
PRIOR: Personalized Prior for Reactivating the Information Overlooked in Federated Learning.

Mingjia Shi^{1*} Yuhao Zhou¹ Kai Wang² Huaizheng Zhang
Shudong Huang¹ Qing Ye^{1†} Jiangcheng Lv^{1†}
¹Sichuan University ²National University of Singapore

Abstract

Classical federated learning (FL) enables training machine learning models without sharing data for privacy preservation, but heterogeneous data characteristic degrades the performance of the localized model. Personalized FL (PFL) addresses this by synthesizing personalized models from a global model via training on local data. Such a global model may overlook the specific information that the clients have been sampled. In this paper, we propose a novel scheme to inject personalized prior knowledge into the global model in each client, which attempts to mitigate the introduced incomplete information problem in PFL. At the heart of our proposed approach is a framework, the *PFL with Bregman Divergence* (pFedBreD), decoupling the personalized prior from the local objective function regularized by Bregman divergence for greater adaptability in personalized scenarios. We also relax the mirror descent (RMD) to extract the prior explicitly to provide optional strategies. Additionally, our pFedBreD is backed up by a convergence analysis. Sufficient experiments demonstrate that our method reaches the *state-of-the-art* performances on 5 datasets and outperforms other methods by up to 3.5% across 8 benchmarks. Extensive analyses verify the robustness and necessity of proposed designs. https://github.com/BDeMo/pFedBreD_public

1 Introduction

Federated learning (FL) [53] has achieved significant success in many fields [72, 43, 71, 76, 69, 59, 32, 5, 34, 49], which include recommendation systems utilized by e-commerce platforms [71], prophylactic maintenance for industrial machinery [76], disease prognosis employed in healthcare [69]. Data heterogeneity is a fundamental characteristic of FL, leading to challenges such as inconsistent training and testing data (data drift) [36]. An efficient solution to these challenges is to fine-tune the global model locally for adaptation on local data [3, 74, 35]. This solution is straightforward and pioneering, but presents a fundamental limitation when dealing with highly heterogeneous data. For examples, heterogeneous data drift may introduce substantial noise [29] and the resulted model may not generalize well to new sample [20, 12]. Thus, heterogeneous data in FL is still challenging [66].

Recently, personalized FL (PFL) is proposed to mitigate the aforementioned negative impact of heterogeneous data [66]. To improve the straightforward solution mentioned above, Per-FedAvg [20] is introduced to train a global model that is easier to fine-tune. Another paper on the similar topic, FedProx [45], aims to resolve the issue of personalized models drifting too far from the global model during training with a dynamic regularizer in the objective during local training. This issue could occur especially in post-training fine-tuning methods without regularization (*e.g.*, Per-FedAvg [20]). Moreover, pFedMe [65], another regularization method modeling local problems using Moreau

*3101ihs@gmail.com

†Equal Corresponding Authors {yeqing, lvjiancheng}@scu.edu.cn

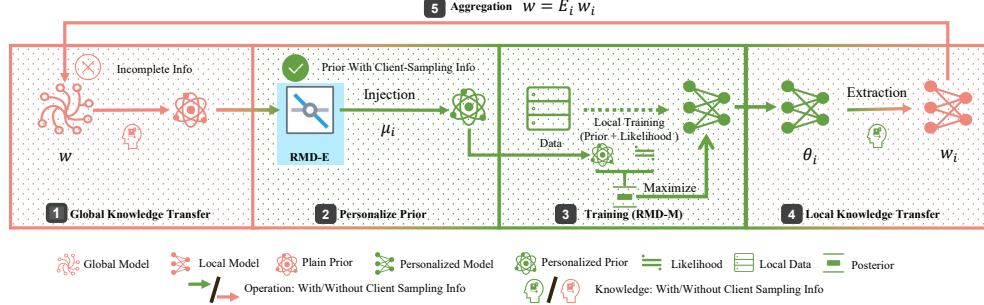


Figure 1: pFedBreD framework: Global-MLE and Local-MAP. The personalized prior knowledge is injected into the global model of the global problem (MLE) in the 2nd step for local training. The local knowledge is extracted from the local problem (MAP) in the 4th step for aggregation.

envelopes, replaces FedProx’s personalized model aggregation method with an interpretable approach for aggregating local models [45]. It also accommodates first-order Per-FedAvg [20].

Although the existing PFL methods have achieved promising results, the prior knowledge from single global model for local training [66] hinders the development of PFL. Specifically, we analyze the shortcomings of current PFL methods as follows: 1) utilizing the same global model for direct local training could potentially disregard the client’s sampling information. As shown in Figure 1, a single global model provides global knowledge directly for local training, which overlooks the client-sampling information when the global knowledge is transferred to specific clients. 2) Explicitly extracting prior knowledge can be a challenging task. Most of the insightful works [17, 52] propose assumptions for recovering this incomplete information, but these assumptions are implicit, which limits the way to use the information to develop personalized strategies.

To address the former issue above, we propose framework pFedBreD to inject personalized prior knowledge into the one provided by a global model. As shown in Figure 1, it is injected in the 2nd step and the local knowledge is transferred into global model via local models instead of directly aggregating personalized models [45]. To address the latter, we introduce *relaxed mirror descent* RMD to explicitly extract the prior for exploring personalized strategies.

Our method is backed up with direct theoretical support from Bayesian modeling in Section 4 and a convergence analysis in Section 5, which provides a linear bound $\mathcal{O}(1/TN)$ with aggregation noise and a quadratic speedup $\mathcal{O}(1/(TNR)^2)$ without.³ Meanwhile, the existence and validity of the injection and extraction aforementioned information is verified in Section 6.2. The remarkable performance of the implements of the proposed method is tested on 5 datasets and 8 benchmarks. Consistently, our method reach the *state-of-the-art*. Especially, the improvement of accuracy on task DNN-FEMNIST [11] is up to 3.5%. Extensive ablation study demonstrate that parts of the hybrid strategy **mh** are complementary to each other. Our contributions can be summarized as follows:

- The problem of overlooking client-sampling information at prior knowledge being transferred is introduced in this paper, and we first investigate the possibility of explicitly expressing the prior knowledge of the information and design personalized strategies on it.
- To express the personalized prior, we model PFL into a Bayesian optimization problem, *Global-MLE and Local-MAP*. A novel framework, pFedBreD, is proposed for computing the modeled problem, and RMD is introduced to explicitly extract prior information.
- Sufficient experiments demonstrate our method surpasses most baselines on public benchmarks, thereby showcasing its robustness to data heterogeneity, particularly in cases involving small aggregation ratios and non-convex local objective settings.

³ TN and TNR : total global / local epoch in FL system. See Appendix A.1.

2 Related Works

Regularization Researchers have developed a variety of approaches based on regularization to handle the PFL challenge in recent years (*e.g.*, FedU [18], pFedMe [65], FedAMP [31], HeurFedAMP [31]). All of these approaches’ personalized objective functions can be expressed as $J(\theta) + R(\theta; \mu)$ where $J(\theta)$ is the loss function of the local problem and $R(\theta; \mu)$ is the regularization term used to restrict the deviation between θ and μ (*e.g.*, $R(\theta; \mu) = \frac{1}{2}\|\theta - \mu\|^2$ in pFedMe).

Meta Learning One of the most representative meta-learning based single-model PFL approach is the well-known Per-FedAvg [20], aiming to find an initialization that is easy to fine-tune. That is, the global model in the FL setting is regarded as a meta model in MAML [21, 19], where the objective function of the local problem is $J(\theta - \eta \nabla J(\theta))$. Researchers also show the connections between FedAvg [53] and Reptile [60], another meta learning framework. [35] shows how to improve personalization of FL via Reptile. Proximal updating is also used in meta-learning based algorithms such as [79]. One of our strategies **meg** is the one motivated by MAML.

Expectation Maximization Two EM-based [15] methods are proposed, *e.g.*, FedSparse in [50] and FedEM in [17]. Both of them focus on communication compression. The latter provides a variance reduce version and assumes complete information (or data) of the global model obeys distribution in X-family. Another FedEM [52] combines Bayesian modeling, Federated Multi-task learning (FTML) and EM. Our framework pFedBreD is a expectation maximization and maximum a posteriori estimate (EM-MAP) [17] algorithm with personalized prior specified.

Bayesian FL In recent years, studies of PFL with Bayesian learning have been proposed. In related approaches, FLOA [48] and pFedGP [1] are proposed with KL divergence regularization in the loss function, which is comparable to applying specific assumption of X-family prior in pFedBreD see Appendix B.2 for details. Our implementation doesn’t use a Bayesian neural network (BNN) model as an inferential model as others do (*e.g.*, pFedGP uses a Gaussian process tree and pFedBayes [77] uses BNN). Instead, to eliminate weight sampling cost in Bayesian methods, prior knowledge is introduced through regularization term.

3 Preliminary

Overlooked Information in Prior Knowledge From a Bayesian and info. perspective, the global knowledge transferred in conventional method with single global model has no mutual information (MI) with client sampling i , *i.e.*, formally, $w = \mathbf{E}_i w_i = \mathbf{E}_i w_i | i \Rightarrow \text{MI } I(w; i) = 0$, in particular when applying reg. $R(w^{(t)}; \dots)$ or local init. $w_{i,0}^{(t)} \leftarrow w$ where w_i is the local model on the i^{th} client. This makes the specific model on each client have to re-obtain this information from scratch solely from the data during training, especially impacted on hard-to-learn representations and datasets.

Bregman-Moreau Envelope Bregman divergence [10] is employed as a general regular term in our local objective that exactly satisfies the computational requirements and prior assumption, and is formally defined in Eq. (1).

$$\mathcal{D}_g(x, y) := g(x) - g(y) - \langle \nabla g(y), x - y \rangle \quad (1)$$

where g is a convex function. For convenience, g is assumed to be strictly convex, proper and differentiable such that Bregman divergence is well-defined. To utilize the computational properties of Bregman Divergence in optimization problems, we introduce the following definition in Eq. (2) [7, 8]: Bregman proximal mapping, Bregman-Moreau envelope, and the relationship between them.

$$\begin{aligned} \mathcal{D}\text{prox}_{g, \lambda^{-1}} f(x) &:= \arg \min_{\theta} \{f(\theta) + \lambda \mathcal{D}_g(\theta, x)\}, \\ \mathcal{D}\text{env}_{g, \lambda^{-1}} f(x) &:= \min_{\theta} \{f(\theta) + \lambda \mathcal{D}_g(\theta, x)\}, \\ \nabla \mathcal{D}\text{env}_{g, \lambda^{-1}} f(x) &= \lambda \nabla^2 g(x) [x - \mathcal{D}\text{prox}_{g, \lambda^{-1}} f(x)], \end{aligned} \quad (2)$$

where $\lambda > 0$ denotes the regular intensity in general and the variance of the prior in our modeling.

Exponential Family The regular exponential family (X-family) is a relatively large family that facilitates calculations. Therefore, to yield the prior, we employ the X-family [6] defined in Eq. (3).

$$\mathbf{P}_{ef}(\mathcal{V}; s, g) = h(\mathcal{V}) \exp\{\langle \mathcal{V}, s \rangle - g(s)\} = h(\mathcal{V}) \exp\{-\mathcal{D}_{g^*}(\mathcal{V}, \mu) + g^*(\mathcal{V})\}, \quad (3)$$

where g is assumed to be convex, $\mathcal{D}_g(\cdot, \cdot)$ is the Bregman divergence, and g^* is the Fenchel Conjugate of g . In Eq. (3), s , $h(\mathcal{V})$ and $g(s)$ are respectively the natural parameter, potential measure and logarithmic normalization factor, where we have the mean parameter $\mu = \nabla g(s)$. Additionally, to highlight the variance, the scaled exponential family (SX-family) is introduced in Eq. (4)

$$\mathbf{P}_{sef}(\mathcal{V}; \lambda, s, g) = h_\lambda(\mathcal{V}) \exp\{\lambda[\langle \mathcal{V}, s \rangle - g(s)]\} = h_\lambda(\mathcal{V}) \exp\{-\lambda \mathcal{D}_{g^*}(\mathcal{V}, \mu) + \lambda g^*(\mathcal{V})\}, \quad (4)$$

where $\log h_\lambda(\mathcal{V})$ is the scaled potential measure, and the scale parameter λ is employed to highlight the variance. Moreover, \mathcal{V} is assumed to be the minimal sufficient statistic of the complete information for local inference, details of which can be found in Section 4.

4 Methodology

In this section⁴, we introduce missing client-sampling information based on classic FL, use EM to reduce the computational cost of the information-introduced FL problem, and propose RMD, a class of prior selection strategies, based on the E-step in EM. The general FL classification problem with KL divergence could be formulated as Eq. (5) [53, 66].

$$\arg \min_w \mathbf{E}_i \mathbf{E}_{d_i} \text{KL}(\mathbf{P}(y_i|x_i) || \hat{\mathbf{P}}(y_i|x_i, w)) = \arg \max_w \mathbf{E}_i \mathbf{E}_{d_i} \mathbf{E}_{y_i|x_i} \log \hat{\mathbf{P}}(y_i|x_i, w), (x_i, y_i) \in d_i, \quad (5)$$

where we rewrite the discriminant model as an maximum likelihood estimation (MLE) problem [41] of $y_i|x_i$ in the right hand side (R.H.S.) of Eq. (5). (x_i, y_i) represent the pairs of input and label respectively in dataset d_i on the i^{th} client, and $\hat{\mathbf{P}}(y_i|x_i, w)$ is the inferential model parameterized by w . Each local data distribution is presuppose to be unique, so using the global model with local data for inference and training could overlook the fact that the client has been sampled before transmitting the global model, and the prior knowledge transmitted directly via the global model as the local training prior knowledge (e.g. via initial points, penalty points in dynamic regular terms, etc.) has no mutual information with the client sampling, i.e., the global model $w = \mathbf{E}_i w_i = \mathbf{E}_i w_i|i$. Thus, to reduce the potential impact of the overlooked information, the complete information Θ_i on the i^{th} client is introduced which turns Eq. (5) into Eq. (6).

$$\arg \max_w \mathbf{E} \log \int_{\Theta_i} \hat{\mathbf{P}}(y_i|x_i, \Theta_i, w) \mathbf{P}(\Theta_i|x_i, w) d\Theta_i, \quad (6)$$

where $\mathbf{E} = \mathbf{E}_i \mathbf{E}_{d_i} \mathbf{E}_{y_i|x_i}$ and the direct calculation of this is computationally expensive [14].

Framework: Leveraging Expectation Maximization for Prior Parameter Extraction The integral term in Eq. (6) makes direct computation impossible [14], so we employ EM to approximate the likelihood with unobserved variables [15] as shown in Eq. (7), where $\mathbf{Q}(\Theta_i)$ is any probability measure.

$$\sum_i \log \hat{\mathbf{P}}(y_i|x_i, w) \geq \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} [\log \hat{\mathbf{P}}(y_i|x_i, \Theta_i, w) + \mathbf{E}_{y_i|x_i, w} \log \mathbf{P}(\Theta_i|d_i, w)]. \quad (7)$$

Assuming that prior $\Theta_i|d_i, w \sim \hat{\mathbf{P}}_{sef}(\Theta_i; \lambda, s_i(w; d_i), g)$ and the local loss function on the i^{th} client $f_i(\Theta_i, w)$ is $\mathbf{E}_{d_i} [-\log \mathbf{P}(y_i|x_i, \Theta_i, w)]$, we have the left hand side (L.H.S.) of the Eq. (8) from (7). Here is an assumption for simplification that θ_i contains all the information for local inference, i.e. $\theta_i = \Theta_i$ and $\mathbf{P}(y_i|x_i, \Theta_i, w) = \mathbf{P}(y_i|x_i, \Theta_i)$. It happens when θ_i is all the parameters of the personalized model and we only use the personalized model for inference. Thus, $f_i(\theta_i) = \mathbf{E}_{d_i} [-\log \hat{\mathbf{P}}(y_i|x_i, \theta_i)]$. Thus, we can optimize an upper bound as a bi-level optimization problem as shown in the R.H.S. of the Eq. (8) to solve Eq. (5) approximately, where mean parameter $\mu_i = \nabla g \circ s_i$ ⁵ [9]. And, we can derivate our framework as shown in Section 5.

$$-\max_{w, \{\theta_i\}} \mathbf{E}_i \{-f_i(\theta_i) - \lambda \mathcal{D}_{g^*}(\theta_i, \mu_i(w))\} \leq \min_w \mathbf{E}_i \min_{\{\theta_i\}} \{f_i(\theta_i) + \lambda \mathcal{D}_{g^*}(\theta_i, \mu_i(w))\}. \quad (8)$$

⁴More details of equations are in Appendix B.

⁵The following d_i is omitted with the same footnote i in μ_i for simplification ($\mu_i(\cdot) \leftarrow \mu_i(\cdot; d_i)$).

Strategies: Relaxing Mirror Descent for Prior Selection To extract the prior strategies and implement μ_i E-step of EM in close-form, we propose a method called relaxed mirror descent (RMD), where the mirror descent (MD) is EM in X-family [40]. MD can be generally written as Eq. (9) from the old \hat{w} to the new one \hat{w}^+ in each iteration [54, 40].

$$\hat{w}^+ \leftarrow \arg \min_{\hat{w}} \{f(\hat{w}) + \langle \nabla f(\hat{w}), \hat{\theta} - \hat{w} \rangle + \hat{\lambda} \mathcal{D}_{\hat{g}}(\hat{\theta}, \hat{w})\}. \quad (9)$$

According to the Lagrangian dual, we rewrite the problem into a more general variant shown in Eq. (10) with relaxed restrictions and superfluous parameter.

$$\arg \min_{\hat{\theta}, \hat{\mu}} \{\Psi(\hat{\theta}, \hat{w}) + \langle \nabla \Phi(\hat{w}), \hat{\mu} - \hat{w} \rangle + \lambda \mathcal{D}_{g^*}(\hat{\theta}, \hat{\mu}) + (2\eta)^{-1} \|\hat{\mu} - \hat{w}\|^2\}. \quad (10)$$

We can transform Eq. (10) back into Eq. (9) by setting $\Phi(\hat{w})$ to satisfy $\nabla \Phi(\hat{w}) = \nabla f(\hat{w})$, and defining $\Psi(\hat{\theta}, \hat{w})$ as a function with $f(\hat{w})$ and a penalty term to make $\hat{\theta}$ and \hat{w} close as possible (e.g., $\hat{\lambda} \mathcal{D}_{\hat{g}}(\hat{\theta}, \hat{w})$). This provides us a way to extract μ_Φ the function to generate mean parameter of the prior, as shown in Eq. (11), which is minimizing an upper bound of the problem in Eq. (10).

$$\begin{aligned} \mathcal{D}\text{env}_{g^*, \lambda^{-1}} \Psi(\cdot, w)(\mu_\Phi(w)) &= \min_{\theta} \{\Psi(\theta, w) + \lambda \mathcal{D}_{g^*}(\theta, \mu_\Phi(w))\} \\ \mu_\Phi(w) &= \arg \min_{\mu} \{\langle \nabla \Phi(w), \mu - w \rangle + (2\eta)^{-1} \|\mu - w\|^2\}. \end{aligned} \quad (11)$$

By optimality condition, we have $\mu_\Phi(w) = w - \eta \nabla \Phi(w)$, which can be specified by Φ . The remaining part is a Bregman-Moreau envelope. Thus, we can optimize the upper bound with an EM-MAP method, alternately computing $\mu_\Phi(w)$ and $\mathcal{D}\text{prox}_{g^*, \lambda^{-1}} \Psi(\cdot, w)(\mu_\Phi(w))$.

