Federated Asymptotics: a model to compare federated learning algorithms

Gary Cheng*† chenggar@stanford.edu
Karan Chadha*† knchadha@stanford.edu
John Duchi†‡ jduchi@stanford.edu

February 21, 2022

Abstract

We propose an asymptotic framework to analyze the performance of (personalized) federated learning algorithms. In this new framework, we formulate federated learning as a multi-criterion objective, where the goal is to minimize each client's loss using information from all of the clients. We analyze a linear regression model where, for a given client, we may theoretically compare the performance of various algorithms in the high-dimensional asymptotic limit. This asymptotic multi-criterion approach naturally models the high-dimensional, many-device nature of federated learning. These tools make fairly precise predictions about the benefits of personalization and information sharing in federated scenarios—at least in our (stylized) model—including that Federated Averaging with simple client fine-tuning achieves the same asymptotic risk as the more intricate meta-learning and proximal-regularized approaches and outperforming Federated Averaging without personalization. We evaluate these predictions on federated versions of the EMNIST, CIFAR-100, Shakespeare, and Stack Overflow datasets, where the experiments corroborate the theoretical predictions, suggesting such frameworks may provide a useful guide to practical algorithmic development.

1 Introduction

In Federated learning (FL), a collection of client machines, or devices, collect data and coordinate with a central server to fit machine-learned models, where communication and availability constraints add challenges [KMA+19]. A natural formulation here, assuming a supervised learning setting, is to assume that among m distinct clients, each client i has distribution P_i, draws observations $Z \sim P_i$, and wishes to fit a model—which we represent abstractly as a parameter vector $\theta \in \Theta$—to minimize a risk, or expected loss, $L_i(\theta) := \mathbb{E}_{P_i}[\ell(\theta; Z)]$, where the loss $\ell(\theta; z)$ measures the performance of θ on example z. Thus, at the most abstract level, the federated learning problem is to solve the multi-criterion problem

$$\underset{\theta_1,\ldots,\theta_m}{\text{minimize}} \left(L_1(\theta_1), \ldots, L_m(\theta_m)\right). \tag{1}$$

At this level, problem (1) is both trivial—one should simply minimize each risk L_i individually—and impossible, as no individual machine has enough data locally to effectively minimize L_i. Consequently, methods in federated learning typically take various departures from the multicriterion objective (1) to provide more tractable problems. Many approaches build off of the empirical risk minimization principal [Vap92, Vap95, HTF09], where we seek a single parameter θ that does well across all machines and data, minimizing a (weighted) average loss

$$\sum_{i=1}^{m} p_i L_i(\theta) \tag{2}$$

*Equal contribution, author order random
†Electrical Engineering Department, Stanford University
‡Statistics Department, Stanford University

over $\theta \in \Theta$, where $p \in \mathbb{R}_+^m$ satisfies $p^T \mathbf{1} = 1$. This "zero personalization" approach has the advantage that data is (relatively) plentiful, and has led to a substantial literature. Much of this work focuses on developing efficient methods that limit possibly expensive and unreliable communication between centralized servers and distributed devices [HM19, RCZ+21, MMR+17, KKM+20, MSS19, LSZ+20]. Given (i) the challenges of engineering such large-scale systems, (ii) the success of large-scale machine learning models without personalization, and (iii) the plausibility that individual devices have similar distributions P_i, the zero personalization approach is natural. However, as distributions across individual devices are typically non-identical, it is of interest to develop methods more closely targeting problem (1).

One natural assumption to make is that the optimal client parameters are "close" to one another and thus must be "close" to the minimizer of the zero-personalization objective (2). Approaches leveraging this assumption [DTN20, SCST17, WMK+19, FMO20] regularize client parameters towards the global parameter. While these methods are intuitive, most focus on showing convergence rates to local minima. While convergence is important, these analyses do little to characterize the performance of solutions attained—to what the methods actually converge. These issues motivate our paper.

Contributions:

1. **New model** (Sec. 2): We propose and analyze a (stylized) high-dimensional linear regression model, where, for a given client, we can characterize the performance of collaborate-then-personalize algorithms in the high-dimensional asymptotic limit.

2. **Precise risk characterization**: We use our stylized model to evaluate the asymptotic test loss of several procedures. These include simple fine-tuned variants of Federated Averaging [MMR+17], where one learns an average global model (2) and updates once using local data; meta-learning variants of federated learning; and proximal-regularized personalization in federated learning.

3. **Precise predictions and experiments**: Our theory makes several percise predictions, including that fairly naive methods—fine-tuning variants—should perform as well as more sophisticated methods, as well as conditions under which federated methods improve upon zero-personalization (2) or zero-collaboration methods. To test these predicted behaviors, we perform several experiments on federated versions of the EMNIST, CIFAR-100, Shakespeare, and Stack Overflow datasets. Perhaps surprisingly, the experiments are quite consistent with the behavior the theory predicts.

Our choice to study linear models in the high-dimensional asymptotic setting (when dimension and samples scale proportionally) takes as motivation a growing phenomenological approach to research in machine learning, where one develops simple models that predict (perhaps unexpected) complex behavior in model-fitting. Such an approach has advantages: by developing simpler models, one can isolate causative aspects of behavior and make precise predictions of performance, leveraging these to provide insights in more complex models. Consider, for instance, [HMRT19], who show that the "double-descent" phenomenon, where (neural-network) models show decreasing test loss as model size grows, exists even in linear regression. In a robust (adversarial) learning setting, [CRS+19] use a two-class Gaussian linear discriminant model to suggest ways that self-supervised training can circumvent hardness results, using the predictions (on the simplified model) to inform a full deep training pipeline substantially outperforming (then) state-of-the-art. [Fel19] develops clustering models where memorization of data is necessary for good learning performance, suggesting new models for understanding generalization. We view our contributions in this intellectual tradition: using a stylized high-dimensional asymptotics to develop statistical insights underpinning Federated Learning (FL). This allows direct comparison between different FL methods—not between upper bounds, but actual losses—and serving as a basic framework to motivate new methodologies in Federated Learning.

Related Work. The tried-and-true method to adapt to new data distributions is fine-tuning [HR18]. In Federated Learning (FL), this broadly corresponds to fine-tuning a global model (e.g., from FedAvg) on a user's local data [WMK+19, YBS20, LHBS21]. While fine-tuning's simplicity and practical efficacy recommend it, we know of little theoretical analysis.

A major direction in FL is to design personalization-incentivizing objectives. [SCST17], for example, build out of the literature on multitask and transfer learning [Car97, PY09] to formulate a multi-task learning objective, treating each machine as an individual task; this and other papers [FMO20, MMRS20, DTN20] show rates of convergence for optimization methods on these surrogates. These methods use the heuristic that personalized, local models should lie "close" to one another, and the authors provide empirical evidence for their improved performance. Yet it is not always clear what conditions are necessary (or sufficient) for these specialized personalization methods to outperform naive zero collaboration—fully local training on available data on each individual device—and zero personalization (averaged) methods. In a related vein, meta-learning approaches [FAL17, FMO20, JKRK19] seek a global model that can quickly "adapt" to local distributions P_i, typically by using a few gradient steps. This generally yields a sophisticated non-convex objective, making it hard to give even heuristic guarantees, leading authors instead to emphasize worst-case convergence rates of iterative algorithms to (near) stationary points.

Other methods of personalization have also been proposed. In contrast to using a global model to help train the local model, [MMRS20] and [ZMM+20] use a mixture of global and local models to incorporate personalized features. [CZLS21], like we do, propose evaluating federated algorithms via the formulation (1); they give minimax bounds to distinguish situations in which zero collaboration and zero personalization (averaged) methods (2) are, respectively, worst-case optimal.

2 The Linear Model

We consider a high-dimensional asymptotic model, where clients solve statistically related linear regression problems, and each client $i \in [m]$ has a local dataset size n_i smaller than (but comparable to) the dimension d of the problem. This choice models the empirical fact that the data on a single client is typically small relative to model dimension (e.g., even training the last layer of a deep neural network).

More concretely, each client $i \in [m]$ will use an overparameterized linear regression problem to recover an unknown parameter $\theta_i^\star \in \mathbb{R}^d$. Client i has n_i i.i.d. observations $(\mathbf{x}_{i,k}, y_{i,k}) \in \mathbb{R}^d \times \mathbb{R}$,

$$y_{i,k} = \mathbf{x}_{i,k}^T \theta_i^\star + \xi_{i,k}, \qquad \mathbf{x}_{i,k} \overset{\text{iid}}{\sim} P_{\mathbf{x}}^i \;\text{ and }\; \xi_{i,k} \overset{\text{iid}}{\sim} P_\xi^i.$$

We make the routine assumption that the features are centered, with $\mathbb{E}[\mathbf{x}_{i,k}] = 0$ and $\mathrm{Cov}(\mathbf{x}_{i,k}) = \Sigma_i$. We also assume that the noise is centered with finite variance, i.e., $\mathbb{E}[\xi_{i,k}] = 0$ and $\mathrm{Var}(\xi_{i,k}) = \sigma_i^2$. For convenience, we let $X_i \in \mathbb{R}^{n_i \times d}$ and $\mathbf{y}_i \in \mathbb{R}^{n_i}$ denote client i's data, and $X \coloneqq [X_1^T, \dots, X_m^T]^T$. We also let $N \coloneqq \sum_{j=1}^m n_j$.

A prior P_θ^i on the parameter $\theta_i^\star$ relates tasks on each client, where conditional on $\theta_0^\star$, $\theta_i^\star$ is supported on $r_i \mathbb{S}^{d-1} + \theta_0^\star$—the sphere of radius r_i (bounded by a constant for all $i \in [m]$) centered at $\theta_0^\star$—with $\mathbb{E}[\theta_i^\star] = \theta_0^\star$. The variation between clients is captured by differences in r_i (label shift) and Σ_i (covariate shift), while the similarity is captured by the shared center $\theta_0^\star$. Intuitively, data from client j is useful to client i as it provides information on the possible location of $\theta_0^\star$. Lastly, we assume that the distributions of $\mathbf{x}, \theta^\star$, and ξ are independent of each other and across clients.

