Understanding Benign Overfitting in Personalized Federated Learning

Lisha Chen Tianyi Chen

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, United States

***Abstract*—In many federated learning problems, the model parameter is often shared across all the clients. However, when data distributions across clients are different, such a shared model may not have desired performance for all the clients. Therefore, personalization is critical to federated learning, especially when data over clients are non-i.i.d. Overparametrized models are typically trained using data on all devices and then fine-tuned to each device for personalization. While the conventional statistical learning theory suggests that overparameterized models overfit, empirical evidence reveals that overparameterized models for personalized federated learning still work well – a phenomenon called "benign overfitting." To better understand this phenomenon, we analyze the generalization error of a meta learning based personalized federated learning model in linear regression settings. Our theory explains the delicate interplay among data heterogeneity, model personalization, and benign overfitting in personalized federated learning.**

***Index Terms*—federated learning, personalization, meta learning, overparameterization**

I. Introduction

Federated learning (FL) is an emerging distributed learning paradigm where a machine learning model is trained across multiple clients holding their local data without exchanging them [1]. In many cases of FL, the model parameter is shared and thus common across all the clients. However, as data distributions across clients are usually heterogeneous, finding a common model for all the clients may not be desired [2]. Therefore, it is critical to develop personalized FL models for each client but still allow knowledge sharing among clients.

To achieve this goal, overparametrized models such as pretrained foundation models are often used as base models for clients to personalize [3]. However, training such base models is difficult in FL because each client only has limited local data and the number of training data is much smaller than the dimension of the model parameter. To address this issue, various approaches have been proposed to train the base models by leveraging the data of all clients.

In this paper, we study a specific variant of personalized federated learning models where the goal is to learn a shared initial global model that can quickly adapt to personalized models by taking several gradient descent steps, which is referred to as the model agnostic meta learning (MAML) [4], [5]. Previous works on MAML-based personalized FL mainly focus on analyzing the optimization and generalization error with sufficient training data [5]. Different from the previous works, we are particularly interested in the *generalization performance of the sought base model in practical scenarios where the total number of data from all clients is smaller than the dimension of the base model*. In those scenarios, the generalization error of the personalized yet overparameterized FL models is not fully understood.

Motivated by this, we ask:

Whether the overparameterized base models lead to overfitting in personalized FL?

In this paper, we take an initial step by answering this question in the meta linear regression setting.

A. Related works

Personalized FL has been actively studied via machine learning techniques such as transfer learning, meta learning, and representation learning; see e.g., [6]. We review related works that are grouped into the following categories.

a) **Personalized federated learning:** Personalized FL has received substantial attention recently. A simple baseline is to finetune the shared global model on local client data [7]. In [8], user-specific models are developed upon a multi-task learning framework. And in [9], a mixture of local and global models are used for personalized models. Another line of personalized FL methods is based on MAML which learns a global model that quickly adapts to local data with a few gradient steps [4], [5]. The empirical success of such methods inspired theoretical analysis to better understand how they work. One line of theoretical works study the generalization error or excess risk of meta learning methods in the linear centroid model [10], [11]. Generalization bound based on information theory [12] and the PAC-Bayes framework [13] has also been provided for meta learning. Recently, overparameterized meta learning has attracted attention. Bernacchia et al. [14] suggests that a negative learning rate in overparameterized MAML is optimal during training. Sun et al. [15] shows that overparameterized representation is optimal in representation based meta learning.

b) **Benign overfitting:** The empirical success of overparameterized deep neural networks inspired theoretical studies

The work was partially supported by and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons) and NSF SCALE-MoDL 2134168.

of benign overfitting. Bartlett et al. [16] analyze overparameterized linear regression model with the minimum norm solution and find that certain data covariance matrices lead to benign overfitting, explaining why overparameterized models which perfectly fit the noisy training data work well during testing. Later on, the analysis has been extended to ridge regression [17] and adversarial learning with linear models [18]. While previous theoretical efforts on benign overfitting have been largely focused on linear models, until recently, the analysis of benign overfitting has been extended to two-layer neural networks [19]–[21]. Existing works mainly study benign overfitting for single level empirical risk minimization problems applicable to conventional FL, rather than nested problems such as MAML, which is the focus of this work.