5 Framework Design

Problem Formulation that Highlights Personalized Prior Inspired by the aforementioned motivation, the personalized models θ_i and mean parameters are respectively the solution of $\mathcal{D}\text{env}_{g^*, \lambda^{-1}} f_i(\mu_i(w))$ and $\mu_i(w)$ on the i^{th} client, where w is the global model. We assume that personalized model contains all the local information required for inference on the i^{th} client, and satisfies $\theta_i | x_i, w \sim \mathbf{P}_{\text{sef}}(\theta_i; \lambda, s_i(w), g)$. The global problem can be written as Eq. (12).

$$\min_w \mathbf{E}_i \{F_i(w) := \mathcal{D}\text{env}_{g^*, \lambda^{-1}} f_i(\mu_i(w))\}. \quad (12)$$

The given g is strictly convex, $\lambda > 0$, f_i is the local loss function, $s_i(w)$ is the natural parameter and $\mu_i(w) = \mathbf{E}_{\theta_i | x_i, w} \theta_i = \nabla g(s_i(w))$ is the mean (or expectation) parameter in Eq. (12).

Framework: pFedBreD To solve the optimization problem in Eq. (12), we use gradient-based methods to solve the global problem using the gradient of F_i :

$$\nabla F_i(w) = \lambda \mathbf{D} \mu_i(w) \nabla^2 g^*(\mu_i(w)) [\mu_i(w) - \mathcal{D}\text{prox}_{g^*, \lambda^{-1}} f_i(\mu_i(w))], \quad (13)$$

where \mathbf{D} is the gradient operator of the vector value function, and ∇^2 is the Hessian operator.⁶ The framework is shown as Algorithm 1, where \mathcal{I} is the client selecting strategy for global model aggregation; w_{init} and θ_{init} are the initialization strategies on the i^{th} client; α_m is the main problem step-size; T, R, N are respectively the total number of iterations, local iterations, and clients. β is used in the same trick as [37, 65]. The strategies to derive the initialization points of w_i and θ_i at each local epoch are $w_{i,0}^{(t)} \leftarrow w^{(t-1)}$ and $\theta_{i,0}^{(t)} \leftarrow \theta_{i,R}^{(t-1)}$.

Implementation: Maximum Entropy and Meta-Step Practically, two main parts of the pFedBreD are needed to be implemented:

- g , the function used to derive the logarithmic normalization factor, determines the type of prior to be used;
- $\{s_i\}$ or $\{\mu_i\}$, the functions used to derive the natural parameter and mean parameter for the personalized local prior, determine which particular prior is used.

⁶The details of first-order methods is in Appendix A.7.

Algorithm 1 Algorithm for pFedBreD

Input: $\mathcal{I}, \{d_i\}, i = 1 \dots N$ **Parameter:** $\alpha_m, g, \lambda, T, R, \{w_{init}, \theta_i, \mu_i\}, i = 1 \dots N$ **Output:** $w^{(T)}, \{\theta_i^{(T)}\}, i = 1 \dots N$

```
1: Initialize  $w^{(0)}, \{\theta_i^{(0)}\}, \{C_{i,R}^0\}$ ;
2: for  $t=1 \dots T$  do
3:   Server sends  $w^{(t-1)}$  to clients ;
4:   for  $i=1 \dots N$  in parallel on each clients do
5:     Initialize  $w_{i,0}^{(t)}$  and  $\theta_{i,0}^{(t)}$  with  $w_{init}$  and  $\theta_{init}$ ;
6:     for  $r=1 \dots R$  do
7:       Generate  $\mu_{i,r}^{(t)} \leftarrow \mu_i(w_{i,r-1}^{(t)}, \dots)$  ;
8:        $\theta_{i,r}^{(t)} \leftarrow \mathcal{D}\text{prox}_{g^*, \lambda^{-1}} f_i(\mu_{i,r}^{(t)})$  ;
9:        $w_{i,r}^{(t)} \leftarrow w_{i,r-1}^{(t)} - \alpha_m \nabla F_i(w_{i,r-1}^{(t)})$  ;
10:    end for
11:  end for
12:  Server collects  $\{w_{i,R}^{(t)}\}$  and calculate  $w^{(t)} \leftarrow (1 - \beta)w^{(t-1)} + \beta \mathbf{E}_{\mathcal{I}} w_{i,R}^{(t)}$  ;
13: end for
14: return  $w^T, \{\theta_i^T\}$ .
```

We propose first-order implementations based on maximum entropy rule [22, 33]. In the SX-family, the Gaussian distribution has the maximum entropy among continuous distributions when g, μ_i (the first-order moment), and λ (the parameter determining the second moment) are given. Thus, we employ the scaled norm square $g = g^* = \frac{1}{2} \|\cdot\|^2$ to turn the prior into a spherical Gaussian, in order to maximize the entropy of the prior on a particular client. With this assumed prior, we have $\nabla g = \nabla g^* = I$, which means $\mu_i = s_i$. We can choose a different Φ_i as shown in Eq. (14)⁷ to generate selection strategies according to Section 4, via $\mu_i(w) = w - \eta \nabla \Phi_i(w)$ (meta-step).

$$\Phi_i = \begin{cases} f_i & \text{lg} \\ F_i & \text{meg} \\ f_i + F_i & \text{mh} \end{cases} \quad \mu_{i,r}^{(t)} \leftarrow \begin{cases} w_{i,r-1}^{(t)} - \eta_\alpha \nabla f_i(w_{i,r-1}^{(t)}), \\ w_{i,r-1}^{(t)} - \eta(w_{i,R}^{(t-1)} - \theta_{i,r-1}^{(t)}), \\ w_{i,r-1}^{(t)} - \eta_\alpha \nabla f_i(w_{i,r-1}^{(t)}) - \eta(w_{i,R}^{(t-1)} - \theta_{i,r-1}^{(t)}), \end{cases} \quad (14)$$

where η_α and η are the meta-step-size parameters. Practical parameter selection strategies with meta-step are shown as $\mu_{i,r}^{(t)}$ in Eq. (14). The three of μ_i , i.e. **lg**, **meg** and **mh**, represent **loss gradient**, **memorized envelope gradient** and **memorized hybrid** respectively.

Convergence Analysis we analyze the convergence of pFedBreD with RMD on a uniform client sampling $\mathbf{E}_i = \frac{1}{N} \sum_{i=1}^N$ setting for simplification. Other sampling methods can be obtained with client sampling expectation $\mathbf{E}_i[F_i] = F$, by changing sampling weights. The assumptions, proof sketch and detailed notations are in Appendix A.1 and Appendix D.

Theorem 1 (pFedBreD's global bound). *Under settings in Section 5 and Appendix D, at global epoch $T \geq \frac{2}{\hat{\mu}_F \hat{\alpha}}$, by properly choose $\tilde{\alpha}_m = \alpha_m \beta R$, $\exists \tilde{\alpha}_m \leq \min\{\frac{\beta}{\sqrt{2}\hat{c}}, \frac{2}{\hat{\mu}_F}, \hat{\alpha}_m\}$, where $A = [\frac{\hat{L}_{g^*}}{\hat{\mu}_F}(\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 (\frac{\hat{\gamma}_f^2}{|\hat{d}_i|} + \hat{c}^2)$, $B = [\hat{L}_\mathcal{E} \hat{\gamma}_\Phi (1 + \sigma_\Phi)(\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2$, $C = \frac{\sigma_\Phi^2 \hat{L}_\mathcal{E}^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2}{\hat{\mu}_F^3}$, $\xi^{(t)} = (1 - \frac{\tilde{\alpha} \hat{\mu}_F}{2})^{-t-1}$, $\bar{w}^{(T)} := \frac{\sum_{t=0}^{T-1} \xi^{(t)} w^{(t)}}{\sum_{t=0}^{T-1} \xi^{(t)}}$ and $\hat{\alpha}_m := \frac{\hat{\mu}_F \beta R}{e(1+\sigma_\Phi) \hat{L}_\mathcal{E} (\hat{u}_m + \eta \hat{\gamma}_\Phi) 2^{R+6\frac{1}{2}} (\frac{1}{R}+2) + 15(\hat{\mu}_F \beta R) \hat{L}_F}$, such that:*

$$\begin{aligned} \mathcal{O}[\mathcal{D}_F(\bar{w}^{(T)}, w^*)] &= \mathcal{O}(\hat{\mu}_F e^{-\tilde{\alpha}_m \hat{\mu}_F T/2} \Delta^{(0)}) + \mathcal{O}\left(\frac{A\lambda^2 + B}{\hat{\mu}_F}\right) \\ &\quad + \mathcal{O}\left(\frac{(N/S-1)\sigma_{F,*}^2}{NT\hat{\mu}_F}\right) + \mathcal{O}\left(\frac{2^R C}{T^2 \beta^2 R^2} [R\sigma_{F,*}^2 + A\lambda^2 + B]\right). \end{aligned}$$

⁷A variant of **mh** is in Appendix C.2.

Theorem 2 (pFedBreD_{ns}'s first-order personalization bound). *Under the same conditions as in Theorem 1, with prior assumption of a spherical Gaussian and first-order approximation, the bound for the gap between the personalized approximate model and global model in the Euclidean space is:*

$$\mathbf{E} \|\tilde{\theta}_i(\bar{w}^T) - w^*\|^2 \leq \mathcal{O}(\dot{\delta}_p) + \mathcal{O}[\dot{c}_p \mathcal{D}_F(\bar{w}^{(T)}, w^*)]$$

where $\dot{\delta}_p = \frac{2}{\hat{\mu}_{F,i}^2} (\frac{\hat{\gamma}_f^2}{|d_i|} + \hat{\epsilon}^2) + \frac{2}{\lambda^2} \epsilon_1^2 + \frac{4}{\lambda^2} \sigma_{F,*}^2 + \frac{1}{2} \eta^2 \mathcal{G}_{\Phi}^2$, and $\dot{c}_p = (\frac{32}{\lambda^2} \hat{L}_F + \frac{8}{\hat{\mu}_F})$.

Remark 1. Theorem 1 shows the main factors that affect the convergence of a global model are as follows: random mini-batch size, client drift error, aggregation error, heterogeneous data, dual space selection, local approximation error, and selection strategy for exponential family prior mean and variance. These can be divided into four categories based on their computational complexity. The first and second term shows that the proper fixed $\tilde{\alpha}_m$ can linearly reduce the influence of initial error $\Delta^{(0)}$ and the global model converges to a ball near the optimal point. The radius of this ball is determined by the personalized strategy and local errors (including local data randomness and envelope approximation errors). The third term implies that a linear convergence rate $\mathcal{O}(1/(NT))$ can be obtained w.r.t. the total global epoch NT in the presence of aggregation noise. Without client sampling $N = S$, according to the fourth term, the quadratic rate $\mathcal{O}(1/(TNR)^2)$ can be obtained with $\beta = \mathcal{O}(N)$ or $\beta = \mathcal{O}(N\sqrt{R})$ (Note that the number of local epoch R cannot be too large due to client drift, according to 2^R). Theorem 2 shows that, with spherical Gaussian prior assumption and first-order methods, the radius of the neighborhood range for the minimum that includes the personalized model on i^{th} client, $\mathcal{O}(C_{\Phi,F,f,d} + \frac{1}{\lambda^2} (\epsilon_1^2 + \sigma_{F,*}^2 + \frac{B\hat{L}_F}{\hat{\mu}_F}) + \lambda^2 \frac{A}{\hat{\mu}_F})$, can be trade-off by λ , and is affected by the prior selection strategies and first-order approximate error besides the elements in Theorem 1. (Note that the Euclidean space is self-dual.)

6 Experiments

6.1 General Settings

Tricks, Datasets and Models: our experiments include several tasks: CNN [28] on CIFAR-10 [18, 39], LSTM [27] on Sent140 [11] and MCLR/DNN on FEMNIST [11]/FMNIST [65, 67]/MNIST [65, 42]. The details of tricks (FT, AM), data heterogeneity and models are in Appendix C.

Baselines: we choose following algorithms as our baselines: FedAvg [53], Per-FedAvg [20], pFedMe [65], FedAMP [31], pFedBayes [77] and FedEM [52]. These baselines are respectively classical FL, MAML-based meta learning, regularization based, FTML methods, variational inference PFL and FMFL with EM.

Global Test and Local Test: the global and personalized model, represented by \mathbf{G} and \mathbf{P} , are evaluated with global and local tests respectively. Global test means all the test data is used in the test. Local test means only the local data is used for the local test and the weight of the sum in local test is the ratio of the number of data. The results of average accuracy per client are shown in Table 1. Each experiment is repeated 5 times. More details are in Appendix C. For readability, we only give the error bar in the main Table 1 and Table 2, and keep one decimal except for the main Table 1.

Hyperparameter Settings The step-size of the main problem, α_m , and the personalized step-size, α , for all methods are 0.01. β is 1, and the number of local epochs, R , is 20 for all datasets. λ is chosen from 15.0 to 60.0. The batch sizes of Sent140 and the other datasets are 400 and 20, respectively, and the aggregation strategy, \mathcal{I} , is uniform sampling. The ratios of aggregated clients per global epoch are 40%, 10%, and 20% for Sent140, FEMNIST, and the other datasets, respectively. The numbers of total clients, N , are 10, 198, 20, and 100 for Sent140, FEMNIST, CIFAR-10, and other datasets. The number of proximal iterations is 5 for all settings with proximal mapping. In our implementations, η_α and η are respectively 0.01 and 0.05.

Summarizing the Effects of Hyper-parameters We test the hyper-parameter effect of η and λ in our implementation pFedBreD_{ns,mh}. The details are in Appendix C. From the results, we find that it will degrade the test accuracy if the values of λ or η are too large or too small. The test accuracy of personalized model is more sensitive than the ones of global model. The test accuracy of personalized

Table 1: Results of average testing accuracy (%) per client of each settings. We mark the best and second best performance by **bold** and underline. Avg and Std : the average results and the standard deviation of them on all tasks; H.Avg and H.Std : the average results and the standard deviation of them on hard tasks (non-linear DNN with complex classification or architecture: DNN / CNN / LSTM on FEMNIST / CIFAR-10 / Sent140). The **G** and **P** are global and personalized model

Methods / Datasets Names - Models	FEMNIST		FMNIST		MNIST		CIFAR-10		Sent140		Statistics			
	MCLR	DNN	MCLR	DNN	MCLR	DNN	CNN	LSTM	Avg	Std	H.Avg	H.Std		
FedAvg [33] -G	53.38 \pm 0.26	57.04 \pm 0.08	82.75 \pm 0.04	80.09 \pm 0.06	86.59 \pm 0.03	88.26 \pm 0.05	57.51 \pm 0.07	70.86 \pm 0.01	72.06	14.34	61.80	7.85		
FedAvg+AM -G	55.34 \pm 0.05	59.03 \pm 0.10	82.58 \pm 0.03	81.03 \pm 0.12	86.74 \pm 0.03	89.31 \pm 0.05	57.07 \pm 0.12	71.27 \pm 0.01	72.80	14.01	62.46	7.70		
FedEM [52] -G	40.75 \pm 0.32	45.47 \pm 0.04	95.78 \pm 0.03	96.42 \pm 0.03	85.75 \pm 0.01	86.49 \pm 0.02	57.67 \pm 0.16	66.72 \pm 0.03	71.88	22.28	56.62	10.66		
pFedBayes [77] -P	49.66 \pm 0.46	-	98.46 \pm 0.05	98.67 \pm 0.05	89.64 \pm 0.06	90.48 \pm 0.12	-	-	-	-	-	-		
FedAMP [31] -P	60.04 \pm 0.08	66.79 \pm 0.04	98.63 \pm 0.02	98.72 \pm 0.01	90.81 \pm 0.02	92.21 \pm 0.02	77.40 \pm 0.04	69.83 \pm 0.05	81.80	15.21	71.34	5.46		
pFedMe [65] -P	50.74 \pm 0.10	53.56 \pm 0.12	97.60 \pm 0.03	98.63 \pm 0.01	88.20 \pm 0.05	90.51 \pm 0.01	72.24 \pm 0.05	69.36 \pm 0.02	77.61	18.96	65.05	10.06		
pFedMe+FT -P	58.04 \pm 0.11	62.93 \pm 0.10	97.63 \pm 0.01	98.39 \pm 0.02	88.36 \pm 0.02	91.71 \pm 0.01	68.17 \pm 0.11	67.82 \pm 0.03	79.13	16.53	66.31	2.93		
pFedMe+AM -P	55.56 \pm 0.09	60.08 \pm 0.05	97.57 \pm 0.02	98.67 \pm 0.00	88.46 \pm 0.02	91.22 \pm 0.00	73.35 \pm 0.09	70.93 \pm 0.05	79.48	16.79	68.12	7.07		
Per-FedAvg [20] -P	54.34 \pm 0.14	62.72 \pm 0.03	94.28 \pm 0.05	97.46 \pm 0.04	87.09 \pm 0.01	90.96 \pm 0.02	78.87 \pm 0.05	70.05 \pm 0.03	79.47	15.74	70.54	8.09		
Per-FedAvg+FT -P	55.34 \pm 0.15	63.34 \pm 0.01	95.76 \pm 0.07	98.10 \pm 0.01	87.56 \pm 0.03	89.58 \pm 0.01	79.68 \pm 0.04	70.20 \pm 0.01	79.95	15.61	71.07	8.20		
Per-FedAvg+AM -P	56.66 \pm 0.09	65.74 \pm 0.02	92.08 \pm 0.10	98.24 \pm 0.02	86.91 \pm 0.04	90.85 \pm 0.02	78.97 \pm 0.03	70.73 \pm 0.05	80.02	14.54	71.81	6.68		
mh (ours) -P	56.34 \pm 0.09	64.93 \pm 0.03	98.44 \pm 0.01	98.73 \pm 0.01	89.83 \pm 0.02	92.04 \pm 0.01	79.44 \pm 0.02	72.04 \pm 0.01	81.47	15.88	72.14	7.26		
mh (ours)+FT -P	59.81 \pm 0.07	67.53 \pm 0.02	98.51 \pm 0.02	98.98 \pm 0.03	90.10 \pm 0.03	92.96 \pm 0.05	79.16 \pm 0.03	71.87 \pm 0.01	82.37	14.92	72.85	5.88		
mh (ours)+AM -P	60.64 \pm 0.02	70.34 \pm 0.01	98.48 \pm 0.01	<u>98.75</u> \pm 0.01	89.88 \pm 0.01	<u>92.32</u> \pm 0.02	80.60 \pm 0.01	73.68 \pm 0.01	83.09	14.01	74.87	<u>5.23</u>		

model is more sensitive to η than to λ . Note that the hyper-parameters are roughly tuned, which shows the insensitivity of **mh**, and better tuning could improve the performance in the Table 1.