Every client i seeks to minimize its local population loss—the squared prediction error of a new sample $\mathbf{x}_{i,0}$ independent of the training set—conditioned on X. The sample loss is $\ell(\theta; (\mathbf{x}, y)) = (\mathbf{x}^T\theta - y)^2 - \sigma_i^2$, giving per client test loss

$$\begin{aligned} L_i(\hat{\theta}_i \mid X) &\coloneqq \mathbb{E}[(\mathbf{x}_{i,0}^T \hat{\theta}_i - \mathbf{x}_{i,0}^T \theta_i^\star)^2 \mid X] \\ &= \mathbb{E}\Big[\big\| \hat{\theta}_i - \theta_i^\star \big\|_{\Sigma_i}^2 \mid X \Big] \end{aligned}$$

where the expectation is taken over $(\mathbf{x}_{i,0}, \theta_i^\star, \xi_i) \sim P_{\mathbf{x}}^i \times P_{\theta^\star}^i \times P_\xi^i$ and $\|x\|_\Sigma^2 = x^T \Sigma x$. It is essential here that we focus on per client performance: the ultimate goal is to improve performance on any given client, as per eq. (1). For analysis purposes, we often consider the equivalent bias-variance decomposition

$$L_i(\hat{\theta}_i | X) = \underbrace{\Big\| \mathbb{E}[\hat{\theta}_i | X] - \theta^\star \Big\|_{\Sigma_i}^2}_{B_i(\hat{\theta}_i | X)} + \underbrace{\mathrm{tr}(\mathrm{Cov}(\hat{\theta}_i | X) \Sigma_i)}_{V_i(\hat{\theta}_i | X)}. \tag{3}$$

Our main asymptotic assumption, which captures the high-dimensional and many-device nature central to modern federated learning problems, follows:

Assumption A1. *As $m \to \infty$, both $d = d(m) \to \infty$ and $n_j = n_j(m) \to \infty$ for $j \in [m]$, and $\lim_m \frac{d}{n_j} = \gamma_j$. Moreover, $1 < \gamma_{\min} \le \lim_m \inf_{j\in[m]} \frac{d}{n_j} \le \lim_m \sup_{j\in[m]} \frac{d}{n_j} \le \gamma_{\max} < \infty$.*

Importantly, individual devices are overparameterized: we always have $\gamma_j > 1$, as is common, when the dimension of models is large relative to local sample sizes, but may be smaller than the (full) sample. Intuitively, γ_j captures the degree of overparameterization of the network for user j. We also require control of the eigenspectrum of our data [cf. HMRT19, Assumption 1].

Definition 2.1. *The* empirical distribution of the eigenvalues *of Σ is the function $\mu(\cdot;\Sigma) : \mathbb{R} \to \mathbb{R}_+$ with*

$$\mu(s;\Sigma) := \frac{1}{d}\sum_{j=1}^{d} \mathbf{1}\left\{s \ge s_i\right\}, \tag{4}$$

where $s_1 \ge s_2 \ge \cdots \ge s_d$ are the eigenvalues of Σ.

Assumption A2. *For each user i, data $\mathbf{x} \sim P_{\mathbf{x}}^i$ have the form $\mathbf{x} = \Sigma_i^{\frac{1}{2}}\mathbf{z}$. For some $q > 2$, $\kappa_q < \infty$, and $M < \infty$,*

(a) The vector $\mathbf{z} \in \mathbb{R}^d$ has independent entries with $\mathbb{E}[z_i] = 0$, $\mathbb{E}[z_i^2] = 1$, and $\mathbb{E}[|z_i|^{2q}] \le \kappa_q < \infty$

(b) $s_1 = \left\|\Sigma_i\right\|_{\text{op}} \le M$, $s_d = \lambda_{\min}(\Sigma_i) \ge 1/M$, and $\int s^{-1} d\mu(s;\Sigma_i) < M$.

(c) $\mu(\cdot;\Sigma_i)$ converges weakly to ν_i

These conditions are standard, guaranteeing sufficient moments for convergence of covariance estimates, that the eigenvalues of Σ_i do not accumulate near 0, and a mode of convergence for the spectrum of Σ_i.

3 Locally fine-tuning a global solution

In this section, we describe and analyze fine-tuning algorithms that use the FedAvg solution (a minimizer of the objective (2)) as a warm start to find personalized models. We compare the test loss of these algorithms with naive, zero personalization and zero collaboration approaches. Among other things, we show that a ridge-regularized locally fine-tuned method outperforms the other methods.

3.1 Fine-tuned Federted Averaging (FTFA)

Fine-tuned Federated Averaging (FTFA) approximates minimizing the multi-criterion loss (1) using a two-step procedure in Algorithm 1 (see Section 7 for detailed pseudocode). Let $\mathcal{S}_i$ denote client i's sample. The idea is to replace the local expected risks L_i in (2) with the local empirical risks

$$\widehat{L}_i(\theta) := \frac{1}{n_i}\sum_{z\in\mathcal{S}_i} \ell(\theta; z),$$

using the FedAvg solution $\hat{\theta}_0^{FA}$ as a warm-start for local training in the second step. Intuitively, FTFA interpolates between zero collaboration and zero personalization algorithms. Each client i can run this local training phase independently of (and in parallel with) all others, as the data is fully local; this separation makes FTFA essentially no more expensive than Federated Averaging.

For the linear model in Section 2, FTFA first minimizes the average weighted loss $\sum_{j=1}^m p_j \frac{1}{2n_j}\|X_j\theta - \mathbf{y}_j\|_2^2$ over all clients. As the local linear regression problem to minimize $\|X_i\theta - \mathbf{y}_i\|_2^2$ is overparameterized, first-order methods on it (including the stochastic gradient method) correspond to solving minimum

Algorithm 1 FTFA & RTFA (details in appendix)

1. The server coordinates (e.g. using FedAvg) to find a global model using data from all clients, solving

$$\hat{\theta}_0^{FA} = \operatorname*{argmin}_{\theta} \sum_{j=1}^{m} p_j \widehat{L}_j(\theta), \tag{5}$$

where $p \in \mathbb{R}_+^m$ are weights satisfying $\sum_{j=1}^m p_j = 1$. The server broadcasts $\hat{\theta}_0^{FA}$ to all clients.

2.a. **FTFA:** Client i optimizes its risk $\widehat{L}_i$ using a first-order method initialized at $\hat{\theta}_0^{FA}$, returning model $\hat{\theta}_i^{FA}$.

2.b. **RTFA:** Client i minimizes a regularized empirical risk to return model

$$\hat{\theta}_i^R(\lambda) = \operatorname*{argmin}_{\theta} \widehat{L}_i(\theta) + \frac{\lambda}{2} \left\| \theta - \hat{\theta}_0^{FA} \right\|_2^2.$$

ℓ_2-norm interpolation problems [GLSS18, Thm. 1]. Thus, when performed to convergence (no matter the first-order method), FTFA is equivalent to the two step procedure

$$\hat{\theta}_0^{FA} = \operatorname*{argmin}_{\theta} \sum_{j=1}^{m} p_j \frac{1}{2n_j} \|X_j\theta - \mathbf{y}_j\|_2^2 \tag{6}$$

$$\hat{\theta}_i^{FA} = \operatorname*{argmin}_{\theta} \left\{ \left\|\hat{\theta}_0^{FA} - \theta\right\|_2 \;\text{ s.t. }\; X_i\theta = \mathbf{y}_i \right\}, \tag{7}$$

outputting model $\hat{\theta}_i^{FA}$ for client i. Before giving our result, we make an additional assumption regarding the asymptotics of the number of clients and the dimension of the data.

Assumption A3. *For a constant c and $q > 2$, $(\log d)^{cq} \sum_{j=1}^m p_j^{q/2+1} n_j \to 0$ as $m, d, n_j \to \infty$.*

In Assumption A3, p_j is the weight associated with the loss of j^{th} client when finding the global model using federated averaging. To ground the assumption, consider two particular cases of interest: (i) $p_j = 1/m$, when every client is weighted equally, and (ii) $p_j = n_j/N$, when each data point is weighted equally. When $p_j = 1/m$, we have $(\log d)^{cq} \sum_{j=1}^m p_j^{q/2+1} n_j = \frac{N}{m} \frac{(\log d)^{cq}}{m^{q/2}}$. When $p_j = n_j/N$, using Assumption A1 we have $(\log d)^{cq} \sum_{j=1}^m p_j^{q/2+1} n_j = (\log d)^{cq} \sum_{j=1}^m \frac{n_j^{q/2+2}}{N^{q/2+1}} \leq (\frac{\gamma_{\max}}{\gamma_{\min}})^{q/2+1} \frac{N}{m} \frac{(\log d)^{cq}}{m^{q/2}}$. Thus, in both the cases, ignoring the polylog factors, if we have $\frac{N}{m} \frac{(\log d)^{cq}}{m^{q/2}} \to 0$ i.e., $m^{q/2}$ grows faster than the average client sample size, N/m, then Assumption A3 holds. With these assumptions defined, we are able to compute the asymptotic test loss of FTFA.

Theorem 1. *Consider the observation model in Sec. 2 and the estimator $\hat{\theta}_i^{FA}$ in (7). Let Assumption A1 hold, and let Assumptions A2 and A3 hold with $c = 2$ and $q > 2$. Additionally, assume that for each m and $j \in [m]$, $\|\mathbb{E}[\hat{\Sigma}_j^2]\|_{\text{op}} \leq \tau_2$, where $\tau_2 < \infty$. Then for client i, the asymptotic prediction bias and variance of FTFA are*

$$\lim_{m\to\infty} B_i(\hat{\theta}_i^{FA}|X) \overset{p}{=} \lim_{m\to\infty} \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2$$

$$\lim_{m\to\infty} V_i(\hat{\theta}_i^{FA}|X) \overset{p}{=} \lim_{m\to\infty} \frac{\sigma_i^2}{n_i} \operatorname{tr}(\hat{\Sigma}_i^\dagger \Sigma_i),$$

where $\Pi_i := I - \hat{\Sigma}_i^\dagger \hat{\Sigma}_i$ and $\|z\|_{\Sigma_i}^2 := z^T \Sigma_i z$. The exact expressions of these limits for general choice Σ in the implicit form can be found in the appendix. For the special case when $\Sigma_i = I$, the closed form

limits are

$$B_i(\hat{\theta}_i^{FA}|X) \xrightarrow{p} r_i^2\left(1-\frac{1}{\gamma_i}\right) \qquad V_i(\hat{\theta}_i^{FA}|X) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i-1},$$

where $\xrightarrow{p}$ *denotes convergence in probability.*

3.2 Ridge-tuned FedAvg (RTFA)

Minimum-norm results provide insight into the behavior of popular algorithms including SGM and mirror descent. Having said that, we can also analyze ridge penalized versions of FTFA. In this algorithm, the server finds the same global model as FTFA, but each client uses a regularized objective to find a local personalized model as in 2b of Algorithm 1. More concretely, in the linear regression setup, for appropriately chosen step size and as the number of steps taken goes to infinity, this corresponds to the two step procedure with the first step (6) and second step

$$\hat{\theta}_i^R(\lambda) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n_i}\|X_i\theta - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\left\|\hat{\theta}_0^{FA} - \theta\right\|_2^2, \tag{8}$$

where RTFA outputs the model $\hat{\theta}_i^R(\lambda)$ for client i. Under the same assumptions as Theorem 1, we can again calculate the asymptotic test loss.

