ground-set of* n *data points* $\{x_1, x_2, \dots x_n\}$ *where* $x_i \in \mathbb{R}^d$. *More formally,* $\mathbf{V} = \{x_i\}_{i=1}^n$. *Let* $\mathbf{A} \subseteq \mathbf{B}$ *where* $\mathbf{A}, \mathbf{B} \subset \mathbf{V}$ *and* $v \in \mathbf{V}$. *A submodular function* $f : 2^V \mapsto \mathbb{R}$ *satisfies the diminishing returns property as follows:*

$$f(\mathbf{A} \cup v) - f(\mathbf{A}) \geq f(\mathbf{B} \cup v) - f(\mathbf{B}) \tag{35}$$

10.7 ALGORITHM FOR ACCELERATED IHT

We first present the accelerated IHT algorithm as proposed in Zhang et al. (2021) in Algorithm 10.7.

Algorithm 2 Accelerated IHT (A-IHT) for Bayesian Coreset Optimization

Input Objective $f(w) = \|y - \Phi w\|_2^2$; sparsity k

1: $t = 0, z_0 = 0, w_0 = 0$
2: repeat
3: $\quad \mathcal{Z} = \text{supp}\left(z_t\right)$
4: $\quad \mathcal{S} = \text{supp}\left(\Pi_{\mathcal{N}_k \setminus \mathcal{Z}}\left(\nabla f\left(z_t\right)\right)\right) \cup \mathcal{Z}$ where $|\mathcal{S}| \leq 3k$
5: $\quad \widetilde{\nabla}_t = \nabla f\left(z_t\right)|_{\mathcal{S}}$
6: $\quad \mu_t = \arg\min_\mu f\left(z_t - \mu \widetilde{\nabla}_t\right) = \frac{\left\|\tilde{\nabla}_t\right\|_2^2}{2\left\|\Phi \tilde{\nabla}_t\right\|_2^2}$
7: $\quad w_{t+1} = \Pi_{\mathcal{C}_k \cap \mathbb{R}_+^n}\left(z_t - \mu_t \nabla f\left(z_t\right)\right)$
8: $\quad \tau_{t+1} = \arg\min_\tau f\left(w_{t+1} + \tau\left(w_{t+1} - w_t\right)\right)$
$= \frac{\langle y - \Phi w_{t+1}, \Phi(w_{t+1} - w_t)\rangle}{2\|\Phi(w_{t+1} - w_t)\|_2^2}$
9: $\quad z_{t+1} = w_{t+1} + \tau_{t+1}\left(w_{t+1} - w_t\right)$
10: $\quad t = t + 1$
11: until Stop criteria met
12: return w_t

The algorithm **Accelerated IHT** above is proposed by Zhang et al. (2021). We share a high level view of the algorithm include some of the important features.

Step Size Selection The authors propose that given the quadratic objective of the coreset optimization, they perform exact line search to obtain the best step size per iteration. $\frac{\left\|\tilde{\nabla}_t\right\|_2^2}{2\left\|\Phi \tilde{\nabla}_t\right\|_2^2}$

Momentum The authors propose adaptive momentum acceleration as is evident from line 8 of the pseudocode. At the end during the next update, Nesterov Accelerated Gradient is applied as shown in line 9.

11 CODE

We share our code on GitHub at Link