6.2 Analysis

Comparative Analysis of Performance We compare our methods and the baselines from different perspectives, including convex or non-convex problems, easy or hard tasks, and text tasks. Additionally, we briefly discuss the absence of BNN on hard tasks.

Convex or non-convex: on non-convex problems, especially in hard tasks, our method significantly outperforms other methods by at least 3.06% employing some simple tricks. On convex problem, FedAMP outperforms our method somewhat on convex problems with simple data sets. One explanation is that the learning landscape is simple in shape for these problems and FedAMP converges faster for this case. One possible reason for this is that since FedAMP uses the distance between models as a similarity in the penalty point selection, giving greater weight to the model that is most similar to the local one. In the later stages of training, since there is only one global optimum, this penalty point tends not to change, and thus the method degenerates into a non-dynamic regular term. Compounding intuition, this method will not be as advantageous for non-convex problems and harder convex problems, as penalty point tends to fall into the local optimum and lead to degradation of the dynamic regular term.

From easy to difficult task: from the difference between the statistics of Avg and H.Avg in Table 1, it can be observed that meta-step methods perform most consistently, with all other methods dropping at least 10%. This is due to the simple and effective local loss design of MAML, with its learning-to-learning design philosophy that enables the method to be more stable in complex situations [19, 20].

Personalized prior on text: text tasks, as opposed to image tasks, generally have relatively rugged learning landscape. [55, 16, 13] This understanding is manifested in specific ways, such as parameter sensitivity, slow convergence, and struggling during the process. Thus, the overlooked prior information seems to be more important, which means that each local iteration not only obtains local knowledge from the data, but also the prior itself already contains some local knowledge. Therefore, there is no need to re-obtain this knowledge from scratch solely from the data during training.

Absence of BNN on hard tasks: complex BNN is not in Table 1, such as LSTM in pFedBayes, because it is difficult to conduct comparative experiments by fixing elements, *e.g.*, inferential models, tricks and optimization methods. In pFedBayes, training often crashes on hard tasks and large datasets, as mentioned in [77]. Our one-step-further research shows that it may be caused by the reparameterization tricks and vanilla Gaussian sampling. If we add tricks on it, the implementation will be very different from the original pFedBayes, and it is beyond this analysis.

Ablation Analysis of Personalized Prior We conduct ablation experiments by dropping the gradient of the Bregman-Moreau envelope, the local loss function, or both, from the personalized

Table 2: Average local test accuracy of personalized model (%) in ablation experiments. (\uparrow/\downarrow : average accuracy is increased/reduced; AC4PP: Additional cost for personalized prior; Grad. and Add.: cost about calculate gradient and addition; Other notations are the same in Table 1.)

Methods	FEMNIST		FMNIST		MNIST		CIFAR-10		Sent140		Statistics		AC4PP
	MCLR	DNN	MCLR	DNN	MCLR	DNN	CNN	LSTM	Avg	H.Avg			
Non-PP	50.7 \pm 0.10	53.6 \pm 0.12	97.6 \pm 0.03	98.6 \pm 0.01	88.2 \pm 0.05	90.5 \pm 0.01	72.2 \pm 0.05	69.4 \pm 0.02	77.6	65.1	None		
lg (ours)	50.8 \pm 0.05	49.1 \pm 0.53	98.3 \pm 0.02	98.4 \pm 0.02	88.4 \pm 0.01	91.0 \pm 0.00	65.7 \pm 0.46	60.7 \pm 0.41	75.3 \downarrow	58.5 \downarrow	Grad. \times R		
meg (ours)	50.3 \pm 0.07	53.9 \pm 0.06	97.8 \pm 0.00	98.6 \pm 0.01	88.4 \pm 0.01	90.6 \pm 0.01	73.8 \pm 0.06	69.4 \pm 0.02	77.9 \uparrow	65.7 \uparrow	Add. \times R		
mh (ours)	56.3 \pm 0.09	64.9 \pm 0.03	98.4 \pm 0.01	98.7 \pm 0.01	89.8 \pm 0.02	92.0 \pm 0.01	79.4 \pm 0.02	72.0 \pm 0.01	81.5 \uparrow	72.1 \uparrow	Both above		

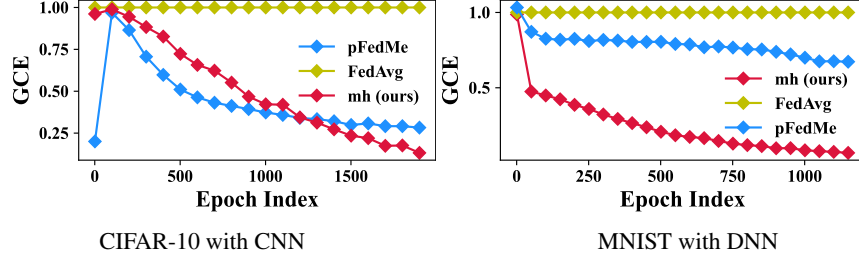


Figure 2: The results of $GCE(\{\nabla F_i(w_i^{(t)})\})$ at each global epoch t after Savitzky-Golay filtering [62].

strategy **mh** as shown in Table 2. The relationship among the three strategies mentioned in Eq. (14) is that **mh** consists of **lg** and **meg**. Moreover, pFedMe can be regarded in our framework as the one which takes the spherical Gaussian as prior and uses vanilla prior selection strategy $\mu_i = I$ without personalization. Thus, pFedMe and the three implementations of pFedBreD are compared. The results reveal the instability of our implementation **lg** and the introduction of **meg** on difficult tasks is about the same as not introducing it. However, introducing both **lg** and **meg** (i.e., **mh**) together shows remarkable performance. This indicates that **lg** and **meg** complement each other. **To explain these results**, by observing the error bars, in most of the settings, **meg** is significantly more stable compared to methods that do not use personalized priors, while **lg** is relatively less stable. Based on this observation, we have reason to believe that **meg** weakens the influence of potential noise, while **lg** introduces new noise. Therefore, we can infer that while the mean parameters are steadily biased towards the personalized model, the introduction of new noise finds a path that is more likely to escape from local optima or saddle points, based on implicit regularization [61, 58, 57].

Generalized Coherence Analysis of Information Injection and Extraction The generalized coherence estimate (GCE) [25] of vectors from personalized to local model (i.e., the envelope gradients in pFedMe and ours) among clients on each global epoch are shown in Figure 2. The smaller the GCE, the less coherent the envelope gradient between individual nodes and the greater the diversity of information in the global model update. As shown in Figure 2, we can observe that during the convergence phase, using a personalized prior method has significantly greater information diversity than not using a personalized prior method, which proves the success of injecting personalized prior knowledge into the global model and extracting local knowledge from the local training.

Variable-Control Analysis of Robustness We analyze the impact of aggregation noise and data heterogeneity [29] on our method, mainly **mh**, by controlling variables. Results are in Table 3 and Table 4. (Details are in Appendix C.8.) We test the performance of global model on different aggregation ratios, where all hyper-parameters except for the aggregation ratios are fixed. Meanwhile, we test the performance of both global and the personalized model on different data heterogeneity settings, where full aggregation (sample client equals total number of clients, $S = N$) and one-step local update (local epoch $R = 1$) are employed to get rid of the effects of aggregation noise and client drift. The experiments demonstrate the instability of the global model in **mh** at small aggregation ratios, which most of the other PFL methods have, by comparing their performances on different aggregation numbers. Comparing to the baselines, the experiments also demonstrate the relative robustness of our method to extreme data heterogeneity.

Table 3: The global test accuracy (%) of the global model with different numbers of clients for aggregation $S \in \{10, 20, 50, 100\}$. (♠:FEMNIST, ◇:FMNIST)

Numbers	small	→	large	Std
♠-DNN	59.0	60.1	60.1	59.8
♠-MCLR	54.4	55.4	55.4	55.5
◇-DNN	75.1	79.6	79.4	79.3
◇-MCLR	80.0	82.6	81.8	82.7

Table 4: The local test accuracy (%) of the personalized model on FMNIST-DNN setting with different data heterogeneity (Non-IID) settings $\alpha \in \{0.01, 0.1, 1, 10, 100, 1000\}$ ($\alpha \downarrow$, Non-IID \uparrow) [29]. The **Bolded** means the best.

Non-IID	small	→	large	Avg
FedAvg-G	18.2	14.8	14.5	11.9
pFedMe-P	89.5	58.2	24.2	12.3
pFedMe-G	17.0	14.3	14.1	12.3
mh(ours)-P	89.6	58.7	25.2	13.1
mh(ours)-G	17.1	14.6	14.6	12.4

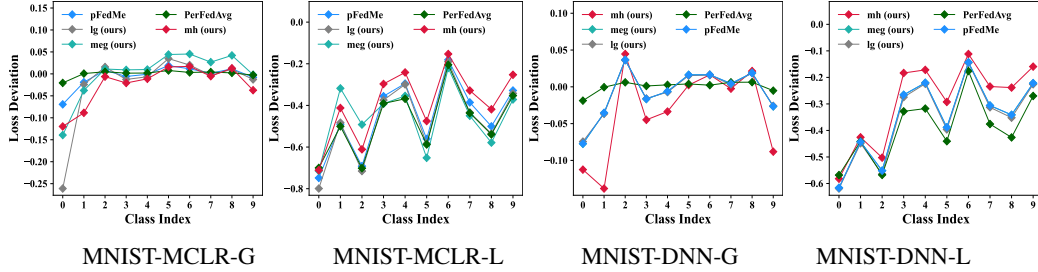


Figure 3: The loss deviation of experiments in Section 6 on the first client, whose major data are on 0th classes. The lower deviation of the available class on global tests and the higher deviation of the unavailable class on local tests demonstrate the superior personalization ability of our methods.

Deviation Analysis of Personalization Deviation represents the difference between an individual and the mean value. We use the deviation of the loss function to reflect the personalization. **On global test**, the lower the deviation, the better the personalized model performance on the corresponding local data. **On local test**, the model is only tested on its own dataset, and because of multiple local iterations, the local test deviation converges to almost the same value, as shown in MNIST-MCLR-L and MNIST-DNN-L in Figure 3. Furthermore, since the local test has a loss of 0 on missing classes, a higher deviation on missing classes reflects a lower mean on these classes. Thus, the lower loss in local testing and better performance are reflected from both of the almost equal deviation in local testing and the higher deviation on missing class. **Summary:** based on Figure 3, we can see that our method has higher deviation on missing classes in local testing and lower deviation in global testing. This means that our method has better personalized performance.

7 Conclusion and Discussion

Conclusion To address the issue of neglecting client-sampling information while providing prior knowledge to local training via direct use of a global model, we propose a general concept: the personalized prior. In this paper, we propose a general framework, pFedBreD, for exploring PFL strategies under the SX-family prior assumption and computation, the RMD to explicitly extract the prior information, and three optional meta-step strategies to personalize the prior. We analyze our proposal both theoretically and empirically. Our strategy **mh** shows remarkable improvement in personalization and robustness to data heterogeneity on non-i.i.d. datasets and the LEAF benchmark [11] with MCLR / DNN / CNN / LSTM as inferential model, which conduct convex / non-convex problems, and image / language benchmarks.

Limitations and Future Work Although **mh** shows remarkable performance and robustness, there is still instability in the global model with aggregation noise. Furthermore, it should be noted that the superficial reason for the improvement of **mh** seems to be that η_α and η and (which are similar to each other) are used simultaneously, resulting in a magnitude in **mh** that is twice as large as the ones in the other two implementations and leading to better performance. However, empirically, simply doubling η_α in **lg** or η in **meg** does not improve performance, and using one more **meg** step used in **lg** significant improvement. Our theoretical analysis cannot explain this phenomenon, and more detailed modeling is needed.

Acknowledge

This work is supported by the Key Program of National Science Foundation of China (Grant No. 61836006) from College of Computer Science (Sichuan University) and Engineering Research Center of Machine Learning and Industry Intelligence (Ministry of Education), Chengdu 610065, P. R. China. This research is also supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-008).

References

- [1] Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. Personalized federated learning with gaussian processes. *Proceedings of the NeurIPS*, 34:8392–8406, 2021.
- [2] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *MACH LEARN*, 50(1):5–43, 2003.
- [3] Manoj Ghuhana Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv:1912.00818*, 2019.
- [4] Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the ICAIT*, volume 117 of *PMLR*, pages 111–132, 08 Feb–11 Feb 2020.
- [5] Ulrich Matchi Aïvodji, Sébastien Gambs, and Alexandre Martin. Iotfla : A secured and privacy-preserving smart home architecture implementing federated learning. In *2019 IEEE SPW*, pages 175–180, 2019.
- [6] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with bregman divergences. *JMLR*, 6(10), 2005.
- [7] Heinz Bauschke, Patrick Combettes, and Dominikus Noll. Joint minimization with alternating bregman proximity operators. *PAC J OPTIM*, 2006.
- [8] Heinz H Bauschke, Jonathan M Borwein, and Patrick L Combettes. Bregman monotone optimization algorithms. *SIAM J CONTROL OPTIM*, 42(2):596–636, 2003.
- [9] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.
- [10] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *COMP MATH MATH PHYS*, 7(3):200–217, 1967.
- [11] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv:1812.01097*, 2018.
- [12] Gary Cheng, Karan Chadha, and John Duchi. Federated asymptotics: a model to compare federated learning algorithms. *arXiv:2108.07313*, 2022.
- [13] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. *NLP*, 2011.
- [14] Katalin Csilléry, Michael G.B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *TRENDS ECOL EVOL*, 25(7):410–418, 2010.
- [15] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *J R STAT SOC B*, 39(1):1–22, 1977.
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *the Proceedings of the CVPR*, pages 248–255, 2009.

- [17] Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated-em with heterogeneity mitigation and variance reduction. *Proceedings of the NeurIPS*, 34:29553–29566, 2021.
- [18] Canh T Dinh, Tung T Vu, Nguyen H Tran, Minh N Dao, and Hongyu Zhang. Fedu: A unified framework for federated multi-task learning with laplacian regularization. *arXiv:2102.07148*, 2021.
- [19] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *Proceedings of the AISTATS*, pages 1082–1092. PMLR, 2020.
- [20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Proceedings of the NeurIPS*, 33:3557–3568, 2020.
- [21] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the ICML*, pages 1126–1135, July 2017. ISSN: 2640-3498.
- [22] Kenneth Friedman and Abner Shimony. Jaynes’s maximum entropy prescription and probability theory. *J STAT PHYS*, 3(4):381–384, 1971.
- [23] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *BRIT J MATH STAT PSY*, 66(1):8–38, 2013.
- [24] Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *STAT SCI*, 1(1):114–135, 1986.
- [25] H. Gish and D. Cochran. Generalized coherence (signal detection). In *ICASSP-88*, pages 2745–2748 vol.5, 1988.
- [26] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General Analysis and Improved Rates. In *Proceedings of the ICML*, pages 5200–5209, May 2019. ISSN: 2640-3498.
- [27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *NEURAL COMPUT*, 9(8):1735–1780, 1997.
- [28] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *Proceedings of the CVPR*, pages 4073–4082, 2015.
- [29] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. *arXiv:1909.06335*, 2019.
- [30] Chenxi Huang, Liang Xie, Yibo Yang, Wenxiao Wang, Binbin Lin, and Deng Cai. Neural collapse inspired federated learning with non-iid data. *arXiv:22303.16066*, 2023.
- [31] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. *AAAI*, 35:7865–7873, 2021.
- [32] Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M. Hadi Amini. A survey on federated learning for resource-constrained iot devices. *IEEE INTERNET THINGS*, 9(1):1–24, 2022.
- [33] Edwin T Jaynes. Information theory and statistical mechanics. *PHYS REV*, 106(4):620, 1957.
- [34] Ji Chu Jiang, Burak Kantarci, Sema Oktug, and Tolga Soyata. Federated learning in smart city sensing: Challenges and opportunities. *Sensors*, 20(21), 2020.
- [35] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv:1909.12488*, 2019.

- [36] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- [37] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the ICML*, pages 5132–5143. PMLR, 2020.
- [38] Robert E Kass and Larry Wasserman. The selection of prior distributions by formal rules. *J AM STAT ASSOC*, 91(435):1343–1370, 1996.
- [39] A Krizhevsky. Learning multiple layers of features from tiny images. *Master’s thesis, University of Tront*, 2009.
- [40] Frederik Kunstner, Raunak Kumar, and Mark Schmidt. Homeomorphic-invariance of em: Non-asymptotic convergence in kl divergence for exponential families via mirror descent. In *Proceedings of the AISTATS*, pages 3295–3303. PMLR, 2021.
- [41] L. Le Cam. Maximum Likelihood: An Introduction. *INT STAT REV*, 58(2):153–171, 1990. Wiley, ISI.
- [42] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [43] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE T KNOWL DATA EN*, 35(4):3347–3366, 2023.
- [44] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the ICML*, pages 6357–6368. PMLR, 2021.
- [45] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *Proceedings of MLSys*, 2:429–450, March 2020.
- [46] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data. In *Proceedings of the ICLR*, April 2020.
- [47] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. *arXiv:2303.10058*, 2023.
- [48] Liangxi Liu, Feng Zheng, Hong Chen, Guo-Jun Qi, Heng Huang, and Ling Shao. A bayesian federated learning framework with online laplace approximation. *arXiv:2102.01936*, 2021.
- [49] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated Learning for Open Banking. In Qiang Yang, Lixin Fan, and Han Yu, editors, *Federated Learning: Privacy and Incentive*, pages 240–254. Springer International Publishing, Cham, 2020.
- [50] Christos Louizos, Matthias Reisser, Joseph Soriaga, and Max Welling. An expectation-maximization perspective on federated learning. *arXiv:2111.10192*, 2021.
- [51] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the ICML*, volume 30, page 3. Citeseer, 2013.
- [52] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated Multi-Task Learning under a Mixture of Distributions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Proceedings of the NeurIPS*, volume 34, pages 15434–15447, 2021.
- [53] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the AISTATS*, pages 1273–1282. PMLR, 2017.