Theorem 2. *Let the conditions of Theorem 1 hold. Then for client i, the asymptotic prediction bias and variance of RTFA are*

$$\lim_{m\to\infty} B_i(\hat{\theta}_i^R(\lambda)|X) \overset{p}{=} \lim_{m\to\infty} \lambda^2 \left\|(\hat{\Sigma}_i+\lambda I)^{-1}(\theta_0^\star - \theta_i^\star)\right\|_{\Sigma_i}^2$$
$$\lim_{m\to\infty} V_i(\hat{\theta}_i^R(\lambda)|X) \overset{p}{=} \lim_{m\to\infty} \frac{\sigma_i^2}{n_i}\operatorname{tr}(\Sigma_i\hat{\Sigma}_i(\lambda I+\hat{\Sigma}_i)^{-2}),$$

The exact expressions of these limits for general choice Σ in the implicit form can be found in the appendix. For the special case when $\Sigma_i = I$, the closed form limits are

$$B_i(\hat{\theta}_i^R(\lambda)|X) \xrightarrow{p} r_i^2\lambda^2 m_i'(-\lambda)$$
$$V_i(\hat{\theta}_i^R(\lambda)|X) \xrightarrow{p} \sigma_i^2\gamma_i(m_i(-\lambda) - \lambda m_i'(-\lambda)),$$

where $m_i(z)$ is the Stieltjes transform of the limiting spectral measure ν_i of the covariance matrix Σ_i. When $\operatorname{Cov}(\mathbf{x}_{j,k}) = I$, $m_i(z)$ has the closed form expression $m_i(z) = (1-\gamma_i-z-\sqrt{(1-\gamma_i-z)^2-4\gamma_i z})/(2\gamma_i z)$. For each client $i \in [m]$, when λ is set to be the minimizing value $\lambda_i^\star = \sigma_i^2\gamma_i/r_i^2$, the expression simplifies to $L_i(\hat{\theta}_i^R(\lambda_i^\star)|X) \to \sigma_i^2\gamma_i m_i(-\lambda_i^\star)$.

With the optimal choice of hyperparameter λ, RTFA has lower test loss than FTFA; indeed, in overparameterized linear regression, the ridge solution with regularization $\lambda \to 0$ converges to the minimum ℓ_2-norm interpolant (7).

3.3 Comparison to Naive Estimators

Three natural baselines to which we may compare FTFA and RTFA are the zero personalization estimator $\hat{\theta}_0^{FA}$, the zero collaboration estimator

$$\hat{\theta}_i^N = \underset{\theta}{\operatorname{argmin}} \|\theta\|_2 \quad s.t. \quad X_i\theta = \mathbf{y}_i, \tag{9}$$

and the ridge-penalized, zero-collaboration estimator

$$\hat{\theta}_i^N(\lambda) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n_i}\|X_i\theta - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2. \tag{10}$$

As with FTFA and RTFA, we can compute the asymptotic test loss explicitly for each. We provide expressions only for identity covariance case for clarity, similar results and comparisons hold for general covariance matrices.

Corollary 3.1. *Let the conditions of Theorem 1 hold and $\Sigma_i = I$ for i. Then for client i,*

$$B_i(\hat{\theta}_0^{FA}|X) \xrightarrow{p} r_i^2 \quad \text{and} \quad V_i(\hat{\theta}_0^{FA}|X) \xrightarrow{p} 0.$$

Consider the estimator $\hat{\theta}_i^N$ defined by eq. (9). In addition to the above conditions, suppose that $\theta_i^\star$ is drawn such that $\|\theta_i^\star\|_2 = \rho_i$ is constant with respect to m. Further assume that for some $q > 2$, for $j \in [m]$ and $k \in [n_j]$, and $l \in [d]$, $\mathbb{E}[(\mathbf{x}_{j,k})_l^{2q}] \le \kappa_q < \infty$. Then

$$B_i(\hat{\theta}_i^N|X) \xrightarrow{p} \rho_i^2\left(1 - \frac{1}{\gamma_i}\right) \quad \text{and} \quad V_i(\hat{\theta}_i^N|X) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}.$$

Consider the estimator $\hat{\theta}_i^N(\lambda)$ defined by eq. (10). Then Under the preceding conditions,

$$B_i(\hat{\theta}_i^N(\lambda)|X) \xrightarrow{p} \rho_i^2\lambda^2 m_i'(-\lambda)$$
$$V_i(\hat{\theta}_i^N(\lambda)|X) \xrightarrow{p} \sigma_i^2\gamma_i(m_i(-\lambda) - \lambda m_i'(-\lambda)).$$

Moreover, if λ is set to be the loss-minimizer $\lambda_i^\star = \sigma_i^2\gamma_i/\rho_i^2$, then $L_i(\hat{\theta}_i^N(\lambda_i^\star); \theta^\star|X) \xrightarrow{p} \sigma_i^2\gamma_i m_i(-\lambda_i^\star)$.

Key to these results are the differences between the radii $r_i^2 = \|\theta_i^\star - \theta_0^\star\|_2^2$ and $\rho_i = \|\theta_i^\star\|_2^2$, where $r_i^2 \le \rho_i^2$, and their relationship to the other problem parameters. First, it is straightforward to see that FTFA outperforms FedAvg, $\hat{\theta}_0^{FA}$, if and only if $\sigma_i^2 < r_i^2(\gamma_i - 1)/\gamma_i$. This makes intuitive sense: if the noise is too large, then local tuning is fitting mostly to noise. FTFA always outperforms the zero-collaboration estimator $\hat{\theta}_i^N$, as $\rho_i \ge r_i$, and the difference $\rho_i^2 - r_i^2$ governs the gap between collaborative and non-collaborative solutions. This remains true for the ridge-based solutions: a first-order expansion comparing Theorem 2 and Corollary 3.1 shows that for ρ_i near r_i, we have

$$\begin{aligned} L_i(\hat{\theta}_i^R(\sigma_i^2\gamma_i/r_i^2) \mid X) - L_i(\hat{\theta}_i^N(\sigma_i^2\gamma_i/\rho_i^2) \mid X) \\ = C \cdot (\rho_i^2 - r_i^2) + o(\rho_i^2 - r_i^2), \end{aligned}$$

where C depends on all the problem parameters.

With appropriate regularization λ, RTFA mitigates the weaknesses of FTFA. Thus, formally, we may show that RTFA with the optimal hyperparameter always outperforms the zero-personalization estimator $\hat{\theta}_0^{FA}$ (see the appendices). Furthermore, since $\rho_i \ge r_i$, its straightforward to see that RTFA outperforms ridgeless zero-collaboration estimator $\hat{\theta}_i^N$, and the ridge-regularized zero-collaboration estimator $\hat{\theta}_i^N(\lambda^\star)$ as well.

4 Meta learning and Proximal Regularized Algorithms

The fine-tuning procedures in the previous section provide a (perhaps) naive baseline, so we consider a few alternative federated learning procedures, both of which highlight the advantages of the high-dimensional asymptotics in its ability to predict performance. While we cannot survey the numerous procedures in FL, we pick two we consider representative: the first adapting meta learning [FMO20] and the second using a proximal-regularized approach [DTN20]. In both cases, the researchers develop convergence rates for their methods (in the former case, to stationary points), but no results on the predictive performance or their *statistical* behavior exists. We develop these in this section, showing that these more sophisticated approaches perform no better, in our asymptotic framework, than the FTFA and RTFA algorithms we outline in Section 3.

4.1 Model-Agnostic Meta-Learning

Model-Agnostic Meta-Learning (MAML) [FAL17] was learns models that adapt to related tasks by minimizing an empirical loss augmented evaluated not at a given parameter θ but at a "one-step-updated" parameter $\theta - \alpha\nabla L(\theta)$, representing one-shot learning. [FMO20], contrasting this MAML approach to the standard averaging objectives (2), adapt MAML to the federated setting, developing a method

we term MAML-FL. We describe their two step procedure in Algorithm 2 (see Section 7 for detailed pseudocode). Algorithm 2 has two variants [FMO20]; in one, the Hessian term is ignored, and in the other, the Hessian is approximated using finite differences. [FMO20] showed that these these algorithms converge to a stationary point of eq. (11) (with $p_j = 1/m$) for general non-convex smooth functions.

Algorithm 2 MAML-FL (details in Appendix 7)

1. Server and clients coordinate to (approximately) solve

$$\hat{\theta}_0^M(\alpha) = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{m} p_j \widehat{L}_j(\theta - \alpha \nabla \widehat{L}_j(\theta)), \tag{11}$$

   where $p_j \in (0,1)$ are weights such that $\sum_{j=1}^m p_j = 1$ and α denotes stepsize. Server broadcasts global model $\hat{\theta}_0^M(\alpha)$ to clients.

2. Client i learns a model $\hat{\theta}_i^M(\alpha)$ by optimizing its empirical risk, $\widehat{L}_i(\cdot)$, using SGM initialized at $\hat{\theta}_0^M(\alpha)$

In our linear model, for appropriately chosen hyperparameters and as the number of steps taken goes to infinity, this personalization method corresponds to the following two step procedure:

$$\begin{aligned} &\hat{\theta}_0^M(\alpha) \\ &= \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{m} \frac{p_j}{2n_j} \left\| X_j \left[\theta - \frac{\alpha}{n_j} X_j^T (X_j \theta - \mathbf{y}_j) \right] - \mathbf{y}_j \right\|_2^2 \end{aligned} \tag{12}$$

$$\hat{\theta}_i^M(\alpha) = \underset{\theta}{\operatorname{argmin}} \left\| \hat{\theta}_0^M(\alpha) - \theta \right\|_2 \quad s.t. \;\; X_i \theta = \mathbf{y}_i \tag{13}$$

As in Section 3.1, we assume that the client model in step 2 of Alg. 2 has fully converged; any convergent first-order method converges to the minimum norm interpolant (13). The representations (12) and (13) allow us to analyze the test loss of the MAML-FL personalization scheme in our asymptotic framework.