B. Our contributions

To our best knowledge, this is the first work to provide sufficient conditions for benign overfitting in MAML based personalized federated learning. Our contributions are summarized as follows.

- **C1)** We derive the upper bound of the excess risk for overparameterized meta learning based personalized FL and analyze the sufficient condition for benign overfitting.
- **C2)** We compare the benign overfitting condition for the overparameterized personalized FL models and that for the conventional FL, and show that overfitting is more likely to happen in MAML based personalized FL than in conventional FL with a shared model.
- **C3)** We show that data heterogeneity across clients will make overfitting more likely to happen compared with learning with a single client.

II. Preliminaries: MAML based personalized FL

In this section, we consider a particular personalized FL technique which is the MAML-based personalized FL [4], [5]. We first introduce the formulation and the algorithm.

Assume the data on the m-th client are drawn from $\mathcal{P}_m$, with input feature $\mathbf{x}_m \in \mathbb{R}^d$ and target label $y_m \in \mathbb{R}$; i.e., $(\mathbf{x}_m, y_m) \sim \mathcal{P}_m$. For each client m, we observe $2N$ i.i.d. samples collected in the dataset $\mathcal{D}_m = \{(\mathbf{x}_{m,n}, y_{m,n})\}_{n=1}^{2N}$, where $\mathcal{D}_m$ is divided into the train and validation datasets, denoted as $\mathcal{D}_m^t$ and $\mathcal{D}_m^v$, respectively. Without loss of generality, here $|\mathcal{D}_m^t| = |\mathcal{D}_m^v| = N$. Each client has its personalized model parameter $\boldsymbol{\theta}_m$. Given the data $\mathcal{D}_m$, we use the empirical loss $\ell_m(\boldsymbol{\theta}_m, \mathcal{D}_m^v)$ of per-client parameter $\boldsymbol{\theta}_m$ as a measure of the performance.

The goal of MAML based personalized FL is to learn an initial parameter $\boldsymbol{\theta}_0$, which can generate a per-client parameter $\boldsymbol{\theta}_m$ by taking one-step gradient descent with step size α on the training data $\mathcal{D}_m^t$, given by

$$\boldsymbol{\theta}_m(\boldsymbol{\theta}_0, \mathcal{D}_m^t) = \boldsymbol{\theta}_0 - \alpha \nabla_{\boldsymbol{\theta}_0} \ell_m(\boldsymbol{\theta}_0, \mathcal{D}_m^t). \tag{1}$$

With M clients and $\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^M$, the objective is to find $\boldsymbol{\theta}_0$ that minimizes the loss averaged over all clients, given by

$$\mathcal{L}(\boldsymbol{\theta}_0, \mathcal{D}) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \ell_m(\boldsymbol{\theta}_m(\boldsymbol{\theta}_0, \mathcal{D}_m^t), \mathcal{D}_m^v) \tag{2}$$

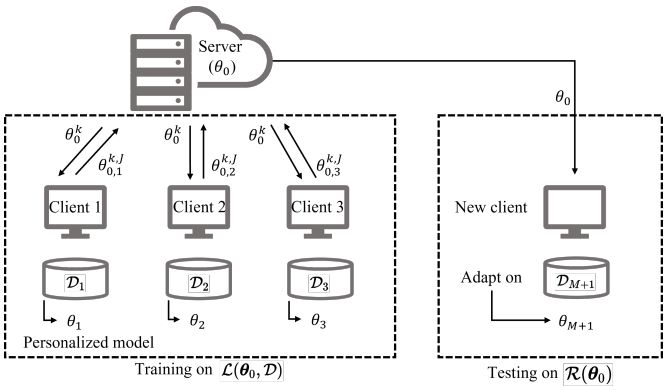

Fig. 1. Diagram of personalized FL.

where $\boldsymbol{\theta}_m(\boldsymbol{\theta}_0, \mathcal{D}_m^t)$ is obtained from the initial parameter $\boldsymbol{\theta}_0$ by taking one-step gradient descent.