- [54] H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The JMLR*, 18(1):3117–3166, 2017.
- [55] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Proceedings of the NeurIPS*, volume 26, 2013.
- [56] Thomas P. Minka. Expectation propagation for approximate bayesian inference. *arXiv:1301.2294*, 2001.
- [57] Behnam Neyshabur. Implicit regularization in deep learning. *arXiv:1709.01953*, 2017.
- [58] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv:1705.03071*, 2017.
- [59] Dinh C. Nguyen, Ming Ding, Pubudu N. Pathirana, Aruna Seneviratne, Jun Li, and H. Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE COMMUN SURV TUT*, 23(3):1622–1658, 2021.
- [60] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv:1803.02999*, 2018.
- [61] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In Marina Meila and Tong Zhang, editors, *Proceedings of the ICML*, volume 139 of *PMLR*, pages 8913–8924, 18–24 Jul 2021.
- [62] Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *ANAL CHEM*, 36(8):1627–1639, July 1964. American Chemical Society.
- [63] Teddy Seidenfeld. Why i am not an objective bayesian; some reflections prompted by rosenkrantz. *THEOR DECIS*, 11(4):413–440, 1979.
- [64] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *Proceedings of the ICML*, pages 9489–9502. PMLR, 2021.
- [65] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Proceedings of the NeurIPS*, 33:21394–21405, 2020.
- [66] Alysia Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *TNNLS*, 2022.
- [67] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.
- [68] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *Proceedings of the ICLR*, 2022.
- [69] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated Learning for Healthcare Informatics. *JHIR*, 5(1):1–19, March 2021.
- [70] Beining Yang, Kai Wang, Qingyun Sun, Cheng Ji, Xingcheng Fu, Hao Tang, Yang You, and Jianxin Li. Does graph distillation see like vision dataset counterpart? In *Proceedings of the NeurIPS*, 2023.
- [71] Liu Yang, Ben Tan, Vincent W. Zheng, Kai Chen, and Qiang Yang. Federated Recommendation Systems. In Qiang Yang, Lixin Fan, and Han Yu, editors, *Federated Learning: Privacy and Incentive*, pages 225–239. Springer International Publishing, Cham, 2020.
- [72] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *TIST*, 10(2), jan 2019.
- [73] Lei Zhang, Guanyu Gao, and Huaizheng Zhang. Spatial-temporal federated learning for lifelong person re-identification on distributed edges. *arXiv:2207.11759*, 2022.

- [74] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the CVPR*, pages 10174–10183, June 2022.
- [75] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. In *Proceedings of the ICLR*, 2020.
- [76] Wei Zhang, Xiang Li, Hui Ma, Zhong Luo, and Xu Li. Federated learning for machinery fault diagnosis with dynamic validation and self-supervision. *KBS*, 213:106679, 2021.
- [77] Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *Proceedings of the ICML*, pages 26293–26310. PMLR, 2022.
- [78] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *arXiv:2211.13976*, 2022.
- [79] Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. *Proceedings of the NeurIPS*, 32, 2019.
- [80] Yuhao Zhou, Mingjia Shi, Yuanxi Li, Yanan Sun, Qing Ye, and Jiancheng Lv. Communication-efficient federated learning with single-step synthetic features compressor for faster convergence. In *Proceedings of the ICCV*, pages 5031–5040, 2023.

A Glossary, Some Basic Knowledge and Details about Implementations

A.1 Glossary

The main notations in this paper are shown in Table 5.

Table 5: The glossary of notations mentioned in this paper

Notation	Implication
\cdot_i	\cdot on i^{th} client
f_i	local loss function
F_i	local objective function
F	global objective function
\mathbf{E}	expectation on \cdot
\mathbf{H}	entropy
\mathbf{P}	probability measure
\mathbf{P}_{ef}	probability in exponential family
\mathbf{P}_{sef}	probability in scaled exponential family
Ω	complete set of \cdot
\mathcal{F}	generic σ -algebra
$\sigma(\cdot)$	σ -algebra derived from \cdot
$\{\Omega, \mathcal{F}\}$	measurable space
$\{\Omega, \sigma(\Omega), \mathbf{P}\}$	probability measurable space
$\hat{\mathbf{P}}$	estimated probability
x_i, y_i, d_i	input data, label data, the pairs of them
w	global model parameter
Θ_i	local information
θ_i	personalized parameters
w_{init}, θ_{init}	function to initialize parameters
μ_i	the function to generate mean parameter
s_i	the function to generate natural parameter
$x, y, \hat{w}, \theta, \mu, s$	generic point notations
$\mathbf{T}, \mathbf{N}, \mathbf{R}, \mathbf{S}$	number of total global epochs, clients, local epochs, number of sampling clients
\mathbf{t}, \mathbf{r}	global epochs, local epochs
$\beta, \eta, \eta_\alpha, \lambda, \hat{\lambda}$	scalar notations
g, h, h_λ	generic function notations
\mathcal{D}_g	Bregman divergence derived from g
\mathcal{D}_{prox}	Bregman divergence proximal mapping
\mathcal{D}_{env}	Bregman Moreau envelope
$\nabla, \mathbf{D}, \nabla^2$	gradient, Jacobian and Hessian operator
Δ	deviation from mean
\cdot^*	the Fenchel conjugate of \cdot
\mathcal{L}	averaged local test loss
\mathcal{G}	averaged global test loss
$\bar{\cdot}$	mean of \cdot over clients
I, I_m	identity mapping, identity matrix

A.2 Bregman Divergence

Bregman divergence is a general distance satisfying that its first-order moment estimation is the point that minimizes the expectation of the distance to all points for all measurable functions on \mathbf{R}^d . In other words, the given distance \mathcal{D} satisfies Condition (15):

$$\forall X \in \{\mathbf{R}^d, \mathcal{F}, \mathbf{P}\}, \mathbf{E}[X] = \arg \min_y \mathbf{E}[\mathcal{D}(X, y)] \quad (15)$$

Eq. (16) is the definition of Bregman divergence:

$$\begin{aligned} \mathcal{D}_g(x, y) &:= g(x) - g(y) - \langle \nabla g(y), x - y \rangle \\ &= \int_y^x \nabla g(t) - \nabla g(y) dt \end{aligned} \quad (16)$$

where g is a convex function. For convenience, in this paper, g is assumed to be strictly convex, proper and differentiable, so that the equation above Eq. (16) are well-defined. In the perspective of Taylor expansion, Bregman divergence is the first-order residual of g expanded at point y valued at point x , which is the natural connection between Bregman divergence and Legendre transformation. The Bregman divergence does not satisfy the distance axiom, but it provides some of the properties we need, such as non-negative distance. Hence, the selected function g should be convex. Furthermore, if one wants the distance to have a good property that $x = y \leftrightarrow \mathcal{D}_g(x, y) = 0$, one needs g to be strictly convex.

A.3 Non-Maximum Entropy

Besides, the non-maximum entropy rule approach is also worth considering, but we focus on maximum entropy prior in this section. See [63, 24, 38, 23] for additional information of non-maximum entropy assumptions.

A.4 Future PFL

Besides the FTML, Bayesian learning, EM, and transfer learning mentioned in the main paper, neural-collapse-motivated methods and life-long learning are also promising methods to handle PFL problem [47, 30, 73].

PFL could also focus on personalizing other characteristics about FL system, e.g., communications, resource-constrained device. For example, this paper [80] gives a data distillation (compression) [78, 70] method to reduce communication cost, and the compressed data itself contains personalized posterior information.

A.5 Personalized Prior and MAML

Based on previous derivations, to obtain a deployable algorithm, our remaining task is to determine Φ . In this section, inspired by MAML, we briefly introduce a meta-step-based implementation method. The mean parameter is used to represent the prior under SX-family prior assumption given any λ and g in this paper. The mean of the SX-family prior in Eq. (8) is used in regular term, which can be personalized in each client i as μ_i , corresponding to μ_Φ in Eq. (11), as shown in Figure 1. Motivated by this, we use MAML to learn the personalized regularization (or personalized prior in Bayesian learning) in Section 5. For example, \mathbf{meg} in Eq. (14) uses MAML on the Bregman-Moreau envelope $\mathcal{D}_{\mathbf{env}_{g^*, \lambda^{-1}} f_i}$ by substituting it into J in Section 2 and Φ in Eq. (11).

A.6 Sampling Method in Bayesian Learning

Bayesian methods are a elegant solution to the complex issue of heterogeneous data, as they operate on a principle whereby the model allocates increasing attention to local data as available, and derives insight from prior information when local information is scarce. Furthermore, Bayesian modeling brings fresh probabilistic insights to PFL regularization techniques, while simultaneously providing a flexible framework for exploring novel strategies. Bayesian modeling, as well as the expectation maximization and maximum a posteriori estimate (EM-MAP) [15], provide our personalized prior approach with straightforward theoretical support, as well as more general perspectives for analysis. Meanwhile, it addresses the cost of additional sampling in the classic and approximate Bayesian learning paradigm with MAP, the regularization method.

In Bayesian modeling, the EP global loss provides more information that we want to use for local training due to its zero-avoiding property. [56]

The sampling methods used to calculate the solution of Bayesian Model mentioned in this paper can be importance sampling, MCMC or others. In this work, we use the approximation Bayesian methods. See more details in [2]. The local training process based on regular terms differs from Bayesian learning based on sampling, *i.e.*, each time a model needs to be obtained by sampling the model distribution under the current parameters. We choose to use Bayesian MAP as a point estimation as our estimation method, thus eliminating steps such as sampling and reparameterization to improve inference efficiency. The personalized model sampled from local training can be seen as the results from random data sampling using SGD or the mean parameter directly.

Table 6: Complexity Comparison

Complexity/Methods	FedEM	FedAvg	pFedMe	Per-FedAvg	FedAMP	pFedBreD (ours)
Sys. Memory	$\mathcal{O}(NM)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$
Sys. Time	$\mathcal{O}(NTRM)$	$\mathcal{O}(NTR)$	$\mathcal{O}(NTRK)$	$\mathcal{O}(NTR)$	$\mathcal{O}(NTRK)$	$\mathcal{O}(NTRK)$

A.7 First-order Methods

There are three parts in Eq. (13) we need to deal with, and the first-order methods are as shown below:

Jacobian Matrix of Mean: specifically, utilizing the prior selection strategy discussed in Section 4, we have $\mathbf{D}\mu_i(w) = I - \eta \nabla^2 \Phi(w)$. Using different Φ functions yields varying results. For instance, with first-order methods and the last term removed, we get the approximation $\mathbf{D}\mu_i \leftarrow I$.

Hessian Matrix: with first-order methods, we let $\nabla^2 g^*(\cdot) = I_m$. It happens when assuming θ_i obeys the spherical Gaussian by letting $g = \frac{1}{2} \|\cdot\|^2$. Moreover, we can assume θ_i obeys the general multivariable Gaussian by letting $g = \langle \cdot, \Sigma^{-1} \cdot \rangle$ and $\nabla^2 g(\cdot) = \Sigma^{-1} \succeq 0$.

Proximal Mapping: given $\mu_i(w)$, the proximal mapping part $\mathcal{D}\text{prox}_{g^*, \lambda^{-1}} f_i(\mu_i(w))$ can be approximately solved with numerical methods, *e.g.*, gradient descent methods. In other words, we can alternately calculate $\mu_i(w)$ on each client and then fix $\mu_i(w)$ in each local epoch with EM.

A.8 Complexity

Since the general process of our implementations, FedAMP and pFedMe are the same as shown in pFedBreD framework, these methods share the same complexity of memory/calculation, $\mathcal{O}(N)/\mathcal{O}(NTRK)$ as shown in Table 6. The complexities of both FedAvg and Per-FedAvg are $\mathcal{O}(N)/\mathcal{O}(NTR)$ since the original methods of them do not need a approximate proximal mapping solution, and therefore are free on K , the number of iterations to calculate the solution. The complexity of FedEM is $\mathcal{O}(NM)/\mathcal{O}(NTRM)$, where M is the components of the distributions we assume, due to the calculation of M components in each global epoch.

A.9 Broader Impacts

In recent years, PFL has found use not only in predictive tasks like mobile device input methods but also in areas where privacy is paramount, such as healthcare and finance. However, before its widespread deployment, several critical factors must be taken into consideration.

One of the primary concerns regarding PFL is its deployment cost. It involves significant computational resources, making it a costly affair. Additionally, client transparency is an important issue that needs attention. Clients have the right to know what data is being collected and how it is used.

Another factor that complicates PFL’s deployment is the differences in user behavior and hardware and software configurations between clients. These differences can affect the performance of the algorithm and require bespoke solutions for each client.

In addition, PFL’s robustness is another essential aspect to consider. Real-world environments are often unpredictable and can interfere with the algorithm’s performance, leading to erroneous results. Therefore, it is necessary to ensure that the algorithm is sufficiently robust before deploying it.

Lastly, even though PFL offers significant benefits, potential drawbacks should not be overlooked. All stakeholders involved in its deployment need to approach this technology with caution and forethought. By considering these factors, we can harness the power of PFL while minimizing its limitations and risks.

B Details of Equations

B.1 Hidden Information

From the definition of KL divergence, we have

$$\begin{aligned}
& \arg \min_w \mathbf{E}_i \mathbf{E}_{d_i} \mathbf{KL}(\mathbf{P}(y_i|x_i) || \hat{\mathbf{P}}(y_i|x_i, w)) \\
&= \arg \min_w \mathbf{E} \log \mathbf{P}(y_i|x_i) - \log \hat{\mathbf{P}}(y_i|x_i, w) \\
&= \arg \min_w \mathbf{E} - \log \hat{\mathbf{P}}(y_i|x_i, w) \\
&= \arg \max_w \mathbf{E} \log \hat{\mathbf{P}}(y_i|x_i, w)
\end{aligned} \tag{17}$$

This is used in Eq. (5) in the main paper.

B.2 Bregman Divergence and X-Family

We use the SX-family due to its computational advantages. While other families of distributions may be able to handle special cases, they may not be as computationally efficient.

If proper and strictly convex function g is differentiable, with g^* the Fenchel conjugate function of g , $\mathcal{D}_g(x, y)$ the Bregman divergence, μ dual point of s , we have:

$$\mathcal{D}_{g^*}(\mathcal{V}, \mu) = g^*(\mathcal{V}) + g(s) - \langle \mathcal{V}, s \rangle = \mathcal{D}_g[s, \nabla g^*(\mathcal{V})] \tag{18}$$

From the definition of Bregman divergence, $\nabla g(s) = \mu$ and definition of g^* Fenchel conjugate on convex function g , we have:

$$\begin{aligned}
\mathcal{D}_{g^*}(\mathcal{V}, \mu) &= g^*(\mathcal{V}) - g^*(\mu) - \langle \nabla g^*(\mu), \mathcal{V} - \mu \rangle \\
&= g^*(\mathcal{V}) - g^*(\mu) - \langle s, \mathcal{V} - \mu \rangle \\
&= g^*(\mathcal{V}) - \langle s, \mathcal{V} \rangle - g^*(\mu) + \langle \mu, s \rangle \\
&= g^*(\mathcal{V}) - \langle s, \mathcal{V} \rangle + g(s)
\end{aligned} \tag{19}$$

Similarly, we have $\mathcal{D}_g[s, \nabla g^*(\mathcal{V})] = g^*(\mathcal{V}) - \langle s, \mathcal{V} \rangle + g(s)$. This property is used in Eq. (3) and (4) in the main paper.

Table 7: Bregman divergence and exponential family. (note $\xi = \langle \cdot, \ln \cdot \rangle$)

Name	Gaussian	Bernoulli	Poisson	Exponential
Domain	\mathbf{R}^d	$\{0, 1\}$	\mathbf{N}	\mathbf{R}_{++}
$g(y)$	$\frac{1}{2} \ y\ _{\Sigma^{-1}}^2$	$\ln(1 + e^y)$	e^y	$-\ln(-y)$
$\nabla g(y)$	y	$\frac{\exp\{y\}}{1 + \exp\{y\}}$	e^y	$-y^{-1}$
$g^*(x)$	$\frac{1}{2} \ x\ _{\Sigma^{-1}}^2$	$\xi(x) + \xi(1 - x)$	$x \ln(x) - x$	$-\ln(x) - 1$
$\nabla g^*(x)$	x	$\ln(\frac{x}{1-x})$	$\ln(x)$	$-x^{-1}$
$\mathcal{D}_{g^*}(x, y)$	$\frac{1}{2} \ x - y\ _{\Sigma^{-1}}^2$	$\ln(1 + e^{(1-2x)y})$	$e^y + \xi(x) - x(y + 1)$	$\frac{x}{y} - \ln \frac{x}{y} - 1$

Table 7 shows parts of the relationship between specific g and related member in exponential family. See [6] for more about the relationships between g that derives Bregman divergence \mathcal{D}_g and related derived divergence (e.g., $\cdot \Sigma^{-1} \cdot$ & Mahalanobis distance, $\sum \cdot \log \cdot$ & KL divergence / generalized I-divergence and etc.).