Theorem 3. *Consider the observation model in Section 2 and the estimator $\hat{\theta}_i^M(\alpha)$ in (13). Let Assumption A1 hold, Assumption A2 hold with $q = 3v$ where $v > 2$, and Assumption A3 hold with $c = 5$ and $q = v$. Additionally, assume that for each m and all $j \in [m]$, $\lambda_{\min}(\mathbb{E}[\hat{\Sigma}_j (I - \alpha \hat{\Sigma}_j)^2]) \geq \lambda_0 > 0$ and $\|\mathbb{E}[\hat{\Sigma}_j^6]\|_{\mathrm{op}} \leq \tau_6 < \infty$. Then for client i, the asymptotic prediction bias and variance of MAML-FL are*

$$\begin{aligned} \lim_{m\to\infty} B_i(\hat{\theta}_i^M(\alpha)|X) &\overset{p}{=} \lim_{m\to\infty} \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 \\ \lim_{m\to\infty} V_i(\hat{\theta}_i^M(\alpha)|X) &\overset{p}{=} \lim_{m\to\infty} \frac{\sigma_i^2}{n_i} \operatorname{tr}(\hat{\Sigma}_i^\dagger \Sigma_i), \end{aligned}$$

where $\Pi_i := I - \hat{\Sigma}_i^\dagger \hat{\Sigma}_i$ and $\|z\|_{\Sigma_i}^2 := z^T \Sigma_i z$. The exact expressions of these limits for general choice Σ in the implicit form are in the appendices. For the special case that $\Sigma_i = I$, the closed form limits are

$$\begin{aligned} B_i(\hat{\theta}_i^M(\alpha)|X) &\xrightarrow{p} r_i^2 \left(1 - \frac{1}{\gamma_i}\right) \\ V_i(\hat{\theta}_i^M(\alpha)|X) &\xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}. \end{aligned}$$

In short, the asymptotic test risk of MAML-FL matches that of FTFA (Theorem 1). In general, the MAML-FL objective (11) is typically non-convex even when $\widehat{L}_j$ is convex, making convergence subtle. Even ignoring convexity, the inclusion of a derivative term in the objective can make the standard smoothness conditions [Nes04] upon which convergence analyses (and algorithms) repose fail to hold. Additionally, computing gradients of the MAML-FL objective (11) requires potentially expensive second-order derivative computations or careful approximations to these, making optimization

more challenging and expensive irrespective of convexity. We provide more discussion in the appendices. Theorem 3 thus suggests that one might be circumspect about choosing MAML-FL or similar algorithms over simpler baselines that do not require such complexity in optimization.

Remark The algorithm [FMO20] propose performs only a single stochastic gradient step for personalization, which is distinct from our analyzed procedure (13), as it is essentially equivalent to running SGM until convergence from the initialization $\hat{\theta}_0^M(\alpha)M\alpha$ (see step 2 of Algorithm 2). We find two main justifications for this choice: first, experimental work of [JKRK19], in addition to our own experiments (see Figures 5 and 6) empirically suggest that the more (stochastic gradient) steps of personalization, the better performance we expect. Furthermore, as we mention above earlier, performing personalization SGM steps locally, in parallel, and asynchronously is no more expensive than running the first step of Algorithm 2. This guaranteed of convergence also presents a fair point of comparison between the algorithms we consider.

4.2 Proximal-Regularized Approach

Instead of a sequential, fine-tuning approach, an alternative approach to personalization involves jointly optimizing global and local parameters. In this vein, [DTN20] propose the pFedMe algorithm (whose details we provide in the appendices) to solve the following coupled optimization problem to find personalized models for each client:

$$\left(\hat{\theta}_0^P(\lambda), \hat{\theta}_1^P(\lambda), \ldots, \hat{\theta}_m^P(\lambda)\right) = \\ \underset{\theta_0,\theta_1,\ldots,\theta_m}{\operatorname{argmin}} \sum_{j=1}^{m} p_j \left(\widehat{L}_i(\theta_j) + \frac{\lambda}{2}\left\|\theta_j - \theta_0\right\|_2^2\right).$$

The proximal penalty encourages the local models θ_i to be close to one another. In our linear model, for appropriately chosen hyperparameters and as the number of steps taken goes to infinity, the proposed optimization problem simplifies to

$$\left(\hat{\theta}_0^P(\lambda), \hat{\theta}_1^P(\lambda), \ldots, \hat{\theta}_m^P(\lambda)\right) = \\ \underset{\theta_0,\theta_1,\ldots,\theta_m}{\operatorname{argmin}} \sum_{j=1}^{m} p_j \left(\frac{1}{2n_j}\left\|X_j\theta_j - \mathbf{y}_j\right\|_2^2 + \frac{\lambda}{2}\left\|\theta_j - \theta_0\right\|_2^2\right), \tag{14}$$

where $\hat{\theta}_0^P(\lambda)$ denotes the global model and $\hat{\theta}_i^P(\lambda)$ denote the local models. We can again use our asymptotic framework to analyze the test loss of this scheme. For this result, we use an additional condition on $\sup_{j\in[m]} \mathbb{P}(\lambda_{\max}(\hat{\Sigma}_j) > R)$ that gives us uniform control over the eigenvalues of all the users.

Theorem 4. *Consider the observation model n section 2 and the estimator $\hat{\theta}_i^P(\lambda)$ in (14). Let Assumption A1 hold, and let Assumptions A3 and A2 hold with $c = 2$ and the same $q > 2$. Additionally, assume that $\mathbb{E}[\|\hat{\Sigma}_j^2\|_{\text{op}}] \leq \tau_3 < \infty$. Further suppose that there exists $R \geq M$ such that $\limsup_{m\to\infty} \sup_{j\in[m]} \mathbb{P}(\lambda_{\max}(\hat{\Sigma}_j) > R) \leq \frac{1}{16M^2\tau_3}$. Then for client i, the asymptotic prediction bias and variance of pFedMe are*

$$\lim_{m\to\infty} B_i(\hat{\theta}_i^P(\lambda)|X) \overset{p}{=} \lim_{m\to\infty} \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}(\theta_0^\star - \theta_i^\star)\right\|_{\Sigma_i}^2$$

$$\lim_{m\to\infty} V_i(\hat{\theta}_i^P(\lambda)|X) \overset{p}{=} \lim_{m\to\infty} \frac{\sigma_i^2}{n_i} \operatorname{tr}(\Sigma_i\hat{\Sigma}_i(\lambda I + \hat{\Sigma}_i)^{-2}),$$

See the appendices for exact expressions of these limits for general Σ. For the special case that $\Sigma_i = I$, the limits are

$$B_i(\hat{\theta}_i^P(\lambda)|X) \xrightarrow{p} r_i^2\lambda^2 m_i'(-\lambda)$$

$$V_i(\hat{\theta}_i^P(\lambda)|X) \xrightarrow{p} \sigma_i^2\gamma(m_i(-\lambda) - \lambda m_i'(-\lambda)),$$

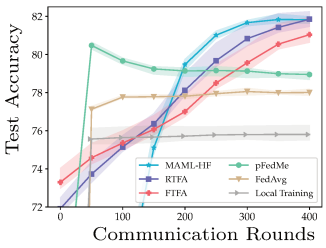

Figure 1. CIFAR-100. Best-average-worst intervals created from different train-val splits.

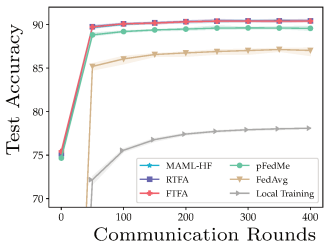

Figure 2. EMNIST. Best-average-worst intervals created from different random seeds.

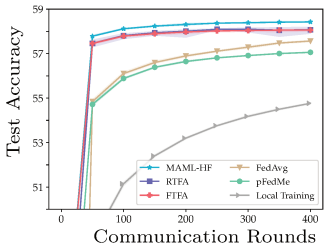

Figure 3. Shakespeare. Best-average-worst intervals created from different random seeds.

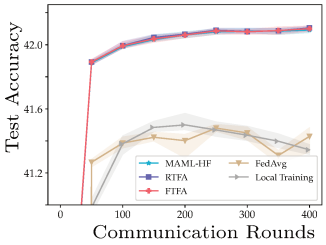

Figure 4. Stack Overflow. Best-average-worst intervals created from different train-val splits.

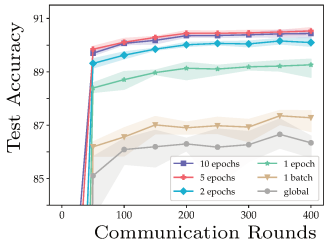

Figure 5. EMNIST. Gains of personalization for FTFA

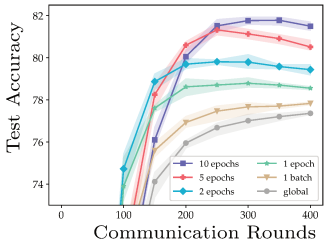

Figure 6. CIFAR. Gains of personalization for Hessian free MAML-FL

where $m_i(z)$ is as in Theorem 2. For each client $i \in [m]$, the minimizing λ is $\lambda_i^\star = \sigma_i^2\gamma_i/r_i^2$ and $L_i(\hat{\theta}_i^P(\lambda_i^\star)|X) \to \sigma_i^2\gamma_i m_i(-\lambda_i^\star)$.

The asymptotic test loss of the proximal-regularized approach is thus identical to the locally ridge-regularized (RTFA) solution; see Theorem 2. [DTN20]'s algorithm to optimize eq. (14) is sensitive to hyperparameter choice, meaning significant hyperparameter tuning may be needed for good performance and even convergence of the method (of course, both methods do require tuning λ). Moreover, a local update step in pFedMe requires approximately solving a proximal-regularized optimization problem, as opposed to taking a single stochastic gradient step. This can make pFedMe much more computationally expensive depending on the properties of $\widehat{L}$. This is not to dismiss more complex proximal-type algorithms, but only to say that, at least in our analytical framework, simpler and embarassingly parallelizable procedures (RTFA in this case) may suffice to capture the advantages of a proximal-regularized scheme.