In the training stage, we obtain $\hat{\boldsymbol{\theta}}_0$ by minimizing (2). And in the testing stage, we evaluate the test error of $\hat{\boldsymbol{\theta}}_0$ on

$$\mathcal{R}(\hat{\boldsymbol{\theta}}_0) \coloneqq \mathbb{E}_m\big[\mathbb{E}_{\mathcal{D}_m}\big[\ell_m(\boldsymbol{\theta}_m(\hat{\boldsymbol{\theta}}_0, \mathcal{D}_m^t), \mathcal{D}_m^v)\big]\big] \tag{3}$$

where the expectation is taken over the training and validation datasets at all clients and the client distributions.

Below we briefly describe the training phase of the MAML based personalized FL algorithm [4], [5]. Starting from the initial base model $\boldsymbol{\theta}_0^0$, the server randomly samples a subset of clients $\mathcal{M}_k$ at the k-th global iteration and sends the current $\boldsymbol{\theta}_0^k$ to these clients. In the local update, the local clients initially set $\boldsymbol{\theta}_{0,m}^{k,0} = \boldsymbol{\theta}_0^k$, then run J steps of local updates on $\boldsymbol{\theta}_{0,m}$ at the selected clients. For all clients $m \in \mathcal{M}_k$, at j-th iteration, $\boldsymbol{\theta}_{0,m}^{k,j-1}$ is locally updated by gradient descent over loss function $\mathcal{L}(\boldsymbol{\theta}_0, \mathcal{D}_m)$ with step size α_{out}, given below

$$\boldsymbol{\theta}_{0,m}^{k,j} = \boldsymbol{\theta}_{0,m}^{k,j-1} - \alpha_{\text{out}} \nabla \mathcal{L}(\boldsymbol{\theta}_{0,m}^{k,j-1}, \mathcal{D}_m). \tag{4}$$

After J local iterations, the m-th client sends $\boldsymbol{\theta}_{0,m}^{k,J}$ to the server and the server updates $\boldsymbol{\theta}_0^k$ by

$$\boldsymbol{\theta}_0^{k+1} = \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} \boldsymbol{\theta}_{0,m}^{k,J}. \tag{5}$$

The process described above is repeated until convergence. Fig. 1 shows the training and testing procedure for personalized FL. In the testing phase, as illustrated by Fig. 1, the trained base model $\hat{\boldsymbol{\theta}}_0$ will be used for new clients joining FL to personalize with one gradient descent step.

III. Main Results: Benign overfitting analysis

In this section, we introduce the data model and some necessary assumptions for the analysis. We present the main results, highlight the key steps of the proof and provide numerical evaluations to verify our results. Due to space limitations, we will defer all the derivations and the proofs of theorems in this section to the journal version.

A. Model and Assumptions

To make a precise analysis, we will assume the following linear data generation model in this version. Denoting the ground truth parameter on client m as $\boldsymbol{\theta}_m^{\text{gt}} \in \mathbb{R}^d$, and ϵ_m as the noise in the data, we assume the data generation model for client m is

$$y_m = \boldsymbol{\theta}_m^{\text{gt}\top} \mathbf{x}_m + \epsilon_m. \tag{6}$$

We make the following basic assumptions.

Assumption. *1. The total number of data is smaller than the dimension of the model parameter; i.e.* $2MN < d$.
2. Noise ϵ_m *is subgaussian with* $\mathbb{E}[\epsilon_m] = 0$ *and* $\mathbb{E}[\epsilon_m^2] = \sigma^2$.
3. Data $\mathbf{x}_m = \mathbf{V}_m \mathbf{\Lambda}_m^{\frac{1}{2}} \mathbf{z}_m$, *where* $\mathbf{z}_m$ *has independent,* σ_x*-subgaussian entries;* $\mathbb{E}[\mathbf{z}_m] = \mathbf{0}, \mathbb{E}[\mathbf{z}_m \mathbf{z}_m^\top] = \mathbf{I}_d$, *where* $\mathbf{I}_d$ *is a* $d \times d$ *identity matrix.*
4. Define $\mathbf{Q}_m \coloneqq \mathbb{E}[\mathbf{x}_m \mathbf{x}_m^\top]$, *which has bounded eigenvalues.*

Given the linear model (6), problem (2) generally has a unique solution when $d \leq 2MN$. However, by Assumption 1, the model is overparameterized. Therefore, problem (2) may have multiple solutions. Since recent advance in training overparameterized models reveal that gradient descent-based methods converge to the minimum norm solution [22], [23], we will analyze the test error of the minimum norm solution in this setting.