B.3 Expectation Maximization

The details of Eq. (7) in the main paper is shown in Eq. (20).

$$\begin{aligned}
\sum_i \log \mathbf{P}(y_i|x_i, w) &= \sum_i \log \int \mathbf{P}(y_i, \Theta_i|x_i, w) d\Theta_i = \sum_i \log \int \mathbf{Q}(\Theta_i) \frac{\mathbf{P}(y_i, \Theta_i|x_i, w)}{\mathbf{Q}(\Theta_i)} d\Theta_i \\
&\geq \sum_i \int \log \mathbf{Q}(\Theta_i) \frac{\mathbf{P}(y_i, \Theta_i|x_i, w)}{\mathbf{Q}(\Theta_i)} d\Theta_i = \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} \log \frac{\mathbf{P}(y_i, \Theta_i|x_i, w)}{\mathbf{Q}(\Theta_i)} \\
&= \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} \log \mathbf{P}(y_i, \Theta_i|x_i, w) - \log \mathbf{Q}(\Theta_i) \\
&\geq \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} \log \mathbf{P}(y_i, \Theta_i|x_i, w) = \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} [\log \hat{\mathbf{P}}(y_i|x_i, \Theta_i, w) + \log \mathbf{P}(\Theta_i|x_i, w)] \\
&= \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} [\log \hat{\mathbf{P}}(y_i|x_i, \Theta_i, w) + \log \int_{y_i} \mathbf{P}(\Theta_i|d_i, w) \mathbf{P}(y_i|x_i, w)] \\
&\geq \sum_i \mathbf{E}_{\mathbf{Q}(\Theta_i)} [\log \hat{\mathbf{P}}(y_i|x_i, \Theta_i, w) + \mathbf{E}_{y_i|x_i, w} \log \mathbf{P}(\Theta_i|d_i, w)]
\end{aligned} \tag{20}$$

In Eq. (20), we use the concavity of logarithmic function for the first inequality and entropy $\mathbf{H}(\mathbf{Q}(\Theta_i)) = \mathbf{E}_{\mathbf{Q}(\Theta_i)} - \log \mathbf{Q}(\Theta_i) \geq 0$ for the second. (probability $\mathbf{Q}(\Theta_i) \in [0, 1]$; The first equal sign holds, when $\mathbf{Q}(\Theta_i) = \mathbf{P}(\Theta_i|d_i, w)$.) The last inequality is derived from the concavity of the logarithmic function.

$$\begin{aligned}
\max_{x,y} f(x, y) &\geq \max_x \max_y f(x, y) \\
\sum_i a_i \max f(x, y_i) &= \max \sum_i a_i f(x, y_i)
\end{aligned} \tag{21}$$

In Eq. (8), we use the two properties of max shown in Eq. (21). Moreover, these properties are also used to build the upper bound of Eq. (10) as Eq. (11).

B.4 Notations of Deviations

The notations are shown as follows:

$\mathcal{L}_{i,c}$: The averaged local test loss of the i^{th} personalized model over its own local test with label c . The value equals zero on the clients without c -labeled data.

$\bar{\mathcal{L}}_c$: The mean of the averaged local test loss over all personalized models. Each $\mathcal{L}_{i,c}$ is weighted by the ratio of the number of own test data with label c .

$\mathcal{G}_{i,c}$: The averaged global test loss of the i^{th} personalized model over the global test with label c .

$\bar{\mathcal{G}}_c$: The mean of the averaged global test loss over all personalized models.

The deviations of the averaged global and local test loss of the i^{th} personalized model on class c : $\Delta \mathcal{G}_{i,c} = \mathcal{G}_{i,c} - \bar{\mathcal{G}}_c$ and $\Delta \mathcal{L}_{i,c} = \mathcal{L}_{i,c} - \bar{\mathcal{L}}_c$.

C More About Experiments

The access of all data and code is available ⁸.

C.1 More about implementations

The three implementations of μ_i , i.e. **lg**, **meg** and **mh**, represent *loss gradient*, *memorized envelope gradient* and *memorized hybrid* respectively. *Memorized* means that we choose the gradient of Bregman-Moreau envelope $\nabla F_i(w_{i,r-1}^{(t)})$ as $\eta[w_{i,R}^{(t-1)} - \theta_{i,r-1}^{(t)}]$, where $\eta \geq 0$ is a step-size-like

⁸https://github.com/BDeMo/pFedBreD_public

hyper-parameter. Each local client memorizes their own local part of the latest global model $w^{(t)}$ at the last global epochs $w_{i,R}^{(t-1)}$, instead of $w_{i,r-1}^{(t)}$ in practice.

C.2 Variant

Based on the facts, the results in Table 1 shows the instability of our personalized models. Here we propose a variant of **mh**, shown in Eq. (22), trying to improve the robustness of personalized model on the original **mh**, which use $\Phi_i \leftarrow f_i + F_i$.

$$\begin{aligned}\Phi_i &\leftarrow \tilde{F}_{i,\tilde{\eta}_\alpha,\tilde{\eta}} := \tilde{\eta}_\alpha f_i \circ (\cdot - \tilde{\eta} \nabla f_i) + F_i \\ \mu_{i,r} &\leftarrow w_{i,r-1}^{(t)} - \eta \nabla \tilde{F}_{i,\tilde{\eta}_\alpha,\tilde{\eta}}(w_{i,r-1}^{(t)}) \\ &= w_{i,r-1}^{(t)} - \eta \{ \tilde{\eta}_\alpha \nabla f_i[w_{i,r-1}^{(t)} - \tilde{\eta} \nabla f_i(w_{i,r-1}^{(t)})] \} - \eta \{ w_{i,R}^{(t-1)} - \theta_{i,r-1}^{(t)} \}\end{aligned}\tag{22}$$

This method in Eq. (22) performance almost the same as the original **mh** when η_α is small, but it provides flexibility to tune the hyper-parameter and decide whether to focus more on the current gradient step or the meta-gradient step by tuning $\tilde{\eta}_\alpha$ and $\tilde{\eta}$. $\tilde{\eta}_\alpha \leftarrow \eta_\alpha/\eta$ and $\tilde{\eta} \leftarrow \eta_\alpha$ are used in practice.

C.3 Implementations of Per-FedAvg

We implement Per-FedAvg with the first-order method [20] and fine-tune the personalized model twice, with each learning step of the global and personalized step sizes.

C.4 Details of Tricks, Datasets and Models

Tricks are shown as follows:

FT: fine-tuning single personalized model one more step for local test.

AM: aggregate momentum, the same trick used in 12th line of Algorithm 1. (To compare more fairly between methods with single global model; $\beta = 2$ for methods and employing AM)

Datasets settings are shown as follows:

CIFAR-10: the whole dataset is separated into 20 clients, and each client has data of 3 classes of label. [18, 39]

FEMNIST: we use non-i.i.d. FEMNIST from LEAF benchmark with fraction of data to sample of 5% and fraction of training data of 90%. [11]

FMNIST: the whole fashion-MNIST dataset is separated into 100 clients, and each client has data of 2 classes of label. [65, 67]

MNIST: the whole MNIST dataset is separated into 100 clients, and each client has data of 3 classes of label. [65, 42]

Sent140: we use non-i.i.d text dataset Sent140 from LEAF benchmark with fraction of data to sample of 5%, fraction of training data of 90% and minimum number of samples per user of 3. Then we re-separate Sent140 into 10 clients with at least 10 samples. [11]

Model settings are shown as follows:

CNN: for the image data, we use convolutional neural network of CifarNet [28].

DNN: the non-linear model is 2 layers deep neural network with 100-dimension hidden layer and activation of leaky ReLU [51] and output of softmax.

MCLR: the linear model, multi-class linear regression, is 1 layer of linear mapping with bias, and then output with softmax.

LSTM: text data model consists of 2 LSTM layers [27] as feature extraction layer of 50-dimension embedding and hidden layer and 2 layers deep neural network as classifier with 100-dimension of hidden layer.

C.5 Non-I.I.D Distribution

Figure 4 shows the non-i.i.d. distribution of MNIST, CIFAR-10, FMNIST, FEMNIST and Sent140. Sent140 is a bi-level classification so each client has two class of label data and we directly use the LEAF benchmark [11] and Dirichlet distribution of $\alpha = 0.5$ to separate users into 10 groups (See the code for more details).

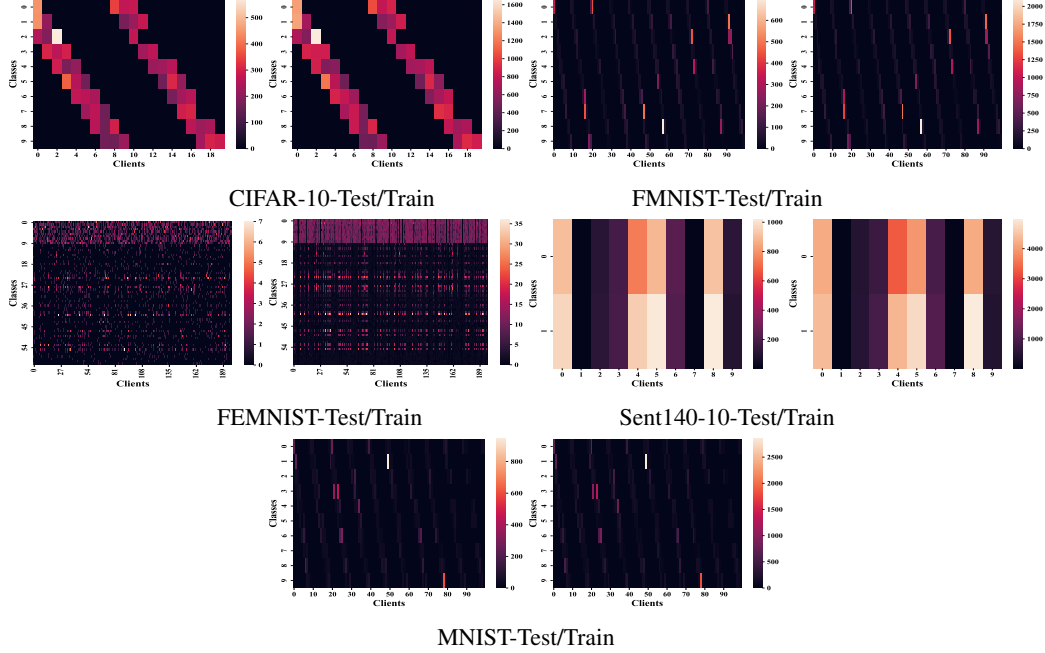


Figure 4: The visualization of the non-i.i.d. data distributions of MNIST, CIFAR-10, FMNIST, FEMNIST and Sent140.

C.6 More About Hyper-Parameter Effect

We post the hyper-parameter effects of η and λ on FEMNIST, FMNIST, MNIST and Sent140 and of η on CIFAR-10 in Figure 5-8. We haven't put the effects of λ on CIFAR-10 for better visualization of the effects of more sensitive eta, as well as our equipment limitations, and the fact that other non-linear models for image classification are already demonstrated on FEMNIST, FMNIST and MNIST. The results of these figures are in the same hyper-parameter settings as mentioned in Section 6.1 except the varying hyper-parameters.

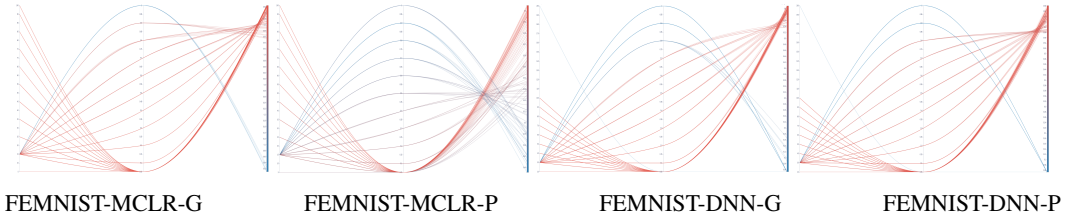


Figure 5: Hyper-parameter effect: The left, middle and right bars in each figure respectively represent λ , η and test accuracy, ranges of which are respectively $[0,100]$, $[0,1]$ and $[0,1]$ increasing from bottom to top (color from blue to red refers to the accuracy from 0 to 1).

C.7 More about Deviation Analysis

The deviations of the global and local test on each settings are shown in Figure 9 mentioned in Section 6.2 in the main paper.

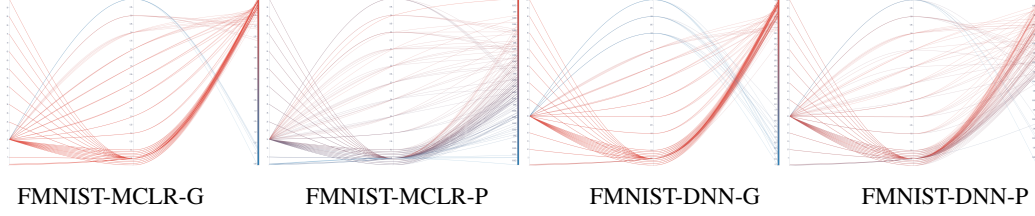


Figure 6: The left, middle and right bars in each figure respectively represent λ , η and test accuracy, ranges of which are respectively $[0,100]$, $[0,1]$ and $[0,1]$ increasing from bottom to top (color from blue to red).

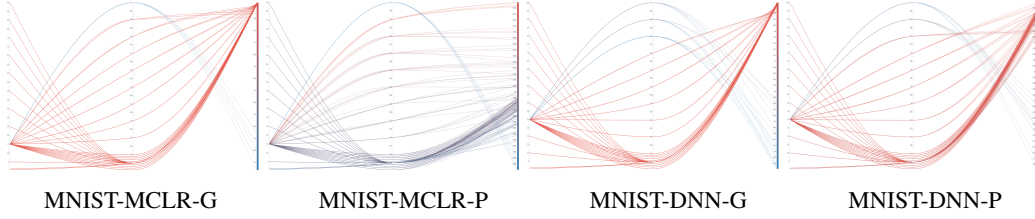


Figure 7: The left, middle and right bars in each figure respectively represent λ , η and test accuracy, ranges of which are respectively $[0,100]$, $[0,1]$ and $[0,1]$ increasing from bottom to top (color from blue to red).

C.8 Experiments about Instability and Robustness on Aggregation Noise and Data Heterogeneity

In this section, we experimentally demonstrate the instability of the global model in **mh** at small aggregation ratios by comparing the performances of clients with different aggregation numbers. Additionally, we also conduct experiments on different data heterogeneity settings.

The experimental settings in Section 6.1 of the main paper have been utilized, with the exception of the client count for aggregation at the culmination of each global epoch. To ensure clarity, we present Table 3 without well-tuning hyper-parameters (which are random selected in a narrow range with Gaussian variance of 0.01). Notably, supplementary experiments have been repeated 5 times to enhance the robustness of our analysis.

The results of experiments about different Non-IID settings are shown in Table 4. The FMNIST in these experiments are equal number of total local data with different local data distribution the distribution are shown in Figure 10. All experiments employ full aggregation of 40 clients and only 1 local epoch to get rid of the effects from aggregation noise and client drift caused by multiple local update.

An interesting example is that if the local classes are only two classes in the case of an extremely unbalanced heterogeneous distribution, the underlying local test accuracy for a personalized model will be at least the probability of the maximum probability class being sampled, say 90% of the first class and 10% of the second class, then a learned knowledge model is at least 90% accurate.

C.9 Additional Experiments

The additional experiments with more baselines are shown in Table 8 with the same settings mentioned in the Table 1.

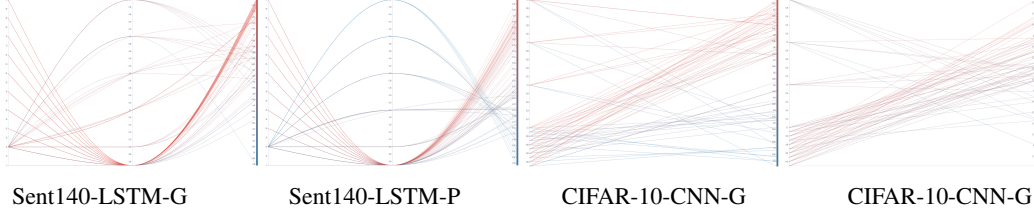


Figure 8: The left, middle and right bars in each figure respectively represent λ and test accuracy, ranges of which are respectively $[0,100]$ and $[0,1]$ increasing from bottom to top (color from blue to red). The ranges of η are respectively $[0,0.5]$ and $[0,0.4]$ in settings of CIFAR-10-CNN and Sent140.

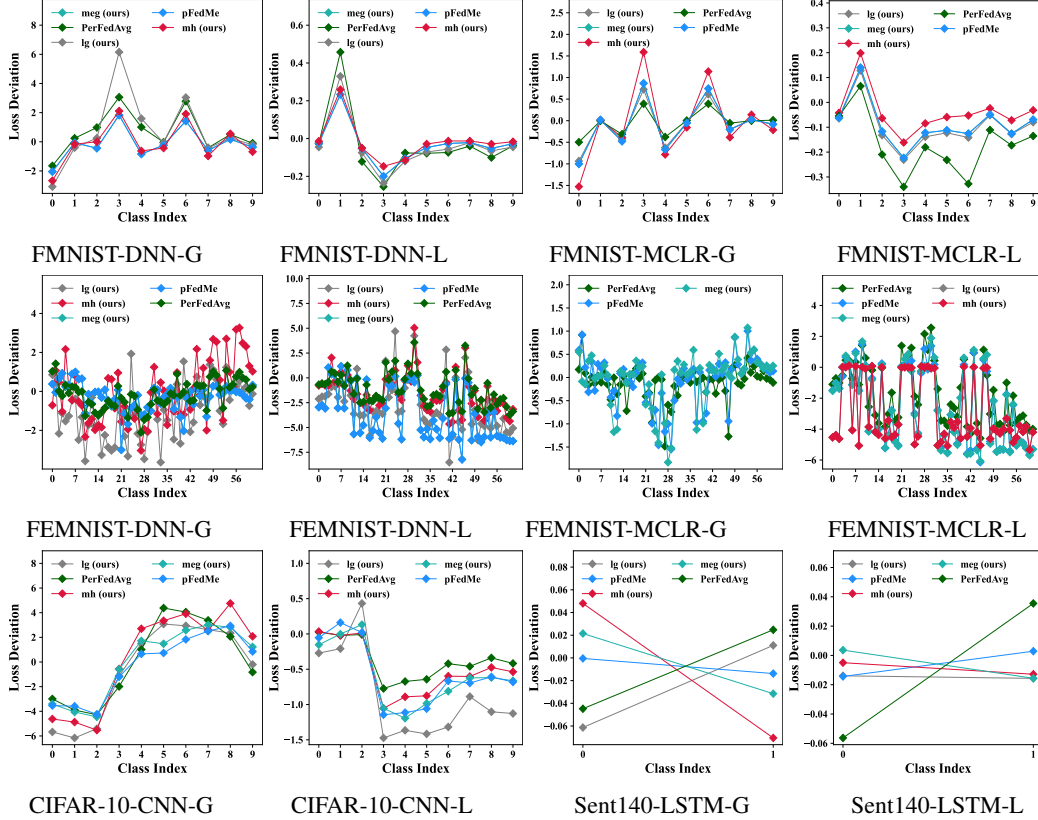


Figure 9: The loss deviation of our experiments in Section 6 on the first client on settings: FEMNIST-DNN/MCLR, FMNIST-DNN/MCLR, CIFAR-10-CNN and Sent140-LSTM.