5 Experiments

While the statistical model we assume in our analytical sections is stylized and certainly will not fully hold, it suggests some guidance in practice, and make precise predictions about the error rates of different methods: that the simpler fine-tuning methods should exhibit performance comparable to more complex federated methods, such as MAML-FL and pFedMe. With this in mind, we turn to several datasets, performing experiments on federated versions of the Shakespeare [MMR+17], CIFAR-100 [KH09], EMNIST [CATvS17], and Stack Overflow [MRR+19] datasets; dataset statistics and details of how we divide the data to make effective "users" are in Section 8. For each dataset, we compare the performance of the following algorithms: Zero Communication (Local Training), Zero Personalization (FedAvg), FTFA, RTFA, MAML-FL, and pFedMe [DTN20]. For each classification task, we use each federated learning algorithm to train the last layer of a pre-trained neural network. We run each algorithm for 400 communication rounds, and we compute the test accuracy (the fraction of correctly classified test data points across machines) every 50 communication rounds. FTFA, RTFA, and MAML-

FL each perform 10 epochs of local training for each client before the evaluation of test accuracy. For each client, pFedMe uses the local models to compute test accuracy. We first hyperparameter tune each method using training and validation splits; again, see Appendix 8 for details. We track the test accuracy of each tuned method over 11 trials using two different kinds of randomness:

1. *Different seeds*: We run each hyperparameter-tuned method on 11 different seeds. This captures how different initializations and batching affect accuracy.

2. *Different training-validation splits*: We generate 11 different training / validation splits (same test data) and run each hyperparameter-tuned method on each split. This captures how variations in user data affect test accuracy.

Experimental setting Our experiments are "semi-synthetic" in that in each, we re-fit the top layer of a pre-trained neural network. While this differs from some practice with experimental work in federated learning, several considerations motivate our choices to take this tack, and we contend they may be valuable for other researchers: (i) our (distributed) models are convex, that is, can be fit via convex optimization. In the context of real-world engineering problems, it is important to know when a method has converged and, if it does not, why it has not; in this vein, non-convexity can be a bugaboo, as it hides the causes of divergent algorithms—is it non-convexity and poor optimization or engineering issues (e.g. communication bugs)? This choice thus can be valuable even in real, large-scale systems. (ii) In our experiments, we achieve state-of-the-art or near state-of-the-art results; using federated approaches to fit full deep models appears to lead to substantial degradation in performance over a single centralized, pre-trained model (see, e.g., [RCZ+21, Table 1], where accuracies on CIFAR-10 using a ResNet 18 are at best 78%, substantially lower than current state-of-the-art). A question whose answer we do not know: if a federated learning method provides worse performance than a downloadable model, what does the FL method's performance tell us about good methodologies in federated learning? (iii) Finally, computing with large-scale distributed models is computationally expensive: the energy use for fitting large distributed models is substantial and may be a poor use of resources [SGM19]. In effort to better approximate the use of a pre-trained model in real federated learning applications, we use held-out data to pre-train a preliminary network in our Stack Overflow experiments, doing the experimental training and validation on an independent dataset.

Results Figures 1 to 4 plot test accuracy against communication rounds. The performance of MAML-FL is similar to that of FTFA and RTFA, and on the Stack Overflow and EMNIST datasets, where the total dataset size is much larger than the other datasets, the accuracies of MAML-FL, FTFA and RTFA are nearly identical. This is consistent with our theoretical claims. The performances of the naive, zero communication and zero personalization algorithms are worse than that of FTFA, RTFA and MAML-FL in all figures. This is also consistent with our theoretical claims. The performance of pFedMe in Figures 1 to 3 is worse than that of FTFA, RTFA and MAML-FL.

In Figures 5 and 6, we plot the test accuracy of FTFA and MAML-FL and vary the number of personalization steps each algorithm takes. In both plots, the global model performs the worst, and performance improves monotonically as we increase the number of personalization steps. As personalization steps are cheap relative to the centralized training procedure, this suggests benefits for clients to locally train to convergence.

References

[BY93] ZD Bai and YQ Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

[Car97] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[CATvS17] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[CGT12] Richard Chen, Alex Gittens, and Joel A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, to appear, 2012.

[CRS+19] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems 32*, 2019.

[CZLS21] Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J. Su. A theorem of the alternative for personalized federated learning. *arXiv:2103.01901 [stat.ML]*, 2021.

[dlPnG99] Victor H. de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, 1999.

[DTN20] Canh Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *arXiv:2006.08848 [cs.LG]*, 2020.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[Fel19] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. *arXiv:1906.05271 [cs.LG]*, 2019.

[FMO20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems 33*, 2020.

[GLSS18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[HLS+20] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. FedML: A research library and benchmark for federated machine learning. *arXiv:2007.13518 [cs.LG]*, 2020.

[HM19] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv:1910.14425 [cs.LG]*, 2019.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan Tibshirani. Surprises in high-dimensional ridgeless linear least squares interpolation. *arXiv:1903.08560 [math.ST]*, 2019.

[HR18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv:1801.06146 [cs.LG]*, 2018.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[JKRK19] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv:1909.12488 [cs.LG]*, 2019.

[KH09] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[KKM+20] 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 37th International Conference on Machine Learning*, 2020.

[KMA+19] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv:1912.04977 [cs.LG]*, 2019.

[LHBS21] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, 2021.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[MMR+17] H. 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 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[MMRS20] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv:2002.10619 [cs.LG]*, 2020.

[MRR+19] Brendan McMahan, Keith Rush, Michael Reneer, Zachary Garrett, and TensorFlow Federated Team. Tensorflow federated stack overflow dataset. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data, 2019.

[MSS19] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[RCZ+21] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the Ninth International Conference on Learning Representations*, 2021.

[SCST17] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems 17*, 2017.

[SGM19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[SS90] G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[Vap92] V. Vapnik. Principles of risk minimization for learning theory. In John E. Moody, Steve J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. Morgan Kaufmann, 1992.

[Vap95] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[wei20] weiaicunzai. Pytorch-cifar100. https://github.com/weiaicunzai/pytorch-cifar100, 2020.

[WMK+19] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv:1910.10252 [cs.LG]*, 2019.

[YBS20] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv:2002.04758 [cs.LG]*, 2020.

[ZMM+20] Edvin Listo Zec, Olof Mogren, John Martinsson, Leon René Sütfeld, and Daniel Gillblad. Specialized federated learning using a mixture of experts. *arXiv:2010.02056 [cs.LG]*, 2020.

6 Proofs

6.1 Additional Notation

To simplify notation, we define some aggregated parameters, $X_i := [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}]^T \in \mathbb{R}^{n_i \times d}$, $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,n_i}]^T \in \mathbb{R}^{n_i}$, $X := [X_1^T, \ldots, X_m^T]^T \in \mathbb{R}^{N \times d}$, and $\mathbf{y} := [\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T]^T \in \mathbb{R}^N$. Additionally, we define $\hat{\Sigma}_i := X_i^T X_i / n_i \in \mathbb{R}^{d \times d}$. We use the notation $a \lesssim b$ to denote $a \leq Kb$ for some absolute constant K.

6.2 Useful Lemmas

Lemma 6.1. *Let $\mathbf{x}_j$ be vectors in $\mathbb{R}^d$ and let ζ_j be Rademacher (± 1) random variables. Then, we have*

$$\mathbb{E}\left[\left\|\sum_{j=1}^m \zeta_j \mathbf{x}_j\right\|_2^p\right]^{1/p} \leq \sqrt{p-1}\left(\sum_{j=1}^m \|\mathbf{x}_j\|_2^2\right)^{1/2},$$

where the expectation is over the Rademacher random variables.

Proof Using Theorem 1.3.1 of [dlPnG99], we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^m \zeta_j \mathbf{x}_j\right\|_2^p\right]^{1/p} &\leq \sqrt{p-1}\mathbb{E}\left[\left\|\sum_{j=1}^m \zeta_j \mathbf{x}_j\right\|_2^2\right]^{1/2} \\
&= \sqrt{p-1}\mathbb{E}\left[\sum_{i,j=1}^m \langle \zeta_j \zeta_i \mathbf{x}_j^T \mathbf{x}_i \rangle\right] \\
&= \sqrt{p-1}\left(\sum_{j=1}^m \|\mathbf{x}_j\|_2^2\right)^{1/2}
\end{aligned}$$

□

Lemma 6.2. *For all clients $j \in [m]$, let the data $\mathbf{x}_{j,k} \in \mathbb{R}^d$ for $k \in [n]$ be such that $\mathbf{x}_{j,k} = \Sigma_j^{1/2} \mathbf{z}_{j,k}$ for some Σ_j, $\mathbf{z}_{j,k}$, and $p > 2$ that satisfy Assumption A2. Let $(\mathbf{x}_{j,k})_l \in \mathbb{R}$ denote the $l \in [d]$ entry of the vector $\mathbf{x}_{j,k} \in \mathbb{R}^d$. Define $\hat{\Sigma}_j = \frac{1}{n_j}\sum_{k \in [n_j]} \mathbf{x}_{j,k}\mathbf{x}_{j,k}^T$. Then, we have*

$$\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^p\right] \leq K(e \log d)^p n_j,$$

where the inequality holds up to constant factors for sufficiently large m.

Proof We first show a helpful fact that $\mathbb{E}[(\mathbf{z}_{j,k})_l^{2p}] \leq \kappa_p < \infty$ implies $\mathbb{E}[\|\mathbf{x}_{j,k}\|_2^{2p}]^{1/(2p)} \lesssim \sqrt{d}$. For any $j \in [m]$, we have by Jensen's inequality

$$\mathbb{E}[\|\mathbf{x}_{j,k}\|_2^{2p}] \leq M^{2p}\mathbb{E}[\|\mathbf{z}_{j,k}\|_2^{2p}] = M^{2p}d^p\mathbb{E}\left[\left(\frac{1}{d}\sum_{l=1}^d (\mathbf{z}_{j,k})_l^2\right)^p\right] \leq M^{2p}d^p\frac{1}{d}\sum_{l=1}^d \mathbb{E}[(\mathbf{z}_{j,k})_l^{2p}] \leq M^{2p}\kappa_p d^p$$

We define some constant $C_4 > M^{2p}\kappa_p$. With this fact and Theorem A.1 from [CGT12], we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^p\right] &= \mathbb{E}\left[\left\|\sum_{k=1}^n \frac{\mathbf{x}_{j,k}\mathbf{x}_{j,k}^T}{n}\right\|_{\text{op}}^p\right] \leq 2^{2p-1}\left(\|\Sigma_j\|_{\text{op}}^p + \frac{(e\log d)^p}{n_j^p}\mathbb{E}\left[\max_k \left\|\mathbf{x}_{j,k}\mathbf{x}_{j,k}^T\right\|_{\text{op}}^p\right]\right) \\
&\leq 2^{2p-1}\left(C + \frac{(e\log d)^p}{n_j^{p-1}}\mathbb{E}\left[\|\mathbf{x}_{j,k}\|_2^{2p}\right]\right) \\
&\leq 2^{2p-1}\left(C + C_4\frac{(e\log d)^p d^p}{n_j^{p-1}}\right)
\end{aligned}$$