Definition 1 (Minimum norm solution)**.** *The minimum norm solution to the empirical personalized FL problem* (2) *with the linear regression loss is expressed by*

$$\begin{aligned} &\min_{\boldsymbol{\theta}_0} \quad \|\boldsymbol{\theta}_0\|^2 \\ &\text{s.t.} \sum_{m=1}^{M} \|\mathbf{X}_m^v \hat{\boldsymbol{\theta}}_m(\boldsymbol{\theta}_0) - \mathbf{y}_m^v\|^2 = \min_{\boldsymbol{\theta}} \sum_{m=1}^{M} \|\mathbf{X}_m^v \hat{\boldsymbol{\theta}}_m(\boldsymbol{\theta}) - \mathbf{y}_m^v\|^2 \\ &\hat{\boldsymbol{\theta}}_m(\boldsymbol{\theta}, \mathcal{D}_m^t) = (\mathbf{I} - \alpha \hat{\mathbf{Q}}_m^t)\boldsymbol{\theta} + \frac{\alpha}{N} \mathbf{X}_m^{t\top} \mathbf{y}_m^t, \quad \forall m \end{aligned} \tag{7}$$

where $\mathbf{X}_m^t = [\mathbf{x}_{m,1}^t, \ldots, \mathbf{x}_{m,N}^t]^\top \in \mathbb{R}^{N \times d}$, $\mathbf{y}_m^t = [y_{m,1}^t, \ldots, y_{m,N}^t]^\top \in \mathbb{R}^N$. *Superscript 't' represents training, which can be replaced by 'v' for validation and 'a' for all.*

With the linear data model (6), the empirical loss, test error, along with their optimal solutions can be computed analytically with closed-form which we summarize in Proposition 1.

Proposition 1. (Empirical and population level solutions) *Under data model* (6)*, the test error of MAML based personalized FL with parameter* $\boldsymbol{\theta}_0$ *can be computed by*

$$\mathcal{R}(\boldsymbol{\theta}_0) = \mathbb{E}_m\big[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_m^{\text{gt}}\|_{\mathbf{W}_m}^2\big]. \tag{8}$$

The optimal solution to the test error in (8) *is given by*

$$\boldsymbol{\theta}_0^* \coloneqq \arg\min_{\boldsymbol{\theta}_0} \mathcal{R}(\boldsymbol{\theta}_0) = \mathbb{E}_m\big[\mathbf{W}_m\big]^{-1} \mathbb{E}_m\big[\mathbf{W}_m \boldsymbol{\theta}_m^{\text{gt}}\big] \tag{9}$$

where $\mathbf{W}_m$ *is defined as*

$$\mathbf{W}_m = (\mathbf{I} - \alpha \mathbf{Q}_m)\mathbf{Q}_m(\mathbf{I} - \alpha \mathbf{Q}_m). \tag{10}$$

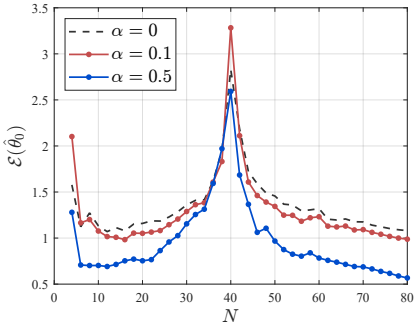

Fig. 2. Excess risk v.s. number of training samples (N) with different α. We fix $M = 10, d = 200$.