D Details of Theorems

D.1 Proof Sketch

We prove the theorems primarily through two supporting lemmas. The first lemma provides the upper bound of the global iterative error, while the second lemma restricts the upper bound of the error between the actual local update and theoretical expectation.

D.2 Related Notations

$\cdot_{i,r}^{(t)}$ represents the \cdot on i^{th} client at r^{th} local epoch of t^{th} global epoch.

The Local Sampled Data $\tilde{d}_i \in d_i$

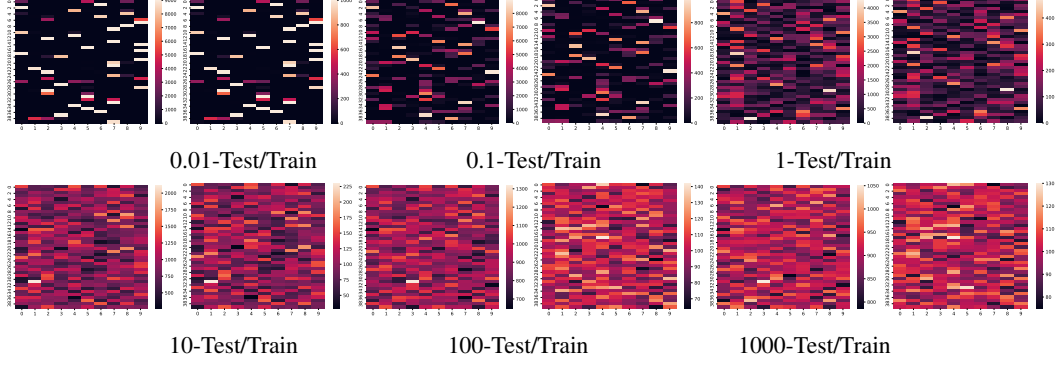


Figure 10: Different heterogeneous distributions of FMNIST. The horizontal and vertical axes represent the different classes and clients respectively.

Table 8: Additional experiments with new baselines. (accuracy)

Methods / Datasets & Models	FEMNIST / DNN	CIFAR-10 / CNN	Sent140 / LSTM	Average Decrease by Noise
Aggregation Ratio	10% \rightarrow 5%	20% \rightarrow 10%	40% \rightarrow 20%	-
FedPAC [68]	62.2% \rightarrow 60.7%	78.9% \rightarrow 77.3%	68.1% \rightarrow 66.8%	1.5%
FedHN [64]	61.1% \rightarrow 59.6%	77.5% \rightarrow 76.9%	71.2% \rightarrow 70.1%	1.1%
Fedfomo [75]	60.1% \rightarrow 58.9%	71.4% \rightarrow 70.6%	70.1% \rightarrow 68.9%	1.1%
Ditto [44]	52.9% \rightarrow 52.2%	72.4% \rightarrow 72.1%	71.0% \rightarrow 70.3%	0.6%
mh(ours)	64.9% \rightarrow 64.3%	79.4% \rightarrow 79.1%	72.0% \rightarrow 71.8%	0.4%

The Approximated Personalized Model $\tilde{\theta}_{i,r}^{(t)} := \tilde{\theta}(\mu_{i,r}^{(t)})$.

The Uniform Local Data Sampling Expectation $\mathbf{E}_{\tilde{d}_i} := \frac{1}{|\tilde{d}_i|} \sum_{\tilde{d}_i \in d_i}$

The Unbiased Empirical First Moment $\mathbf{E}_{\tilde{d}_i} \nabla \tilde{f}_i(\theta; \tilde{d}_i) = \nabla f_i(\theta)$

The Global Minimizer w^* .

The Local Minimizer $\theta_{i,r}^{*(t)} := \mathcal{D}\text{prox}_{g^*, \lambda^{-1}} f_i(\mu_{i,r}^{(t)})$.

The Local Approximate Error $\Delta_{i,r}^{(t)} := \tilde{\theta}_{i,r}^{(t)} - \theta_{i,r}^{*(t)}$.

The Global Approximate Squared Error $\Delta^{(t)} := \mathbf{E} \|w^{(t)} - w^*\|^2$

The Approximated Global Gradient $\mathbf{g}_{i,r}^{(t)} = \lambda \mathbf{D} \mu_{i,r}^{(t)} \nabla^2 g^*(\mu_{i,r}^{(t)}) [\mu_{i,r}^{(t)} - \tilde{\theta}_{i,r}^{(t)}]$

The (first-order) Approximated Envelope Gradient: $\nabla \tilde{F}_i(w)$.

$\|\cdot\|_m$ is any matrix norm, with $\|I\|_m = \hat{u}_m$.

\mathbb{I}_E , indicator function on event E .

The Virtual Global Gradient: $\mathbf{g}^{(t)} = \frac{1}{SR} \sum_{i \in \mathcal{S}^{(t)}} \sum_{r=1}^R \mathbf{g}_{i,r}^{(t)}$.

The Virtual Global Step-size: $\tilde{\alpha}_m = \alpha_m \beta R$.

The Expected Smooth [26] Coefficient of F and F_i : L_F, L_{F_i} .

Bounded Deviation Ratio of Strategy Disturbance Coefficient σ_Φ .

D.3 Basic Propositions

Proposition 1 (μ -strongly convex). *If f is μ -strongly convex, we have:*

$$\begin{aligned} \langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq \mu \|x - y\|^2 \\ \|\nabla f(x) - \nabla f(y)\| &\geq \mu \|x - y\| \end{aligned}$$

Proposition 2 (*L-smooth*). *If f is L -smooth, we have:*

$$\begin{aligned}\langle \nabla f(x) - \nabla f(y), x - y \rangle &\leq L \|x - y\|^2 \\ \|\nabla f(x) - \nabla f(y)\| &\leq L \|x - y\| \\ \|\nabla f(x) - \nabla f(y)\|^2 &\leq 2L \mathcal{D}_f(x, y)\end{aligned}$$

Proposition 3 (Jensen's inequality). *If f is convex, we have:*

$$\mathbf{E}_X f(X) \geq f(\mathbf{E}_X X)$$

. *A variant of the general one shown above:*

$$\left\| \sum_{i=1}^{\mathcal{N}} x_i \right\|^2 \leq \mathcal{N} \sum_i \|x_i\|^2$$

Proposition 4 (triangle inequality). *The triangle inequality:*

$$\|A + B\| \leq \|A\| + \|B\|$$

Proposition 5 (matrix norm compatibility). *The matrix norm compatibility, $A \in \mathbf{R}^{a \times b}, B \in \mathbf{R}^{b \times c}, v \in \mathbf{R}^b$:*

$$\begin{aligned}\|AB\|_m &\leq \|A\|_m \|B\|_m \\ \|Av\|_m &\leq \|A\|_m \|v\|_m\end{aligned}$$

Proposition 6 (Peter Paul inequality).

$$2\langle x, y \rangle \leq \frac{1}{\epsilon} \|x\|^2 + \epsilon \|y\|^2$$

D.4 General Assumptions for Analysis

Assumption 1 (Prior selection). *The given g^* is $\hat{\mu}_{g^*}$ -strongly convex and \hat{L}_{g^*} -smooth: $\hat{\mu}_{g^*} \|x - y\| \leq \|\nabla g^*(x) - \nabla g^*(y)\| < \hat{L}_{g^*} \|x - y\|$, and $\|\nabla^2 g^*(\cdot)\|_m \leq \hat{L}_{g^*}$ (Examples are in Appendix A.7).*

Assumption 2 (Smooth envelope assumption). *For each local envelope $\mathcal{E}_i(\cdot) = [F_i \circ \mu_i^{-1}](\cdot) = \text{Denv}_{g^*, \lambda^{-1}}(\cdot)$, we have $\|\nabla \mathcal{E}_i(x) - \nabla \mathcal{E}_i(y)\|^2 \leq 2\hat{L}_{\mathcal{E}_i} \mathcal{D}_{\mathcal{E}_i}(x, y)$, note that \mathcal{E}_i is convex, $\mathcal{D}_{\mathcal{E}_i}(x, y) := \mathcal{E}_i(x) - \mathcal{E}_i(y) - \langle \nabla \mathcal{E}_i(y), x - y \rangle$. For simplification, we take $\hat{L}_{\mathcal{E}} := \max \hat{L}_{\mathcal{E}_i}, \forall i$ and bounded difference on optimal point $0 \leq \frac{\mathcal{D}_{\mathcal{E}_i}(\mu_i(w), \mu_i(w^*))}{\mathcal{D}_{F_i}(w, w^*)} \leq \tau$.*

Assumption 3 (Strongly convex envelope settings). *f_i is $\hat{\mu}_{f_i}$ -strongly convex: $\hat{\mu}_{f_i} \|x - y\| \leq \|\nabla f_i(x) - \nabla f_i(y)\|$, $\hat{\mu}_f = \min_i \hat{\mu}_{f_i}, \forall i$; f_i is \hat{L}_{f_i} -smooth and non-convex: $\hat{L}_{f_i} \|x - y\| \geq \|\nabla f_i(x) - \nabla f_i(y)\|$, $\hat{L}_f = \max_i \hat{L}_{f_i}, \forall i$. Therefore, we have F_i is $\hat{\mu}_{F_{sc}} := \lambda \hat{\mu}_g + \hat{\mu}_f$ -strongly convex or $\hat{\mu}_{F_{nc}} := \lambda \hat{\mu}_g - \hat{L}_f$ -strongly convex, by tuning λ to make $\lambda \hat{\mu}_g - \hat{L}_f > 0$. We use $\hat{\mu}_F$ as the unified notation for both, for simplification.*

Assumption 4 (Bounded local error). *Since classical gradient descent is used locally, we assume a unified local error bound, $\forall (i, r, t), \|\nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i) + \lambda \nabla \mathcal{D}_{g^*}(\tilde{\theta}_{i,r}^{(t)}, \mu_i)\| \leq c_{i,r}^{(t)} \leq \hat{c}, \forall i$, and a local data sampling shift variance bound $\forall \theta, \mathbf{d} \in d_i, \mathbf{E}_{\mathbf{d}} \|\nabla \tilde{f}_i(\theta; \mathbf{d}) - \nabla f_i(\theta; d_i)\| \leq \gamma_{f_i} \leq \hat{\gamma}_f := \max\{\gamma_{f_i}\}, \forall i$.*

Assumption 5 (RMD meta-step function bound). *$\forall i, \Phi_i$ with limited gradient, $\|\nabla \Phi_i(\cdot)\| \leq \mathcal{G}_{\Phi}$, and Hessian $\|\nabla^2 \Phi_i(w)\|_m \leq \hat{\gamma}_{\Phi}$, therefore, $\|\mathbf{D}\mu_i(w)\|_m = \|I - \eta \nabla^2 \Phi_i(w)\|_m \leq \hat{u}_m + \eta \hat{\gamma}_{\Phi}$.*

Assumption 6 (Bounded deviation ratio of strategy disturbance). *We assume the local training is not affected too much by the personalized prior strategies, which means we don't want a large discrepancy between the results of local strategies formulation and the calculation of local envelope gradients given the prior on each client, which may cause a significant disturbance in the local optimization objective due to the haphazard formulation of prior strategies. Given \forall, w, w' , we have:*

$$\frac{\|\mathbf{D}\mu_i(w) - \mathbf{D}\mu_i(w')\|_m}{\|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\|} \leq \sigma_{\Phi} \frac{\max\{\|\mathbf{D}\mu_i(w)\|_m, \|\mathbf{D}\mu_i(w')\|_m\}}{\max\{\|\nabla \mathcal{E}(\mu_i(w))\|, \|\nabla \mathcal{E}(\mu_i(w'))\|\}}$$

Assumption 7 (Optimal global gradient noise bound). $\|\nabla F_i(w^*)\|^2 \leq \sigma_{F_i, *}^2$, let $\sigma_{F, *}^2 = \max_i \sigma_{F_i, *}^2, \forall i$.

Assumption 8 (First-order approximate bound). $\|\nabla F_i(w) - \nabla \tilde{F}_i(w)\| \leq \epsilon_1$

D.5 General Lemmas

Lemma 1 (Local Sampling Proximal Bound). *Under settings and assumptions in Section 5 and Section D.4, if f is $\hat{\mu}_f$ -strongly convex, $\mathbf{E}_{\tilde{d}_i} \|\Delta_{i,r}^{(t)}\|^2 \leq \frac{2}{(\hat{\mu}_f + \lambda \hat{\mu}_{g^*})^2} [\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \hat{\epsilon}^2]$ holds; if f is \hat{L}_f -smooth and non-convex, $\mathbf{E}_{\tilde{d}_i} \|\Delta_{i,r}^{(t)}\|^2 \leq \frac{2}{(\lambda \hat{\mu}_{g^*} - \hat{L}_f)^2} [\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \hat{\epsilon}^2]$ holds, such that:*

$$\mathbf{E}_{\tilde{d}_i} \|\Delta_{i,r}^{(t)}\|^2 \leq \frac{2}{\hat{\mu}_{F_i}^2} [\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \hat{\epsilon}^2]$$

Proof. With Proposition 1, Assumption 3 and optimal condition of $F_i(\mu_{i,r}^{(t)})$ on $\theta_{i,r}^{*(t)}$, we have:

$$\|\Delta_{i,r}^{(t)}\|^2 = \|\tilde{\theta}_{i,r}^{(t)} - \theta_{i,r}^{*(t)}\|^2 \leq \frac{1}{\mu_{F_i}^2} \|\mathbf{g}_{i,r}^{(t)}\|^2$$

Note that, $\mathbf{g}_{i,r}^{(t)} = \nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \tilde{d}_i) + \lambda \nabla \mathcal{D}_{g^*}(\tilde{\theta}_{i,r}^{(t)}, \mu_{i,r}^{(t)})$. With Proposition 3 and Assumption 4, we have:

$$\begin{aligned} \|\mathbf{g}_{i,r}^{(t)}\|^2 &= \|\nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \tilde{d}_i) - \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i) + \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i) + \lambda \nabla \mathcal{D}_{g^*}(\tilde{\theta}_{i,r}^{(t)}, \mu_i)\|^2 \\ &\leq 2\{\|\nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \tilde{d}_i) - \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i)\|^2 + \|\nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i) + \lambda \nabla \mathcal{D}_{g^*}(\tilde{\theta}_{i,r}^{(t)}, \mu_i)\|^2\} \\ &\leq 2\{\|\nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \tilde{d}_i) - \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i)\|^2 + \hat{\epsilon}^2\} \end{aligned}$$

Taking expectation on both sides, combining both inequalities above, we have:

$$\begin{aligned} \mathbf{E}_{\tilde{d}_i} \|\Delta_{i,r}^{(t)}\|^2 &\leq 2\{\frac{1}{|\tilde{d}_i|^2} \mathbf{E}_{\mathbf{d}} \|\sum_{\mathbf{d} \in \tilde{d}_i} \nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \mathbf{d}) - \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i)\|^2 + \hat{\epsilon}^2\} \\ &\leq 2\{\frac{1}{|\tilde{d}_i|^2} \sum_{\mathbf{d} \in \tilde{d}_i} \mathbf{E}_{\mathbf{d}} \|\nabla \tilde{f}_i(\tilde{\theta}_{i,r}^{(t)}; \mathbf{d}) - \nabla f_i(\tilde{\theta}_{i,r}^{(t)}; d_i)\|^2 + \hat{\epsilon}^2\} \\ &\leq 2[\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \hat{\epsilon}^2] \end{aligned}$$

□

Lemma 2 (Expected-Smooth Personalized Local Object). *Under settings and assumptions in Section 5 and Section D.4, the personalized local objective function is expected-smooth, such that:*

$$\begin{aligned} \|\nabla F_i(w) - \nabla F_i(w')\| &\leq (1 + \sigma_\Phi)(\hat{u}_m + \eta \hat{\gamma}_\Phi) \|\nabla \mathcal{E}_i(\mu_i(w)) - \nabla \mathcal{E}_i(\mu_i(w'))\|, \forall w, w'; \\ \|\nabla F_i(w) - \nabla F_i(w^*)\|^2 &\leq 2(1 + \sigma_\Phi)^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2 \hat{L}_{\mathcal{E}_i} \mathcal{D}_{\mathcal{E}}(\mu_i(w), \mu_i(w^*)) \leq 2\hat{L}_{F_i} \mathcal{D}_{F_i}(w, w^*); \\ \mathbf{E}_i \|\nabla F_i(w) - \nabla F_i(w^*)\|^2 &\leq 2\hat{L}_F [F(w) - F(w^*)] = 2\hat{L}_F \mathcal{D}_F(w, w^*), \end{aligned}$$

where $\hat{L}_{F_i} := \tau(1 + \sigma_\Phi)^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2 \hat{L}_{\mathcal{E}_i}$ and $\hat{L}_F = \max \hat{L}_{F_i}, \forall i$.

Proof. With Assumption 5 and Assumption 6, we have:

$$\begin{aligned} \|\nabla F_i(w) - \nabla F_i(w')\|^2 &= \|\mathbf{D}\mu_i(w) \nabla \mathcal{E}(\mu_i(w)) - \mathbf{D}\mu_i(w') \nabla \mathcal{E}(\mu_i(w'))\|^2 \\ &= \|\mathbf{D}\mu_i(w) \nabla \mathcal{E}(\mu_i(w)) - \mathbf{D}\mu_i(w') \nabla \mathcal{E}(\mu_i(w')) + \mathbf{D}\mu_i(w) \nabla \mathcal{E}(\mu_i(w')) - \mathbf{D}\mu_i(w') \nabla \mathcal{E}(\mu_i(w'))\|^2 \\ &\leq \|\mathbf{D}\mu_i(w) [\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))]\|^2 + \|\mathbf{D}\mu_i(w') - \mathbf{D}\mu_i(w)\|^2 \|\nabla \mathcal{E}(\mu_i(w'))\|^2 \\ &\leq \|\mathbf{D}\mu_i(w)\|_m \|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\| + \|\mathbf{D}\mu_i(w') - \mathbf{D}\mu_i(w)\|_m \|\nabla \mathcal{E}(\mu_i(w'))\| \\ &\leq \max\{\|\mathbf{D}\mu_i(w)\|, \|\mathbf{D}\mu_i(w')\|\} \|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\| \\ &\quad + \max\{\|\nabla \mathcal{E}(\mu_i(w))\|, \|\nabla \mathcal{E}(\mu_i(w'))\|\} \|\mathbf{D}\mu_i(w') - \mathbf{D}\mu_i(w)\|_m \\ &\leq \max\{\|\mathbf{D}\mu_i(w)\|, \|\mathbf{D}\mu_i(w')\|\} \|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\| \\ &\quad + \sigma_\Phi \max\{\|\mathbf{D}\mu_i(w)\|_m, \|\mathbf{D}\mu_i(w')\|_m\} \|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\| \\ &\leq (1 + \sigma_\Phi)(\hat{u}_m + \eta \hat{\gamma}_\Phi) \|\nabla \mathcal{E}(\mu_i(w)) - \nabla \mathcal{E}(\mu_i(w'))\| \end{aligned}$$

where the first two inequalities is by Proposition 4 and Proposition 5.