Now, $2^{2p-1}\left(C + C_4\frac{(e\log d)^p d^p}{n_j^{p-1}}\right) \leq K(e\log d)^p n_j$ for some absolute constant K since $\frac{d}{n_j} \to \gamma_i$. □

Lemma 6.3. *For all clients $j \in [m]$, let the data $\mathbf{x}_{j,k} \in \mathbb{R}^d$ for $k \in [n]$ be such that $\mathbf{x}_{j,k} = \Sigma_j^{1/2}\mathbf{z}_{j,k}$ for some Σ_j, $\mathbf{z}_{j,k}$, and $q' > 2$ that satisfy Assumption A2. Further let $q' = pq$ where $p \geq 1$ and $q \geq 2$. Let $\hat{\Sigma}_j = \frac{1}{n_j}\sum_{k\in[n_j]}\mathbf{x}_{j,k}\mathbf{x}_{j,k}^T$ and $\mu_j = \mathbb{E}[\hat{\Sigma}_j^p]$. Additionally assume that $\left\|\mathbb{E}[\hat{\Sigma}_j^{2p}]\right\|_{\text{op}} \leq C_3$ for some constant C_3. Let d, n_j grow as in Assumption A1. Then we have for sufficiently large m,*

$$\mathbb{P}\left(\left\|\sum_{j=1}^m p_j\left(\hat{\Sigma}_j^p - \mu_j\right)\right\|_{\text{op}} > t\right) \leq \frac{2^{q-1}C_2}{t^q}\left[(\log d)^{q/2}\sum_{j=1}^m p_j^{q/2+1} + (\log d)^{pq+q}\sum_{j=1}^m p_j^q n_j\right].$$

Further supposing that $(\log d)^{pq+q}\sum_{j=1}^m p_j^q n_j \to 0$, we get that $\left\|\sum_{j=1}^m p_j\left(\hat{\Sigma}_j^p - \mu_j\right)\right\|_{\text{op}} \xrightarrow{p} 0$.

Proof Using Markov's inequality, Jensen's inequality, and symmeterization, we have with ζ_j iid Rademacher

$$\mathbb{P}\left(\left\|\sum_{j=1}^m p_j\left(\hat{\Sigma}_j^p - \mu_j\right)\right\|_{\text{op}} > t\right) \leq \frac{\mathbb{E}\left[\left\|\sum_{j=1}^m p_j(\hat{\Sigma}_j^p - \mu_j)\right\|_{\text{op}}^q\right]}{t^q} \leq 2^q\frac{\mathbb{E}\left[\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j^p\zeta_j\right\|_{\text{op}}^q\right]}{t^q}$$

We use the second part of Theorem A.1 with $q \geq 2$ from [CGT12] to bound the RHS.

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j^p\zeta_j\right\|_{\text{op}}^q\right] &\leq \left(\sqrt{e\log d}\left\|\mathbb{E}[\sum_{j=1}^m p_j^2\hat{\Sigma}_j^{2p}]^{1/2}\right\|_{\text{op}} + (2e\log d)\mathbb{E}\left[\max_j\left\|p_j\hat{\Sigma}_j^p\right\|_{\text{op}}^q\right]^{1/q}\right)^q \\
&\leq 2^{q-1}(e\log d)^{q/2}\left\|\mathbb{E}[\sum_{j=1}^m p_j^2\hat{\Sigma}_j^{2p}]^{1/2}\right\|_{\text{op}}^q + 2^{q-1}(e\log d)^q\mathbb{E}\left[\max_j p_j^q\left\|\hat{\Sigma}_j^p\right\|_{\text{op}}^q\right] \\
&\leq 2^{q-1}(e\log d)^{q/2}\left\|\mathbb{E}\left[\sum_{j=1}^m p_j^2\hat{\Sigma}_j^{2p}\right]\right\|_{\text{op}}^{q/2} + 2^{q-1}(e\log d)^q\mathbb{E}\left[\sum_{j=1}^m p_j^q\left\|\hat{\Sigma}_j\right\|_{\text{op}}^{pq}\right]
\end{aligned}$$

Now we bound the RHS of this quantity using the first part of Theorem A.1. For each $j \in [m]$, we have by Lemma 6.2 for sufficiently large m,

$$\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^{pq}\right] \leq K(e\log d)^{pq}n_j,$$

for some absolute constant K. Supposing that $\left\|\mathbb{E}\left[\hat{\Sigma}_j^{2p}\right]\right\|_{\text{op}} \leq C_3$ exist for all j. Combining all the inequalities, we have for sufficiently large m,

$$\mathbb{E}\left[\left\|\sum_{j=1}^{m} p_j \hat{\Sigma}_j^p \xi_j\right\|_{\text{op}}^q\right] \leq 2^{q-1}(e\log d)^{q/2}\left(\sum_{j=1}^{m} p_j^{q/2+1}\left\|\mathbb{E}\left[\hat{\Sigma}_j^{2p}\right]\right\|_{\text{op}}^{q/2}\right)$$
$$+ 2^{q-1}(e\log d)^q \sum_{j=1}^{m} p_j^q K (e\log d)^{pq} n_j$$
$$\leq C_2\left[(\log d)^{q/2}\sum_{j=1}^{m} p_j^{q/2+1} + (\log d)^{pq+q}\sum_{j=1}^{m} p_j^q n_j\right],$$

where in the first term of the first inequality, we use Jensen's inequality to pull out $\sum_{j=1}^m p_j$ of the expectation.

To prove the second part of the lemma, we observe that if $(\log d)^{(p+1)q}\sum_{j=1}^m p_j^q n_j \to 0$ as $m \to \infty$ such that $d/n_i \to \gamma_i > 1$ for all devices $i \in [m]$, then $(\log d)^{q/2}\sum_{j=1}^m p_j^{q/2+1} \to 0$. To see this, we first observe

$$(\log d)^{(p+1)q}\sum_{j=1}^{m} p_j^q n_j \geq (\max_{j\in[m]} p_j (\log d)^{p+1})^q,$$

so we know that $\max_{j\in[m]} p_j(\log d)^{p+1} \to 0$. Further, by Holder's inequality, we know that

$$(\log d)^{q/2}\sum_{j=1}^{m} p_j^{q/2+1} \leq (\max_{j\in[m]} p_j \log d)^{q/2}.$$

By the continuity of the $q/2$ power, we get the result. □

Lemma 6.4. *Let $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{d\times d}$ be positive semidefinite matrices such that $\lambda_{\min}(U) \geq \lambda_0$ for some constant λ_0. Let $d, n_j, m \to \infty$ as in Assumption A1. Suppose $\|V-U\|_{\text{op}} \xrightarrow{p} 0$, then $\|V^{-1}-U^{-1}\|_{\text{op}} \xrightarrow{p} 0$.*

Proof For any $t > 0$, we have by Theorem 2.5 (from Section III) of [SS90]

$$P\left(\|V^{-1}-U^{-1}\|_{\text{op}} > t\right)$$
$$\leq P\left(\|V^{-1}-U^{-1}\|_{\text{op}} > t \cap \|V-U\|_{\text{op}} < \frac{1}{\|U^{-1}\|_{\text{op}}}\right) + P\left(\|V-U\|_{\text{op}} \geq \lambda_0\right)$$
$$\leq P\left(\|U^{-1}(V-U)\|_{\text{op}} > \frac{t}{t+\|U^{-1}\|_{\text{op}}}\right) + o(1)$$
$$\leq P\left(\|V-U\|_{\text{op}} > \frac{t\lambda_0}{t+\lambda_0^{-1}}\right) + o(1)$$

We know this quantity goes to 0 by assumption. □

6.3 Some useful definitions from previous work

In this section, we recall some definitions from [HMRT19] that will be useful in finding the exact expressions for risk. The expressions for asymptotic risk in high dimensional regression problems (both ridge and ridgeless) are given in an implicit form in [HMRT19]. It depends on the geometry of the covariance matrix Σ and the true solution to the regression problem $\theta^\star$. Let $\Sigma = \sum_{i=1}^d s_i v_i v_i^T$ denote

the eigenvalue decomposition of Σ with $s_1 \geq s_2 \cdots \geq s_d$, and let $(c, \ldots, v_d^T\theta^\star)$ denote the inner products of $\theta^\star$ with the eigenvectors. We define two probability distributions which will be useful in giving the expressions for risk:

$$\widehat{H}_n(s) \coloneqq \frac{1}{d}\sum_{i=1}^{d} 1\{s \geq s_i\}\,, \qquad \widehat{G}_n(s) \coloneqq \frac{1}{\|\theta^\star\|_2^2}\sum_{i=1}^{d}(v_i^T\theta^\star)^2 1\{s \geq s_i\}\,.$$

Note that $\widehat{G}_n$ is a reweighted version of $\widehat{H}_n$ and both have the same support (eigenvalues of Σ).