Denote $\hat{\mathbf{Q}}_m^a \coloneqq \frac{1}{2N} \mathbf{X}_m^{a\top} \mathbf{X}_m^a$, *and the weight matrix* $\hat{\mathbf{W}}_m$ *as*

$$\hat{\mathbf{W}}_m = (\mathbf{I} - \alpha \hat{\mathbf{Q}}_m^t)\hat{\mathbf{Q}}_m^v(\mathbf{I} - \alpha \hat{\mathbf{Q}}_m^t). \tag{11}$$

The minimum-norm solution obtained from (7) *is given by*

$$\hat{\boldsymbol{\theta}}_0 = \Big(\sum_{m=1}^{M} \hat{\mathbf{W}}_m\Big)^\dagger \Big(\sum_{m=1}^{M} \hat{\mathbf{W}}_m \boldsymbol{\theta}_m^{\text{gt}}\Big) + \Delta_M \tag{12}$$

where † *denotes the Moore–Penrose pseudo inverse, and the error term* Δ_M *is a polynomial function of* M, N, d.

To study overfitting in the personalized FL model, we need to measure its generalization ability. A widely used metric to quantify the generalization ability of a model is the excess risk. The excess risk of a solution $\hat{\boldsymbol{\theta}}_0$ is defined as

$$\mathcal{E}(\hat{\boldsymbol{\theta}}_0) \coloneqq \mathcal{R}(\hat{\boldsymbol{\theta}}_0) - \mathcal{R}(\boldsymbol{\theta}_0^*). \tag{13}$$

From (13), we can see that excess risk measures the difference between the test error of the empirical solution from finite samples, $\hat{\boldsymbol{\theta}}_0$ and the optimal test error. The larger the excess risk, the further the empirical solution $\hat{\boldsymbol{\theta}}_0$ is from the optimal population solution $\boldsymbol{\theta}_0^*$, indicating more severe *overfitting*.

Next, we use the definition in (13) and the solutions in Proposition 1 to compute the excess risk analytically and analyze its upper bound.

B. Main results

With the closed-form solutions given in Proposition 1, we are ready to bound the excess risk of personalized FL.

Theorem 1 (Personalized FL excess risk bound)**.** *Suppose Assumptions 1-4 hold. Let* $\mu_1(\cdot) \geq \mu_2(\cdot) \ldots$ *denote the eigenvalues of a matrix in the descending order. Define* $\overline{\mathbf{W}}_M \coloneqq \frac{1}{M} \sum_{m=1}^{M} \mathbf{W}_m$. *For meta linear regression problem with the minimum norm solution in* (7)*, define the effective ranks as*

$$r_k\left(\overline{\mathbf{W}}_M\right) \coloneqq \frac{\sum_{i>k} \mu_i\left(\overline{\mathbf{W}}_M\right)}{\mu_{k+1}\left(\overline{\mathbf{W}}_M\right)};\ R_k\left(\overline{\mathbf{W}}_M\right) \coloneqq \frac{\left(\sum_{i>k} \mu_i(\overline{\mathbf{W}}_M)\right)^2}{\sum_{i>k} \mu_i^2(\mathbf{W})}. \tag{14}$$

Define the cross-client data heterogeneity in terms of the difference of eigenvalues of $\mathbf{Q}_m$, *given by*

$$\mathbb{V}(\{\mathbf{\Lambda}_m\}_{m=1}^M) \coloneqq \Big|\max_{i,m} \lambda_i - \lambda_{m,i}\Big|.$$

For a universal constant b, if the effective dimension $k^* = \min\{k \geq 0 : r_k(\overline{\mathbf{W}}_M) \geq bMN\}$, then with high probability, the excess risk satisfies

$$\mathcal{E}(\hat{\boldsymbol{\theta}}_0) \leq \xi + \sigma^2 c_1 \Big(\frac{k^*}{2MN} + \frac{2MN}{R_{k^*}(\overline{\mathbf{W}}_M)} \Big) (\mathbb{V}(\{\boldsymbol{\Lambda}_m\}_{m=1}^M) + 1) \tag{15}$$

where c_1 is a constant, ξ depends on $\boldsymbol{\theta}_m^{\mathrm{gt}}, \mathbf{Q}_m, M, N$, and is not increasing with M, N.