With the first inequality in our lemma is proven. With the proven one and Assumption 2, we have:

$$\begin{aligned} \|\nabla F_i(w) - \nabla F_i(w^*)\|^2 &\leq 2(1 + \sigma_\Phi)^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2 \hat{L}_{\mathcal{E}_i} \mathcal{D}_{\mathcal{E}}(\mu_i(w), \mu_i(w^*)) \\ &\leq 2(1 + \sigma_\Phi)^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2 \hat{L}_{\mathcal{E}_i} \tau \mathcal{D}_{F_i}(w, w^*) \\ &\leq 2\hat{L}_{F_i} \mathcal{D}_{F_i}(w, w^*); \end{aligned}$$

$$\mathbf{E}_i \|\nabla F_i(w) - \nabla F_i(w^*)\|^2 \leq 2\hat{L}_F [F(w) - F(w^*)] = 2\hat{L}_F \mathcal{D}_F(w, w^*)$$

where the client sampling expectation is taken in the final inequality. \square

Lemma 3 (RMD Personalized Prior Bound). *Under settings and assumptions in Section 5 and Section D.4, the relationship between $\|\nabla F_i(w)\|$ and $\|\nabla \mathcal{E}_i(\mu_i(w))\|$ is:*

$$\|\nabla F_i(w)\| \leq (\hat{u}_m + \eta \hat{\gamma}_\Phi) \|\nabla \mathcal{E}_i(\mu_i(w))\| \leq \lambda \hat{L}_{g^*} (\hat{u}_m + \eta \hat{\gamma}_\Phi) \|\mu_i(w) - \theta_i^*\|$$

Proof. Applying Proposition 5 and Assumption 5, $\nabla F_i(w) = \mathbf{D}\mu_i(w) \nabla \mathcal{E}_i(\mu_i(w))$, it's easy to prove the first inequality. Rewriting $\nabla \mathcal{E}_i(\mu_i(w))$ in detail as shown following, applying Proposition 5 and Assumption 1, the final inequality is proven:

$$\nabla \mathcal{E}_i(\mu_i(w)) = \lambda \nabla^2 g^*(\mu_i(w)) [\mu_i(w) - \theta_i^*]$$

\square

Lemma 4 (Local Objective's Client Sampling Error Bound). *Under settings and assumptions in Section 5 and Section D.4, the upper bound of local sampling error is:*

$$\mathbf{E}_{S^t} \left\| \frac{1}{S} \sum_{i \in S^t} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 \leq \frac{N/S - 1}{N - 1} \sum_i^N \frac{1}{N} \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2$$

, where $|S^{(t)}| = S, \forall t$.

Proof. This lemma is the same lemma in [46, 65].

$$\begin{aligned} \mathbf{E}_{S^t} \left\| \frac{1}{S} \sum_{i \in S^t} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 &= \frac{1}{S^2} \mathbf{E}_{S^t} \left\| \sum_{i \in S^t} \mathbb{I}_{i \in S^t} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 \\ &= \frac{1}{S^2} \left[\sum_{i \in [N]} \mathbf{E}_{S^t} [\mathbb{I}_{i \in S^t}] \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 \right. \\ &\quad \left. + \sum_{i \neq j} \mathbf{E}_{S^t} [\mathbb{I}_{i \in S^t}, \mathbb{I}_{j \in S^t}] \langle \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}), \nabla F_j(w^{(t)}) - \nabla F(w^{(t)}) \rangle \right] \\ &= \frac{1}{SN} \sum_i^N \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 \\ &\quad + \sum_{i \neq j} \frac{S - 1}{SN(N - 1)} \langle \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}), \nabla F_j(w^{(t)}) - \nabla F(w^{(t)}) \rangle \\ &= \frac{1}{SN} \left(1 - \frac{S - 1}{N - 1} \right) \sum_{i \in [N]} \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 \\ &= \frac{N/S - 1}{N - 1} \sum_{i \in [N]} \frac{1}{N} \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 \end{aligned}$$

where $\mathbb{I} \in \{0, 1\}$ is indicator function, $\mathbf{E}_{S^t} [\mathbb{I}_{i \in S^t}] = \frac{S}{N}$ and $\mathbf{E}_{S^t} [\mathbb{I}_{i \in S^t}, \mathbb{I}_{j \in S^t}] = \frac{S(S-1)}{N(N-1)}, \forall i \neq j$. Note that:

$$\sum_i^N \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 + \sum_{i \neq j} \langle \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}), \nabla F_j(w^{(t)}) - \nabla F(w^{(t)}) \rangle = 0.$$

\square

Lemma 5 (Variance of Global Aggregation on Client Sampling Bound). *Under settings and assumptions in Section 5 and Section D.4, the upper bound of gradient aggregation variance is:*

$$\mathbf{E}_i \|\nabla F_i(w) - \nabla F(w)\|^2 \leq \mathbf{E}_i \|\nabla F_i(w)\|^2 \leq 4\hat{L}_F \mathcal{D}_F(w, w^*) + 2\sigma_{F,*}^2$$

Proof.

$$\begin{aligned} \mathbf{E}_i \|\nabla F_i(w) - \nabla F(w)\|^2 &\leq \mathbf{E}_i \|\nabla F_i(w)\|^2 \leq 2\mathbf{E}_i [\|\nabla F_i(w) - \nabla F_i(w^*)\|^2 + \|\nabla F_i(w^*)\|^2] \\ &\leq 4\hat{L}_F \mathcal{D}_F(w, w^*) + 2\sigma_{F,*}^2 \end{aligned}$$

where the first inequality is by $\mathbf{E}[\|X\|^2] = \mathbf{E}[\|X - \mathbf{E}[X]\|^2] + \mathbf{E}[\|\mathbf{E}[X]\|^2]$, the second one is by Proposition 3 and the final one is by Lemma 2 and Assumption 7. \square

D.6 Supporting Lemmas

Lemma 6 (Global Iteration Bound). *Under settings and assumptions in Section 5 and Section D.4, the upper bound of global iteration error is:*

$$\begin{aligned} \mathbf{E}_{\cdot|t} \|w^{(t+1)} - w^*\|^2 &\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \|w^{(t)} - w^*\|^2 + \frac{3\tilde{\alpha}_m^2 + 2\tilde{\alpha}_m/\hat{\mu}_F}{NR} \sum_{i,r}^{N,R} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\| \\ &\quad + 3\tilde{\alpha}_m^2 \mathbf{E}_{\cdot|t} \left\| \frac{1}{S} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 \\ &\quad + (3\tilde{\alpha}_m^2 \hat{L}_F - 2\tilde{\alpha}_m) \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) \end{aligned}$$

Proof. To separate the norm, we have:

$$\begin{aligned} \mathbf{E}_{\cdot|t} \|w^{(t+1)} - w^*\|^2 &= \mathbf{E}_{\cdot|t} [\|w^{(t)} - \tilde{\alpha}_m \mathbf{g}^{(t)} - w^*\|^2] \\ &= \|w^{(t)} - w^*\|^2 - 2\tilde{\alpha}_m \mathbf{E}_{\cdot|t} [\langle \mathbf{g}^{(t)}, w^{(t)} - w^* \rangle] + \tilde{\alpha}_m^2 \mathbf{E}_{\cdot|t} [\|\mathbf{g}^{(t)}\|^2] \end{aligned}$$

The second term:

$$\begin{aligned} -2\tilde{\alpha}_m \mathbf{E}_{\cdot|t} [\langle \mathbf{g}^{(t)}, w^{(t)} - w^* \rangle] &= -2\tilde{\alpha}_m \langle \mathbf{E}_{\cdot|t} \mathbf{g}^{(t)}, w^{(t)} - w^* \rangle \\ &= -2\tilde{\alpha}_m \frac{1}{NR} \sum_{i,r}^{N,R} [\langle \mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle + \langle \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle] \\ &= \frac{\tilde{\alpha}_m}{NR} \sum_{i,r}^{N,R} [-2\langle \mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle - 2\tilde{\alpha}_m \mathbf{E}_i \langle \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle] \end{aligned}$$

Each of the two factors of the second term is bounded (note that $\mathbf{E}_i = \frac{1}{N} \sum_{i=1}^N$ is discussed):

$$\begin{aligned} -\mathbf{E}_i \langle \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle &\leq -\mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) - \mathbf{E} \frac{\hat{\mu}_F}{2} \|w^{(t)} - w^*\|^2 \\ -2\langle \mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)}), w^{(t)} - w^* \rangle &\leq \frac{2}{\hat{\mu}_F} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\| + \frac{\hat{\mu}_F}{2} \|w^{(t)} - w^*\|^2 \end{aligned}$$

where the first inequality is by Proposition 1 and the second one is by Proposition 6.

The third term:

$$\begin{aligned} \mathbf{E}_{\cdot|t} \|\mathbf{g}^{(t)}\|^2 &= \mathbf{E}_{\cdot|t} \left\| \frac{1}{SR} \sum_{i,r}^{\mathcal{S}^{(t)},R} \mathbf{g}_{i,r}^{(t)} \right\|^2 \leq 3\mathbf{E}_{\cdot|t} \left\| \frac{1}{SR} \sum_{i,r}^{\mathcal{S}^{(t)},R} \mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)}) \right\|^2 \\ &\quad + \left\| \frac{1}{S} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 + \|\nabla F(w^{(t)})\|^2 \\ &\leq 3 \left[\frac{1}{NR} \sum_{i,r}^{N,R} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 \right. \\ &\quad \left. + \mathbf{E}_{\cdot|t} \left\| \frac{1}{S} \sum_{i \in \mathcal{S}^{(t)}} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 + \hat{L}_F \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) \right] \end{aligned}$$

where the first inequality is by Proposition 3 the second one is by Lemma 5.

Thus, if we combine each term back into the separation at the very beginning of this proof, the lemma is proven. \square

Lemma 7 (Local-Global Client Drift Bound). *Under settings and assumptions in Section 5 and Section D.4, by choosing a proper $\tilde{\alpha}_m \leq \frac{\beta}{\sqrt{2\dot{c}}}$, the client drift bound is:*

$$\frac{1}{NR} \sum_{i,r}^{N,R} \mathbf{E}_{\cdot|t,i} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 \leq \dot{\delta} + \dot{c} \alpha_m^2 2^{R+1} \{(1+2R) \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] + \dot{\delta}\}$$

where $\dot{\delta} = 4[\lambda \frac{\hat{L}_{g^*}}{\hat{\mu}_F} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 (\frac{\hat{\gamma}_f^2}{|\hat{d}_i|} + \hat{\epsilon}^2) + 16[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}} (\hat{u}_m + \eta \hat{\gamma}_\Phi) \hat{\gamma}_\Phi]^2$ and $\dot{c} = 4[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2$.

Proof.

$$\begin{aligned} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 &\leq 2[\|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w_{i,r}^{(t)})\|^2 + \|\nabla F_i(w_{i,r}^{(t)}) - \nabla F_i(w^{(t)})\|^2] \\ &\leq 2\{[\lambda \hat{L}_{g^*} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \|\Delta_{i,r}^{(t)}\|^2 \\ &\quad + (1 + \sigma_\Phi)^2 (\hat{u}_m + \eta \hat{\gamma}_\Phi)^2 \|\nabla \mathcal{E}_i(\mu_i(w_{i,r}^{(t)})) - \nabla \mathcal{E}_i(\mu_i(w^{(t)}))\|^2\} \\ &\leq 2\{[\lambda \hat{L}_{g^*} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \|\Delta_{i,r}^{(t)}\|^2 + [(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m \\ &\quad + \eta \hat{\gamma}_\Phi)]^2 \|\mu_i(w_{i,r}^{(t)}) - \mu_i(w^{(t)})\|^2\} \\ &\leq 2\{[\lambda \hat{L}_{g^*} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \|\Delta_{i,r}^{(t)}\|^2 + [(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m \\ &\quad + \eta \hat{\gamma}_\Phi)]^2 [2\|w_{i,r}^{(t)} - w^{(t)}\|^2 + 2\|\nabla^2 \Phi(w_{i,r}^{(t)}) - \nabla^2 \Phi(w^{(t)})\|^2]\} \\ &\leq 2\{[\lambda \hat{L}_{g^*} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \|\Delta_{i,r}^{(t)}\|^2 + 2[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \|w_{i,r}^{(t)} - w^{(t)}\|^2 \\ &\quad + 8[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m + \eta \hat{\gamma}_\Phi) \hat{\gamma}_\Phi]^2\} \end{aligned} \tag{23}$$

where the first inequality is by Proposition 3, the second one is by Lemma 2, the third one is by Assumption 2 and Proposition 2, the fourth one is by Proposition 3 and bringing in Equation (11) and the final one is by Assumption 5.

With Lemma 1, we have:

$$\begin{aligned} \mathbf{E}_{\cdot|t,i} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 &\leq 4[\lambda \frac{\hat{L}_{g^*}}{\hat{\mu}_F} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 (\frac{\hat{\gamma}_f^2}{|\hat{d}_i|} + \hat{\epsilon}^2) \\ &\quad + 16[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m + \eta \hat{\gamma}_\Phi) \hat{\gamma}_\Phi]^2 + 4[(1 + \sigma_\Phi) \hat{L}_{\mathcal{E}_i} (\hat{u}_m + \eta \hat{\gamma}_\Phi)]^2 \mathbf{E}_{\cdot|t,i} \|w_{i,r}^{(t)} - w^{(t)}\|^2 \end{aligned}$$

For simplification: $\mathbf{E}_{\cdot|t,i} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 \leq \dot{\delta} + \dot{c} \mathbf{E}_{\cdot|t,i} \|w_{i,r}^{(t)} - w^{(t)}\|^2$

The second term:

$$\begin{aligned}
\mathbf{E}_{\cdot|t,i} \|w_{i,r}^{(t)} - w^{(t)}\|^2 &= \mathbf{E}_{\cdot|t,i} [\|w_{i,r-1}^{(t)} - w^{(t)} - \alpha_m \mathbf{g}_{i,r-1}^{(t)}\|^2] \\
&\leq 2\mathbf{E}_{\cdot|t,i} [\|w_{i,r-1}^{(t)} - w^{(t)} - \alpha_m \nabla F_i(w^{(t)})\|^2] \\
&\quad + \alpha_m^2 \|\mathbf{g}_{i,r-1}^{(t)} - \nabla F_i(w^{(t)})\|^2] \\
&\leq 2(1 + \frac{1}{2R})\mathbf{E}_{\cdot|t,i} [\|w_{i,r-1}^{(t)} - w^{(t)}\|^2] + 2(1 + 2R)\alpha_m^2 \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] \\
&\quad + 2\alpha_m^2 [\dot{\delta} + \dot{c}\mathbf{E}_{\cdot|t,i} \|w_{i,r-1}^{(t)} - w^{(t)}\|^2] \\
&\leq 2(1 + \frac{1}{2R} + \alpha_m^2 \dot{c})\mathbf{E}_{\cdot|t,i} [\|w_{i,r-1}^{(t)} - w^{(t)}\|^2] \\
&\quad + 2(1 + 2R)\alpha_m^2 \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] + 2\alpha_m^2 \dot{\delta} \\
&\leq 2(1 + \frac{1}{R})\mathbf{E}_{\cdot|t,i} [\|w_{i,r-1}^{(t)} - w^{(t)}\|^2] + 2(1 + 2R)\alpha_m^2 \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] \\
&\quad + 2\alpha_m^2 \dot{\delta}
\end{aligned}$$

where the first inequality is by Proposition 3, the second one is by Proposition 6 and the simplified inequality and the final one is by choose $\tilde{\alpha}_m^2 \leq \frac{\beta^2}{2\dot{c}}$, and $\alpha_m^2 \dot{c} \leq \frac{1}{2R^2} \leq \frac{1}{2R}$.