Definition 6.1. *For $\gamma \in \mathbb{R}^+$, let $c_0 = c_0(\gamma, \widehat{H}_n)$ be the unique non-negative solution of*

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0\gamma s} d\widehat{H}_n(s),$$

the predicted bias and variance is then defined as

$$\mathscr{B}(\widehat{H}_n, \widehat{G}_n, \gamma) \coloneqq \|\theta^\star\|_2^2 \left\{1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)} d\widehat{H}_n(s)}\right\} \cdot \int \frac{s}{(1 + c_0\gamma s)} d\widehat{G}_n(s), \tag{15}$$

$$\mathscr{V}(\widehat{H}_n, \gamma) \coloneqq \sigma^2\gamma \frac{\int \frac{s^2}{(1+c_0\gamma s)} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)} d\widehat{H}_n(s)}. \tag{16}$$

Definition 6.2. *For $\gamma \in \mathbb{R}^+$ and $z \in \mathbb{C}_+$, let $m_n(z) = m(z; \widehat{H}_n, \gamma)$ be the unique solution of*

$$m_n(z) \coloneqq \int \frac{1}{s[1 - \gamma - \gamma z m_n(z)] - z} d\widehat{H}_n(s).$$

Further, define $m_{n,1}(z) = m_{n,1}(z; \widehat{H}_n, \gamma)$ as

$$m_{n,1}(z) \coloneqq \frac{\int \frac{s^2[1-\gamma-\gamma z m_n(z)]}{[s[1-\gamma-\gamma z m_n(z)]-z]^2} d\widehat{H}_n(s)}{1 - \gamma\int \frac{zs}{[s[1-\gamma-\gamma z m_n(z)]-z]^2} d\widehat{H}_n(s)}$$

The definitions are extended analytically to $Im(z) = 0$ whenever possible, the predicted bias and variance are then defined by

$$\mathscr{B}(\lambda; \widehat{H}_n, \widehat{G}_n, \gamma) \coloneqq \lambda^2 \|\theta^\star\|_2 (1 + \gamma m_{n,1}(-\lambda)) \int \frac{s}{[\lambda + (1 - \gamma + \gamma\lambda m_n(-\lambda))s]^2} d\widehat{G}_n(s), \tag{17}$$

$$\mathscr{V}(\lambda; \widehat{H}_n, \gamma) \coloneqq \sigma^2\gamma \int \frac{s^2((1 - \gamma + \gamma\lambda m_n'(-\lambda)))}{[\lambda + (1 - \gamma + \gamma\lambda m_n(-\lambda))s]^2} d\widehat{H}_n(s). \tag{18}$$

6.4 Proof of Theorem 1

On solving (6) and (7), the closed form of the estimators $\hat{\theta}_0^{FA}$ and $\hat{\theta}_i^{FA}$ is given by

$$\begin{aligned}\hat{\theta}_0^{FA} &= \operatorname*{argmin}_\theta \sum_{j=1}^{m} p_j \frac{1}{2n_j} \|X_j\theta - \mathbf{y}_j\|_2^2 = \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \frac{X_j^T \mathbf{y}_j}{n_j} \\ &= \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j \theta_j^\star + \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \frac{X_j^T \xi_j}{n_j}\end{aligned} \tag{19}$$

and

$$\begin{aligned}\hat{\theta}_i^{FA} &= (I - \hat{\Sigma}_i^\dagger \hat{\Sigma}_i)\hat{\theta}_0^{FA} + X_i^\dagger \mathbf{y}_i = (I - \hat{\Sigma}_i^\dagger \hat{\Sigma}_i)\hat{\theta}_0^{FA} + \hat{\Sigma}_i^\dagger \hat{\Sigma}_i \theta_i^\star + \frac{1}{n_i}\hat{\Sigma}_i^\dagger X_i^T \xi_i \\ &= \Pi_i \left[\left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j \theta_j^\star + \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \frac{X_j^T \xi_j}{n_j}\right] + \hat{\Sigma}_i^\dagger \hat{\Sigma}_i \theta_i^\star + \frac{1}{n_i}\hat{\Sigma}_i^\dagger X_i^T \xi_i\end{aligned}$$

We now calculate the risk by splitting it into two parts as in (3), and then calculate the asymptotic bias and variance.

Bias:

$$B_i(\hat{\theta}_i^{FA}|X) \coloneqq \left\|\mathbb{E}[\hat{\theta}_i^{FA}|X] - \theta_i^\star\right\|_{\Sigma_i}^2 = \left\|\Pi_i \left[\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j\hat{\Sigma}_j(\theta_j^\star - \theta_i^\star)\right]\right\|_{\Sigma_i}^2$$
$$= \left\|\Sigma_i^{1/2}\Pi_i \left[\theta_0^\star - \theta_i^\star + \left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j\hat{\Sigma}_j(\theta_j^\star - \theta_0^\star)\right]\right\|_2^2$$

The idea is to show that the second term goes to 0 and use results from [HMRT19] to find the asymptotic bias. For simplicity, we let $\Delta_j := \theta_j^\star - \theta_0^\star$, and we define the event:

$$B_t := \left\{\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1} - \left(\sum_{j=1}^m p_j\Sigma_j\right)^{-1}\right\|_{\text{op}} > t\right\}$$
$$A_t := \left\{\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_{\Sigma_i} > t\right\}$$

The proof proceeds in the following steps:

Bias Proof Outline

Step 1. We first show for any $t > 0$, the $\mathbb{P}(B_t) \to 0$ as $d \to \infty$

Step 2. Then we show for any $t > 0$, the $\mathbb{P}(A_t) \to 0$ as $d \to \infty$

Step 3. We show that for any $t \in (0, 1]$on event A_t^c, $B_i(\hat{\theta}_i^{FA}|X) \leq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + ct$ and $B_i(\hat{\theta}_i^{FA}|X) \geq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 - ct$

Step 4. Show that $\lim_{d\to\infty} \mathbb{P}(|B_i(\hat{\theta}_i^{FA}|X) - \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2| \leq \varepsilon) = 1$

Step 5. Finally, using the asymptotic limit of $\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2$ from Theorem 1 of [HMRT19], we get the result.

Step 1 Since we have $\lambda_{\min}\left(\sum_{j=1}^m p_j\Sigma_j\right) > 1/M > 0$, it suffices to show by Lemma 6.4 that the probability of

$$C_t := \left\{\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j - \sum_{j=1}^m p_j\Sigma_j\right\|_{\text{op}} > t\right\}$$

goes to 0 as $d, m \to \infty$ (obeying Assumption A1). Using Lemma 6.3 with $p = 1$, we have that

$$\mathbb{P}(C_t) \leq \frac{2^{q-1}C_2}{t^q}\left[(\log d)^{q/2}\sum_{j=1}^m p_j^{q/2+1} + (\log d)^{2q}\sum_{j=1}^m p_j^q n_j\right]$$

Since $(\log d)^{2q}\sum_{j=1}^m p_j^q n_j \to 0$, this quantity goes to 0.

Step 2 Fix any $t > 0$,

$$\mathbb{P}(A_t) \leq \mathbb{P}\left(\left\{\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_{\Sigma_i} > t\right\} \cap B_{c_1}^c\right) + \mathbb{P}\left(B_{c_1}\right)$$
$$\leq \mathbb{P}\left(M(c_1+M)\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_2 > t\right) + \mathbb{P}\left(B_{c_1}\right)$$

By Step 1, we know that $\mathbb{P}\left(B_{c_1}\right) \to 0$. The second inequality comes from $\|Ax\|_2 \leq |\!|\!|A|\!|\!|_{\text{op}} \|x\|_2$ and triangle inequality. Now to bound the first term, we use Markov and a Khintchine inequality (Lemma 6.1). We have that

$$\mathbb{P}\left(M(c_1+M)\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_2 > t\right) \leq \frac{(M(c_1+M))^q\mathbb{E}\left[\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_2^q\right]}{t^q}$$
$$\leq \frac{\left(2M(c_1+M)\sqrt{q}\right)^q\mathbb{E}\left[\left(\sum_{j=1}^m\left\|p_j\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q}$$

Using Jensen's inequality and the definition of operator norm, we have

$$\frac{\left(2M(c_1+M)\sqrt{q}\right)^q\mathbb{E}\left[\left(\sum_{j=1}^m p_j^2\left\|\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q} = \frac{\left(2M(c_1+M)\sqrt{q}\right)^q\mathbb{E}\left[\left(\sum_{j=1}^m p_j\cdot p_j\left\|\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q}$$
$$\leq \frac{\left(2M(c_1+M)\sqrt{q}\right)^q\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j\Delta_j\right\|_2^q\right]}{t^q}$$
$$\leq \frac{\left(2M(c_1+M)\sqrt{q}\right)^q\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^q\right]\mathbb{E}\left[\|\Delta_j\|_2^q\right]}{t^q}$$

Lastly, we can bound this using Lemma 6.2 as follows.

$$\mathbb{P}\left(M(c_1+M)\left\|\sum_{j=1}^m p_j\hat{\Sigma}_j\Delta_j\right\|_2 > t\right) \leq \frac{K\left(2M(c_1+M)\sqrt{q}\right)^q(e\log d)^q\sum_{j=1}^m n_j p_j^{q/2+1}\mathbb{E}[\|\Delta_j\|_2^q]}{t^q} \to 0,$$

using $(\log d)^q\sum_{j=1}^m p_j^{q/2+1}n_j r_j^q \to 0$.

Step 3 For any $t \in (0,1]$, on the event A_t^c, we have that

$$B(\hat{\theta}_i^{FA}|X) = \|\Pi_i[\theta_0^\star - \theta_i^\star + E]\|_{\Sigma_i}^2$$

for some vector E where we know $\|E\|_2 \leq t$(which means $\|E\|_2 \leq t\sqrt{M}$).

Thus, we have

$$\|\Pi_i[\theta_0^\star - \theta_i^\star + E]\|_{\Sigma_i}^2 \leq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + \|\Pi_i E\|_{\Sigma_i}^2 + 2\|\Pi_i E\|_{\Sigma_i}\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}$$
$$\leq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + M^2t^2 + 2tM^{3/2}r_i^2$$
$$\|\Pi_i[\theta_0^\star - \theta_i^\star + E]\|_{\Sigma_i}^2 \geq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + \|\Pi_i E\|_{\Sigma_i}^2 - 2\|\Pi_i E\|_2\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}$$
$$\geq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 - 2tM^2r_i^2$$

Since $t \in (0,1]$, we have that $t^2 \leq t$ and thus we can choose $c = M^2 + 2M^{3/2}r_i^2$

Step 4 Reparameterizing $\varepsilon \coloneqq ct$, we have that for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|B_i(\hat{\theta}_i^{FA}|X) - \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2| \le \varepsilon) \ge \lim_{n\to\infty} \mathbb{P}(|B_i(\hat{\theta}_i^{FA}|X) - \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2| \le \varepsilon \wedge c)$$
$$\ge \lim_{n\to\infty} \mathbb{P}(A^c_{\frac{\varepsilon}{c}\wedge 1}) = 1$$

Step 5 Using Theorem 3 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2$ is given by (15) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Theorem 1 of [HMRT19] we have $B_i(\hat{\theta}_i^{FA}|X) = \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_2^2 \to r_i^2\left(1 - \frac{1}{\gamma_i}\right)$.