In the case of overparameterized models, the "benign overfitting" refers to the situation where the variance in the excess risk will still vanish when M and N increase. From Theorem 1, it can be further shown that the excess risk depends on both the eigenvalues of the data covariance matrix $\mathbf{Q}_m$, and the cross-client data heterogeneity, measured by $\mathbb{V}(\{\boldsymbol{\Lambda}_m\}_{m=1}^M)$. Fig. 2 plots the test error versus the number of the training data. A "double descent" curve is formed in Fig. 2. It shows that as N increases, $\mathcal{E}(\hat{\boldsymbol{\theta}}_0)$ first gets better, then worse and then better; see the non-FL setting [24]. The trend in Fig. 2 is similar to the trend observed in [25]. When $d/(2MN) > 1$, the model is overparameterized, which can overfit the training data, leading to larger excess risk as N decreases. However, Fig. 2 shows the excess risk does not become too large as N decreases, indicating that overfitting does not severely harm the test error in this case.

We say the data matrix $\mathbf{Q}_m$ satisfies the *benign overfitting condition*, if for all m,

$$\lim_{MN\to\infty} \frac{k^*}{MN} = \lim_{MN\to\infty} \frac{MN}{R_{k^*}(\overline{\mathbf{W}}_M)} = 0. \tag{16}$$

This guarantees the variance term in the excess risk (15) goes to zero with sufficient training data from all clients.

To compare benign overfitting in personalized FL with that in conventional FL, we can set the step size $\alpha = 0$ in problem (7), which reduces to conventional FL without personalization. Compared to Theorem 1, the benign overfitting condition in (16) is less restrictive since it does not impose constraints on α. Intuitively, benign overfitting is more likely to happen in personalized FL than in conventional FL.

An example to better understand what kind of data matrix satisfies the benign overfitting condition is given below.

Example 1 (Data covariance)**.** *Suppose $\mathbf{Q}_m = \mathrm{diag}(\mathbf{I}_{d_1}, \beta \mathbf{I}_{d-d_1})$ for all m. Set $M = 10, d = 200, d_1 = 20, \alpha = 0.01$. Then it satisfies the benign overfitting condition for personalized FL. We plot the test error with different choice of β in Fig. 3.*

From Fig. 3 we can observe that given a fixed number of training data N, the test error increases with β. This observation verifies our theory since larger β results in a smaller $R_k^*(\overline{\mathbf{W}}_M)$, leading to a larger value of the upper bound on the variance term in (15).

Example 1 demonstrates how the per-client data matrix $\mathbf{Q}_m$ affects the excess risk. We consider another example that demonstrates how the data heterogeneity across clients affects the excess risk.

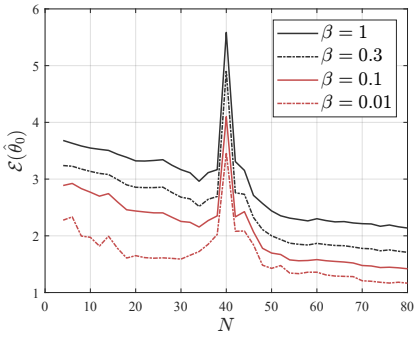

Fig. 3. Excess risk v.s. number of training samples (N) for $\mathbf{Q}_m = \mathrm{diag}(\mathbf{I}_{d_1}, \beta \mathbf{I}_{d-d_1})$ with different β.

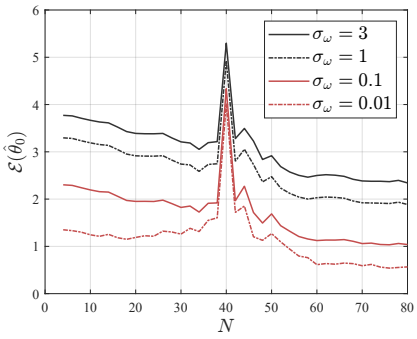

Fig. 4. Excess risk v.s. number of training samples (N) for $\mathbf{Q}_m = |\omega_m + 1| \,\mathrm{diag}(\mathbf{I}_{d_1}, \beta \mathbf{I}_{d-d_1}), \omega_m \sim \mathcal{N}(0, \sigma_\omega^2)$ with different σ_ω.