To recursively unroll:

$$\begin{aligned}
&\mathbf{E}_{\cdot|t,i} \|w_{i,r}^{(t)} - w^{(t)}\|^2 \\
&\leq \{(1 + 2R)\alpha_m^2 \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] + \alpha_m^2 \dot{\delta}\} \sum_{\tilde{r}=0}^r 2^{\tilde{r}+1} (1 + \frac{1}{R})^{\tilde{r}} \\
&\leq \{(1 + 2R)\alpha_m^2 \mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] + \alpha_m^2 \dot{\delta}\} \sum_{\tilde{r}=0}^{R-1} 2^{\tilde{r}+1} (1 + \frac{1}{R})^{\tilde{r}} \\
&\leq \alpha_m^2 e 2^{R+1} \{(1 + 2R)\mathbf{E}_{\cdot|t,i} [\|\nabla F_i(w^{(t)})\|^2] + \dot{\delta}\}
\end{aligned} \tag{24}$$

Thus, bringing in the recursively unrolled inequality back into the simplified one, the lemma's proven. \square

D.7 Proof of Theorems

D.7.1 Proof of Theorem 1

The proof of Theorem 1 is shown as followings:

Proof. With Lemma 6, we have:

$$\begin{aligned}
\Delta^{(t+1)} &:= \mathbf{E} \|w^{(t+1)} - w^*\|^2 \\
&\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} + \frac{3\tilde{\alpha}_m^2 + 2\tilde{\alpha}_m / \hat{\mu}_F}{NR} \sum_{i,r}^{N,R} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\| \\
&\quad + 3\tilde{\alpha}_m^2 \mathbf{E}_{\cdot|t} \|\frac{1}{S} \sum_{i \in S^{(t)}} \nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 + (3\tilde{\alpha}_m^2 \hat{L}_F - 2\tilde{\alpha}_m) \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*)
\end{aligned}$$

With Lemma 4, we have:

$$\begin{aligned}
\Delta^{(t+1)} &\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} + \frac{3\tilde{\alpha}_m^2 + 2\tilde{\alpha}_m / \hat{\mu}_F}{NR} \sum_{i,r}^{N,R} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\| \\
&\quad + 3\tilde{\alpha}_m^2 \frac{N/S - 1}{N - 1} \mathbf{E}_i \|\nabla F_i(w^{(t)}) - \nabla F(w^{(t)})\|^2 + (3\tilde{\alpha}_m^2 \hat{L}_F - 2\tilde{\alpha}_m) \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*)
\end{aligned}$$

With Lemma 5, we have:

$$\mathbf{E}_i \|\nabla F_i(w) - \nabla F(w)\|^2 \leq 4\hat{L}_F \mathcal{D}_F(w, w^*) + 2\sigma_{F,*}^2$$

Thus, the inequality is:

$$\begin{aligned}\Delta^{(t+1)} &\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} + \frac{3\tilde{\alpha}_m^2 + 2\tilde{\alpha}_m/\hat{\mu}_F}{NR} \sum_{i,r}^{N,R} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\| \\ &\quad + 3\tilde{\alpha}_m^2 \frac{N/S-1}{N-1} [4\hat{L}_F \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) + 2\sigma_{F,*}^2] + (3\tilde{\alpha}_m^2 \hat{L}_F - 2\tilde{\alpha}_m) \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*)\end{aligned}\quad (25)$$

With Lemma 7 and $\tilde{\alpha}_m \leq \frac{\beta}{\sqrt{2\hat{c}}}$, by taking full expectation of all variables noted by \mathbf{E} , we have:

$$\begin{aligned}\frac{1}{NR} \sum_{i,r}^{N,R} \mathbf{E} \|\mathbf{g}_{i,r}^{(t)} - \nabla F_i(w^{(t)})\|^2 &\leq e\dot{c}\alpha_m^2 2^{R+1} (1+2R) \mathbf{E}_i [\|\nabla F_i(w^{(t)})\|^2] + (e\dot{c}\alpha_m^2 2^{R+1} + 1)\dot{\delta} \\ &\leq e\dot{c}\alpha_m^2 2^{R+1} (1+2R) \mathbf{E}_i [2\|\nabla F_i(w^{(t)}) - \nabla F_i(w^*)\|^2 \\ &\quad + 2\|\nabla F_i(w^*)\|^2] + (e\dot{c}\alpha_m^2 2^{R+1} + 1)\dot{\delta} \\ &\leq e\dot{c}\alpha_m^2 2^{R+3} (1+2R) \hat{L}_F \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) \\ &\quad + e\dot{c}\alpha_m^2 2^{R+2} (1+2R) \sigma_{F,*}^2 + (e\dot{c}\alpha_m^2 2^{R+1} + 1)\dot{\delta}\end{aligned}$$

where the second inequality is by Proposition 3 and the final one is using Lemma 2 and Assumption 7. With this inequality, Equation (25) turns into:

$$\begin{aligned}\Delta^{(t+1)} &\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} + 6\tilde{\alpha}_m^2 \frac{N/S-1}{N-1} \sigma_{F,*}^2 \\ &\quad + (3\tilde{\alpha}_m^2 + 2\tilde{\alpha}_m/\hat{\mu}_F) [e\dot{c}\tilde{\alpha}_m^2 \frac{2^{R+2}(1+2R)}{\beta^2 R^2} \sigma_{F,*}^2 + (e\dot{c}\tilde{\alpha}_m^2 \frac{2^{R+1}}{\beta^2 R^2} + 1)\dot{\delta}] \\ &\quad + \{e\dot{c}(3\tilde{\alpha}_m + 2/\hat{\mu}_F) \tilde{\alpha}_m^3 \frac{2^{R+3}(1+2R)}{\beta^2 R^2} \hat{L}_F + 12\tilde{\alpha}_m^2 \frac{N/S-1}{N-1} \hat{L}_F \\ &\quad + 3\tilde{\alpha}_m^2 \hat{L}_F - 2\tilde{\alpha}_m\} \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*)\end{aligned}$$

To simplify this inequality with condition $\tilde{\alpha}_m \leq \min\{\frac{\beta}{\sqrt{2\hat{c}}}, \frac{2}{\hat{\mu}_F}\}$, we have:

$$\begin{aligned}\Delta^{(t+1)} &\leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} + 6\tilde{\alpha}_m^2 \frac{N/S-1}{N-1} \sigma_{F,*}^2 \\ &\quad + \frac{2^{R+4}e\dot{c}}{\hat{\mu}_F \beta^2 R^2} [2(1+2R) \sigma_{F,*}^2 + \dot{\delta}] \tilde{\alpha}_m^3 + \frac{8\dot{\delta}}{\hat{\mu}_F} \tilde{\alpha}_m \\ &\quad - \{2 - \tilde{\alpha}_m [\frac{e(1+\sigma_\Phi) \hat{L}_\mathcal{E}(\hat{u}_m + \eta \hat{\gamma}_\Phi) 2^{R+6\frac{1}{2}} (\frac{1}{R} + 2)}{\hat{\mu}_F \beta R} \\ &\quad + 12 \frac{N/S-1}{N-1} + 3] \hat{L}_F\} \tilde{\alpha}_m \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*)\end{aligned}$$

where we use $3\tilde{\alpha}_m^2 \leq \frac{6}{\hat{\mu}_F}$ and $\dot{c}\tilde{\alpha}_m \leq \sqrt{2}\beta(1+\sigma_\Phi) \hat{L}_\mathcal{E}(\hat{u}_m + \eta \hat{\gamma}_\Phi)$

Let $\dot{c}_1 := 2 - \tilde{\alpha}_m [\frac{e(1+\sigma_\Phi) \hat{L}_\mathcal{E}(\hat{u}_m + \eta \hat{\gamma}_\Phi) 2^{R+6\frac{1}{2}} (\frac{1}{R} + 2)}{\hat{\mu}_F \beta R} + 12 \frac{N/S-1}{N-1} + 3] \hat{L}_F$, and we have $\dot{c}_1 \geq 1$, when $\tilde{\alpha}_m$ satisfies:

$$\begin{aligned}\tilde{\alpha}_m &\leq \hat{\alpha}_m := \frac{\hat{\mu}_F \beta R}{e(1+\sigma_\Phi) \hat{L}_\mathcal{E}(\hat{u}_m + \eta \hat{\gamma}_\Phi) 2^{R+6\frac{1}{2}} (\frac{1}{R} + 2) + 15(\hat{\mu}_F \beta R) \hat{L}_F} \\ &\leq \frac{1}{[\frac{e(1+\sigma_\Phi) \hat{L}_\mathcal{E}(\hat{u}_m + \eta \hat{\gamma}_\Phi) 2^{R+6\frac{1}{2}} (\frac{1}{R} + 2)}{\hat{\mu}_F \beta R} + 12 \frac{N/S-1}{N-1} + 3] \hat{L}_F}\end{aligned}\quad (26)$$

By setting $\tilde{\alpha}_m$ with Equation (26), then let $\xi^{(t)} = (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2})^{-t-1}$ and $\mathcal{X}^{(T)} = \sum_{t=0}^{T-1} \xi^{(t)}$, $\tilde{\alpha}T \geq \frac{2}{\hat{\mu}_F}$, $\tilde{\alpha}_m \leq \min\{\frac{\beta}{\sqrt{2\hat{c}}}, \frac{2}{\hat{\mu}_F}\}$, we have:

$$\Delta^{(t+1)} \leq (1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}) \Delta^{(t)} - \tilde{\alpha}_m \mathbf{E} \mathcal{D}_F(w^{(t)}, w^*) + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^j$$

where $\dot{\delta}_1 := \frac{8\dot{\delta}}{\hat{\mu}_F}$, $\dot{\delta}_2 := 6\frac{N/S-1}{N-1}\sigma_{F,*}^2$ and $\dot{\delta}_3 := \frac{2^{R+4}e\dot{c}}{\hat{\mu}_F\beta^2R^2}[2(1+2R)\sigma_{F,*}^2 + \dot{\delta}]$.

Reformulate it as following:

$$\mathbf{E}\mathcal{D}_F(w^{(t)}, w^*) \leq \frac{1}{\tilde{\alpha}_m} \left[\left(1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}\right) \mathbf{\Delta}^{(t)} - \mathbf{\Delta}^{(t+1)} \right] + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1}$$

Multiply both sides with $\xi^{(t)}$ and accumulate over t :

$$\begin{aligned} \mathbf{E}\mathcal{D}_F\left(\frac{\sum_{t=0}^{T-1} \xi^{(t)} w^{(t)}}{\mathcal{X}^{(T)}}, w^*\right) &\leq \frac{\sum_{t=0}^{T-1} \xi^{(t)}}{\mathcal{X}^{(T)}} \mathbf{E}\mathcal{D}_F(w^{(t)}, w^*) \\ &\leq \frac{1}{\tilde{\alpha}_m \mathcal{X}^{(T)}} \sum_{t=0}^{T-1} \left[\left(1 - \frac{\tilde{\alpha}_m \hat{\mu}_F}{2}\right) \xi^{(t)} \mathbf{\Delta}^{(t)} - \xi^{(t)} \mathbf{\Delta}^{(t+1)} \right] + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \\ &= \frac{1}{\tilde{\alpha}_m \mathcal{X}^{(T)}} \mathbf{\Delta}^{(0)} - \frac{\xi^{(T-1)}}{\tilde{\alpha}_m \mathcal{X}^{(T)}} \mathbf{\Delta}^{(T)} + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \\ &= \frac{\hat{\mu}_F}{2\xi^{(T-1)}[1 - (1 - \tilde{\alpha}_m \hat{\mu}_F/2)^T]} \mathbf{\Delta}^{(0)} - \frac{\xi^{(T-1)}}{\tilde{\alpha}_m \mathcal{X}^{(T)}} \mathbf{\Delta}^{(T)} + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \\ &\leq \hat{\mu}_F e^{-\tilde{\alpha}_m \hat{\mu}_F T/2} \mathbf{\Delta}^{(0)} - \frac{\hat{\mu}_F}{2} \mathbf{\Delta}^{(T)} + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \\ &\leq \hat{\mu}_F e^{-\tilde{\alpha}_m \hat{\mu}_F T/2} \mathbf{\Delta}^{(0)} + \sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \\ &\leq \mathcal{O}[\mathcal{D}_F(\bar{w}^{(T)}, w^*)] \end{aligned}$$

where $\bar{w}^{(T)} := \frac{\sum_{t=0}^{T-1} \xi^{(t)}}{\mathcal{X}^{(T)}} w^{(t)}$, we use convexity of \mathcal{D}_F and F for the first inequality, the second one is by the reformulated inequality and the third one is by setting $\tilde{\alpha}_m T \geq \frac{2}{\hat{\mu}_F}$ and the fact $\frac{2\xi^{(T-1)}}{\tilde{\alpha}_m \hat{\mu}_F} \geq \mathcal{X}^{(T)} = \frac{2\xi^{(T-1)}[1 - (1 - \frac{\tilde{\alpha}_m}{2})^T]}{\tilde{\alpha}_m \hat{\mu}_F} \geq \frac{\xi^{(T-1)}}{\tilde{\alpha}_m \hat{\mu}_F}$ and $0 \leq (1 - \frac{\tilde{\alpha}_m}{2})^T \leq e^{-\frac{1}{2}\tilde{\alpha}_m \hat{\mu}_F T} \leq e^{-1} \leq \frac{1}{2}$.

To tighten this bound, we recommend [4], which discusses the range and strategy of step sizes in detail rather than our unified bound.

With $\tilde{\alpha}_m \geq \frac{2}{\hat{\mu}_F T}$, we have:

$$\sum_{j=1}^3 \dot{\delta}_j \tilde{\alpha}_m^{j-1} \leq \mathcal{O}(\dot{\delta}_1) + \mathcal{O}\left(\frac{\dot{\delta}_2}{T\hat{\mu}_F}\right) + \mathcal{O}\left(\frac{\dot{\delta}_3}{T^2\hat{\mu}_F^2}\right)$$

Thus,

$$\begin{aligned} \mathcal{O}[\mathcal{D}_F(\bar{w}^{(T)}, w^*)] &= \mathcal{O}(\hat{\mu}_F e^{-\tilde{\alpha}_m \hat{\mu}_F T/2} \mathbf{\Delta}^{(0)}) + \mathcal{O}\left(\frac{\dot{\delta}}{\hat{\mu}_F}\right) \\ &\quad + \mathcal{O}\left(\frac{(N/S-1)\sigma_{F,*}^2}{NT\hat{\mu}_F}\right) + \mathcal{O}\left(\frac{2^{R+4}e\dot{c}}{T^2\hat{\mu}_F^3\beta^2R^2}[2(1+2R)\sigma_{F,*}^2 + \dot{\delta}]\right) \end{aligned}$$

where, $\dot{\delta} = 4[\lambda \frac{\hat{L}_{g^*}}{\hat{\mu}_F}(\hat{u}_m + \eta\hat{\gamma}_\Phi)]^2(\frac{\hat{\gamma}_f^2}{|\hat{d}_i|} + \hat{\epsilon}^2) + 16[(1 + \sigma_\Phi)\hat{L}_\mathcal{E}(\hat{u}_m + \eta\hat{\gamma}_\Phi)\hat{\gamma}_\Phi]^2$ and $\dot{c} = 4[(1 + \sigma_\Phi)\hat{L}_\mathcal{E}(\hat{u}_m + \eta\hat{\gamma}_\Phi)]^2$.

For simplification, letting $A = [\frac{\hat{L}_{g^*}}{\hat{\mu}_F}(\hat{u}_m + \eta\hat{\gamma}_\Phi)]^2(\frac{\hat{\gamma}_f^2}{|\hat{d}_i|} + \hat{\epsilon}^2)$, $B = [(1 + \sigma_\Phi)\hat{L}_\mathcal{E}(\hat{u}_m + \eta\hat{\gamma}_\Phi)\hat{\gamma}_\Phi]^2$ and $C = \frac{\sigma_\Phi^2 \hat{L}_\mathcal{E}^2(\hat{u}_m + \eta\hat{\gamma}_\Phi)^2}{\hat{\mu}_F^3}$, we have:

$$\begin{aligned} \mathcal{O}[\mathcal{D}_F(\bar{w}^{(T)}, w^*)] &= \mathcal{O}(\hat{\mu}_F e^{-\tilde{\alpha}_m \hat{\mu}_F T/2} \mathbf{\Delta}^{(0)}) + \mathcal{O}\left(\frac{A\lambda^2 + B}{\hat{\mu}_F}\right) \\ &\quad + \mathcal{O}\left(\frac{(N/S-1)\sigma_{F,*}^2}{NT\hat{\mu}_F}\right) + \mathcal{O}\left(\frac{2^R C}{T^2\beta^2R^2}[R\sigma_{F,*}^2 + A\lambda^2 + B]\right). \end{aligned}$$

□

D.7.2 Proof of Theorem 2

The proof of Theorem 2 is shown as followings:

Proof. With Gaussian prior and first-order methods, we have the bound between personalized model and optimal global model, with $\dot{\delta}_p = \frac{2}{\hat{\mu}_{F_i,\cdot}^2}(\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \epsilon^2) + \frac{2}{\lambda^2}\epsilon_1^2 + \frac{4}{\lambda^2}\sigma_{F,*}^2 + \frac{1}{2}\eta^2\mathcal{G}_\Phi^2$, and $\dot{c}_p = (\frac{32}{\lambda^2}\hat{L}_F + \frac{8}{\hat{\mu}_F})$:

$$\begin{aligned}
\mathbf{E}||\tilde{\theta}_i(\bar{w}^T) - w^*||^2 &\leq 4[\mathbf{E}||\tilde{\theta}_i(\bar{w}^T) - \theta_i^*(\bar{w}^T)||^2 \\
&\quad + \mathbf{E}||\theta_i^*(\bar{w}^T) - \mu_i(\bar{w}^T)||^2 + \mathbf{E}||\mu_i(\bar{w}^T) - w^{(T)}||^2 + \mathbf{E}||w^{(T)} - w^*||^2] \\
&\leq 4[\frac{2}{\hat{\mu}_{F_i,\cdot}^2}(\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \epsilon^2) + \frac{1}{\lambda^2}\mathbf{E}2[||\nabla\tilde{F}_i(w^{(T)}) - \nabla F_i(w^{(T)})||^2 \\
&\quad + ||\nabla F_i(w^{(T)})||^2] + \frac{1}{2}\mathbf{E}||\eta\nabla\Phi_i(w^{(T)})||^2 + \mathbf{E}||w^{(T)} - w^*||^2] \\
&\leq 4[\frac{2}{\hat{\mu}_{F_i,\cdot}^2}(\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \epsilon^2) + \frac{2}{\lambda^2}\mathbf{E}\{\epsilon_1^2 + 2[||\nabla F_i(w^{(T)}) - \nabla F_i(w^*)||^2 + ||\nabla F_i(w^*)||^2]\} \\
&\quad + \frac{1}{2}\eta^2\mathcal{G}_\Phi^2 + \mathbf{E}||w^{(T)} - w^*||^2] \\
&\leq 4[\frac{2}{\hat{\mu}_{F_i,\cdot}^2}(\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \epsilon^2) + \frac{2}{\lambda^2}\epsilon_1^2 + \frac{8}{\lambda^2}\hat{L}_F\mathcal{D}_F(w^{(T)}, w^*) + \frac{4}{\lambda^2}\sigma_{F,*}^2 \\
&\quad + \frac{1}{2}\eta^2\mathcal{G}_\Phi^2 + \mathbf{E}||w^{(T)} - w^*||^2] \\
&\leq 4[\frac{2}{\hat{\mu}_{F_i,\cdot}^2}(\frac{\hat{\gamma}_f^2}{|\tilde{d}_i|} + \epsilon^2) + \frac{2}{\lambda^2}\epsilon_1^2 + \frac{4}{\lambda^2}\sigma_{F,*}^2 + \frac{1}{2}\eta^2\mathcal{G}_\Phi^2 + (\frac{8}{\lambda^2}\hat{L}_F + \frac{2}{\hat{\mu}_F})\mathcal{D}_F(w^{(T)}, w^*)] \\
&\leq \mathcal{O}(\dot{\delta}_p) + \mathcal{O}[\dot{c}_p\mathcal{D}_F(\bar{w}^{(T)}, w^*)]
\end{aligned}$$

where the first inequality is by Proposition 3, the second one is by Lemma 1 and Proposition 3, the third one is by Assumption 8 and Lemma 5, the fourth one is by Lemma 2 and Assumption 7 and the final one is by Theorem 1. □