Example 2 (Data heterogeneity)**.** *Let $\omega_m \sim \mathcal{N}(0, \sigma_\omega^2)$. Suppose $\mathbf{Q}_m = |\omega_m + 1| \,\mathrm{diag}(\mathbf{I}_{d_1}, \beta \mathbf{I}_{d-d_1})$ for all m. Set $M = 10, d = 200, d_1 = 20, \alpha = 0.01, \beta = 0.3$. Then it satisfies the benign overfitting condition for personalized FL. Fig. 4 plots the test error with different choices of σ_ω.*

Observe from Fig. 4 that the larger σ_ω^2, the higher the test error, and the more difficult for the benign overfitting condition to be satisfied. Therefore, compared to FL with a single client, the benign overfitting condition for FL is more restrictive as it imposes constraints for both the expected data covariance matrix $\mathbf{Q}_m$, and the data heterogeneity $\mathbb{V}(\{\boldsymbol{\Lambda}_m\}_{m=1}^M)$.

C. Sketch of proof

In this section, we highlight the key steps of the proof for Theorem 1. The first step is to decompose the excess risk of MAML based personalized FL defined in (13) into bias and variance, summarized in Lemma 1.

Lemma 1. *With probability at least $1-\delta$ over ϵ, the excess risk of the minimum norm solution of personalized FL is bounded above by*

$$\mathcal{E}(\hat{\boldsymbol{\theta}}_0) \leq \underbrace{2 \Big\| \Big(\sum_m \hat{\mathbf{W}}_m \Big)^\dagger \Big(\sum_m \hat{\mathbf{W}}_m (\boldsymbol{\theta}_m^{\mathrm{gt}} - \boldsymbol{\theta}_0^*) \Big) \Big\|_{\mathbf{W}}^2}_{\textit{per-client parameter distance}} + \underbrace{2 \boldsymbol{\theta}_0^{*\top} \mathbf{B} \boldsymbol{\theta}_0^*}_{\textit{bias}} + \underbrace{2 c_1 \sigma^2 \log \frac{1}{\delta} \mathrm{tr}(\mathbf{C})}_{\textit{variance}} \tag{17}$$

where $\mathbf{W} \coloneqq \mathbb{E}_m[\mathbf{W}_m]$, and $\mathbf{B}, \mathbf{C}$ are matrices depending on $\mathbf{X}_m, \mathbf{W}_m$.

The first term on the right hand side of (17) is the average weighted distance between the optimizer of the test error and the ground truth parameter on client m. The second term is the bias of the minimum norm solution in overparameterized personalized FL. The first two terms can be bounded based on the concentration inequalities on subgaussian random variables, given in Lemmas 2 and 3 below.

Lemma 2 (Bound on per-client parameter distance)**.** *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\Big\| \Big(\sum_m \hat{\mathbf{W}}_m \Big)^\dagger \Big(\sum_m \hat{\mathbf{W}}_m (\boldsymbol{\theta}_m^{\rm gt} - \boldsymbol{\theta}_0^*) \Big) \Big\|_{\mathbf{W}}^2 \leq \widetilde{O}\Big(\frac{1}{M}\Big)$$

where $\widetilde{O}(\cdot)$ hides the log polynomial dependence on N, M, d, δ.

Lemma 3 (Bound on $\boldsymbol{\theta}_0^{*\top} \mathbf{B} \boldsymbol{\theta}_0^*$)**.** *There is a constant c_2 that depends only on σ_x, such that for any $1 < t < MN$, with probability at least $1 - e^{-t}$,*

$$\boldsymbol{\theta}_0^{*\top} \mathbf{B} \boldsymbol{\theta}_0^* \leq c_2 \|\boldsymbol{\theta}_0^*\|^2 \|\mathbf{W}\| \max \Big\{ \sqrt{\frac{r_0(\mathbf{W})}{MN}}, \frac{r_0(\mathbf{W})}{MN}, \sqrt{\frac{t}{MN}} \Big\}.$$

These two terms correspond to ξ in (15) of Theorem 1, which does not go to infinity as M, N, d increase. And the dominating term in the excess risk is the last variance term. The upper bound of $\mathrm{tr}(\mathbf{C})$ in the dominating variance term in (17) is given below.

Lemma 4 (Bound on $\mathrm{tr}(\mathbf{C})$)**.** *There are constants b, c_1 such that for $0 \leq k \leq 2MN/c_1$, $r_k(\overline{\mathbf{W}}_M) \geq bMN$, and $l \leq k$, with probability at least $1 - 7e^{-2MN/c_1}$, it follows*

$$\mathrm{tr}(\mathbf{C}) \leq c_1 \Big(\frac{l}{2MN} + \frac{2MN}{R_l(\overline{\mathbf{W}}_M)} \Big) (\mathbb{V}(\{\boldsymbol{\Lambda}_m\}_{m=1}^M) + 1).$$

Plugging the results of Lemmas 2, 3 and 4 into (17), we reach the results in Theorem 1.

IV. Conclusions

In this paper, we study overparameterized meta learning based personalized federated learning. For a precise analysis, we focus on linear regression where the total number of data from all clients is smaller than the dimension of the model parameter. We show that when the data heterogeneity across clients is small, the per-client data covariance matrices with certain properties lead to benign overfitting for MAML based personalized federated learning with minimum norm solution. This explains why overparameterized personalized federated learning models can generalize well in new data and new clients. Furthermore, our theory shows that overfitting is more likely to happen in MAML based personalized federated learning than in conventional federated learning with a single shared model. In addition, data heterogeneity across clients makes overfitting more likely in personalized federated learning.

References

[1] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[2] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. International Conference on Machine Learning*, virtual, 2021, pp. 6357–6368.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[4] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.

[5] A. Fallah, A. Mokhtari, and A. E. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Advances in Neural Information Processing Systems*, virtual, 2020.

[6] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[7] G. Cheng, K. Chadha, and J. Duchi, "Fine-tuning in federated learning: a simple but tough-to-beat baseline," *arXiv preprint arXiv:2108.07313*, 2021.

[8] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, Dec 2017, pp. 4427–4437.

[9] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive personalized federated learning," *arXiv preprint arXiv:2003.13461*, 2020.

[10] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil, "Learning to learn around a common mean," in *Proc. Advances in Neural Information Processing Systems*, vol. 31, Montreal, Canada, 2018.

[11] L. Chen and T. Chen, "Is Bayesian model agnostic meta learning better than model agnostic meta learning, provably?" in *Proc. International Conference on Artificial Intelligence and Statistics*, virtual, 2022.

[12] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, vol. 23, no. 1, p. 126, 2021.

[13] J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause, "Pacoh: Bayes-optimal meta-learning with pac-guarantees," in *Proc. International Conference on Machine Learning*, virtual, 2021, pp. 9116–9126.

[14] A. Bernacchia, "Meta-learning with negative learning rates," in *Proc. International Conference on Learning Representations*, virtual, 2020.

[15] Y. Sun, A. Narang, H. I. Gulluk, S. Oymak, and M. Fazel, "Towards sample-efficient overparameterized meta-learning," in *Proc. Advances in Neural Information Processing Systems*, virtual, 2021.

[16] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[17] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *arXiv preprint arXiv:2009.14286*, 2020.

[18] J. Chen, Y. Cao, and Q. Gu, "Benign overfitting in adversarially robust linear classification," *arXiv preprint arXiv:2112.15250*, 2021.

[19] Z. Li, Z.-H. Zhou, and A. Gretton, "Towards an understanding of benign overfitting in neural networks," *arXiv preprint arXiv:2106.03212*, 2021.

[20] Y. Cao, Z. Chen, M. Belkin, and Q. Gu, "Benign overfitting in two-layer convolutional neural networks," *arXiv preprint arXiv:2202.06526*, 2022.

[21] S. Frei, N. S. Chatterji, and P. L. Bartlett, "Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data," *arXiv preprint arXiv:2202.05928*, 2022.

[22] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *Proc. International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 1832–1841.

[23] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[24] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

[25] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, "Optimal regularization can mitigate double descent," in *Proc. International Conference on Learning Representations*, virtual, 2020.