Variance:

We let $\xi_i = [\xi_{i,1}, \dots, \xi_{i,n}]$ denote the vector of noise.

$$V_i(\hat{\theta}_i^{FA}|X) = \operatorname{tr}(\operatorname{Cov}(\hat{\theta}_i^{FA}|X)\Sigma_i) = \mathbb{E}\left[\left\|\hat{\theta}_i^{FA} - \mathbb{E}\left[\hat{\theta}_i^{FA}|X\right]\right\|_{\Sigma_i}^2 |X\right]$$
$$= \mathbb{E}\left[\left\|\Pi_i\left[\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\sum_{j=1}^m p_j\frac{X_j^T\xi_j}{n_j}\right] + \frac{1}{n_i}\hat{\Sigma}_i^\dagger X_i^T\xi_i\right\|_{\Sigma_i}^2 |X\right]$$
$$= \underbrace{\sum_{j=1}^m \frac{p_j^2}{n_j}\operatorname{tr}\left(\Pi_i\Sigma_i\Pi_i\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\hat{\Sigma}_j\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right)\sigma_j^2}_{(i)} +$$
$$\underbrace{2\operatorname{tr}\left(\Sigma_i\hat{\Sigma}_i^\dagger\frac{X_i^TX_i}{n_i^2}\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\Pi_i\right)\sigma_i^2}_{(ii)} + \underbrace{\frac{1}{n_i^2}\operatorname{tr}\left(\hat{\Sigma}_i^\dagger X_i^TX_i\hat{\Sigma}_i^\dagger\Sigma_i\right)\sigma_i^2}_{(iii)}$$

We now study the asymptotic behavior of each of the terms $(i), (ii)$ and (iii) separately.

(i) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the fact that the nuclear norm of a projection matrix is at most d, we get

$$\sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}\operatorname{tr}\left(\Pi_i\Sigma_i\Pi_i\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\hat{\Sigma}_j\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right)$$
$$\le \sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}\|\Pi_i\|_1\|\Sigma_i\|_{\text{op}}\|\Pi_i\|_{\text{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\text{op}}\left\|\hat{\Sigma}_j\right\|_{\text{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\text{op}}$$
$$\le C_3\sigma_{\max}^2\gamma_{\max}\left(\sum_{j=1}^m p_j^2\left\|\hat{\Sigma}_j\right\|_{\text{op}}\right), \tag{20}$$

where the last inequality holds with probability going to 1 for some constant C_3 because $\mathbb{P}(B_t) \to 0$. Lastly, we show that $\mathbb{P}\left(\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} > t\right) \to 0$. Using Markov's and Jensen's inequality, we have

$$\mathbb{P}\left(\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} > t\right) \leq \frac{\mathbb{E}\left[\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}}\right]^q}{t^q} \leq \frac{\sum_{j=1}^m p_j^{q+1} \mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^q\right]}{t^q}$$

Using Lemma 6.2, we have

$$\mathbb{P}\left(\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} > t\right) \leq K \frac{\sum_{j=1}^m p_j^{q+1} (e \log d)^q n_j}{t^q}$$

Finally, since we know that $\sum_{j=1}^m p_j^{q+1} (e \log d)^q n_j \to 0$, we have $\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} \xrightarrow{p} 0$. Thus,

$$\sum_{j=1}^m \frac{p_j^2 \sigma_j^2}{n_j} \operatorname{tr}\left(\Pi_i \Sigma_i \Pi_i \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1} \hat{\Sigma}_j \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1}\right) \xrightarrow{p} 0$$

(ii) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the fact that the nuclear norm of a projection matrix is $d - n$, we get

$$\begin{aligned}\frac{2p_i\sigma^2}{n_i} \operatorname{tr}\left(\Pi_i \Sigma_i \hat{\Sigma}_i^\dagger \hat{\Sigma}_i \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1}\right) &\leq \frac{2p_i\sigma^2}{n_i} \|\Pi_i\|_1 \|\Sigma_i\|_{\text{op}} \left\|\hat{\Sigma}_i^\dagger \hat{\Sigma}_i\right\|_{\text{op}} \left\|\left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1}\right\|_{\text{op}} \\ &\leq C_4 p_i,\end{aligned}$$

where the last inequality holds with probability going to 1 for some constant C_4 because $\mathbb{P}(B_t) \to 0$ and using Assumption A2. Since $p_i \to 0$, we have

$$\frac{2p_i\sigma^2}{n_i} \operatorname{tr}\left(\Pi_i \Sigma_i \hat{\Sigma}_i^\dagger \hat{\Sigma}_i \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1}\right) \to 0$$

(iii)

$$\frac{1}{n_i^2} \operatorname{tr}(\hat{\Sigma}_i^\dagger X_i^T X_i \hat{\Sigma}_i^\dagger \Sigma_i) \sigma_i^2 = \frac{1}{n_i} \operatorname{tr}(\hat{\Sigma}_i^\dagger \Sigma_i) \sigma_i^2$$

Using Theorem 3 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\frac{\sigma_i^2}{n_i} \operatorname{tr}(\hat{\Sigma}_i^\dagger \Sigma_i)$ is given by (16) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Theorem 1 of [HMRT19] we have $V_i(\hat{\theta}_i^{FA}|X) = \frac{\sigma_i^2}{n_i} \operatorname{tr}(\hat{\Sigma}_i^\dagger) \to \frac{\sigma_i^2}{\gamma_i - 1}$.

6.5 Proof of Theorem 2

We use the global model from (6) and the personalized model from (8). The closed form of the estimators $\hat{\theta}_0^{FA}$ and $\hat{\theta}_i^R(\lambda)$ is given by

$$\begin{aligned}\hat{\theta}_0^{FA} &= \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{m} p_j \frac{1}{2n_j} \|X_j\theta - y_j\|_2 = \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \frac{X_j^T y_j}{n_j} \\ &= \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j \theta_j^\star + \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \frac{X_j^T \xi_j}{n_j}\end{aligned}$$

and

$$\begin{aligned}\hat{\theta}_i^R(\lambda) &= \underset{\theta}{\operatorname{argmin}} \frac{1}{2n_i} \|X_i\theta - y_i\|_2^2 + \frac{\lambda}{2} \left\|\hat{\theta}_0^{FA} - \theta\right\|_2^2 \\ &= (\hat{\Sigma}_i + \lambda I)^{-1} \left(\lambda \hat{\theta}_{FA} + \hat{\Sigma}_i \theta_i^\star + \frac{1}{n_i} X_i^T \xi_i\right)\end{aligned}$$

We now calculate the risk by splitting it into two parts as in (3), and then calculate the asymptotic bias and variance.

Bias:

$$\begin{aligned}B(\hat{\theta}_i^R(\lambda)|X) &:= \left\|\mathbb{E}[\hat{\theta}_i^R(\lambda)|X] - \theta_i^\star\right\|_{\Sigma_i}^2 = \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1} \left[\left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j (\theta_j^\star - \theta_i^\star)\right]\right\|_{\Sigma_i}^2 \\ &= \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1} \left[\theta_0^\star - \theta_i^\star + \left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j (\theta_j^\star - \theta_0^\star)\right]\right\|_{\Sigma_i}^2\end{aligned}$$

The idea is to show that the second term goes to 0 and use results from [HMRT19] to find the asymptotic bias. For simplicity, we let $\Delta_j := \theta_j^\star - \theta_0^\star$, and we define the event:

$$\begin{aligned}B_t &:= \left\{\left\|\left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} - \left(\sum_{j=1}^{m} p_j \Sigma_j\right)^{-1}\right\|_{\text{op}} > t\right\} \\ A_t &:= \left\{\left\|\left(\sum_{j=1}^{m} p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^{m} p_j \hat{\Sigma}_j \Delta_j\right\|_{\Sigma_i} > t\right\}\end{aligned}$$

The proof proceeds in the following steps:

Bias Proof Outline

Step 1. We first show for any $t > 0$, the $\mathbb{P}(B_t) \to 0$ as $d \to \infty$

Step 2. We show for any $t > 0$, the $\mathbb{P}(A_t) \to 0$ as $d \to \infty$

Step 3. We show that for any $t \in (0, 1]$ on event A_t^c, $B(\hat{\theta}_i^R(\lambda)|X) \leq \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + ct$ and $B(\hat{\theta}_i^R(\lambda)|X) \geq \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 - ct$

Step 4. Show that $\lim_{d\to\infty} \mathbb{P}(|B(\hat{\theta}_i^R(\lambda)|X) - \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 | \leq \varepsilon) = 1$

Step 5. Finally, using the asymptotic limit of $\lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2$ from Corollary 5 of [HMRT19], we get the result.

Step 1 and **Step 2** follow from **Step 1** and **Step 2** of proof of Theorem 1.

Step 3 For any $t \in (0, 1]$, on the event A_t^c where $T_i^{-1} = (\hat{\Sigma}_i + \lambda I)^{-1}$, we have that

$$B(\hat{\theta}_i^R(\lambda)|X) = \lambda^2 \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star + E]\right\|_{\Sigma_i}^2$$

for some vector E where we know $\|E\|_{\Sigma_i} \leq t$ (which means $\|E\|_2 \leq t\sqrt{M}$).

We can form the bounds

$$\begin{aligned}
\left\|T_i^{-1}[\theta_0^\star - \theta_i^\star + E]\right\|_{\Sigma_i}^2 &\leq \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + \left\|T_i^{-1}E\right\|_{\Sigma_i}^2 + 2\left\|T_i^{-1}E\right\|_{\Sigma_i}\left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i} \\
&\leq \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + M^2\lambda^{-2}t^2 + 2M^{3/2}t\lambda^{-2}r_i^2 \\
\left\|T_i^{-1}[\theta_0^\star - \theta_i^\star + E]\right\|_{\Sigma_i}^2 &\geq \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + \left\|T_i^{-1}E\right\|_{\Sigma_i}^2 - 2\left\|T_i^{-1}E\right\|_{\Sigma_i}\left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i} \\
&\geq \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 - 2M^2t\lambda^{-2}r_i^2.
\end{aligned}$$

Since $t \in (0, 1]$, we have that $t^2 \leq t$ and thus we can choose $c = \lambda^{-2}(M^2 + 2M^{3/2}r_i^2)$.

Step 4 Reparameterizing $\varepsilon := ct$, we have that for any $\varepsilon > 0$

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(|B(\hat{\theta}_i^R(\lambda)|X) - \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 | \leq \varepsilon) \\
&\geq \lim_{n\to\infty} \mathbb{P}(|B(\hat{\theta}_i^R(\lambda)|X) - \lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 | \leq \varepsilon \wedge c) \\
&\geq \lim_{n\to\infty} \mathbb{P}(A_{\frac{\varepsilon}{c}\wedge 1}^c) = 1.
\end{aligned}$$

Step 5 Using Theorem 6 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2$ is given by (17) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Corollary 5 of [HMRT19] we have $B_i(\hat{\theta}_i^R(\lambda)|X) = \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_2^2 \to r_i^2\lambda^2 m_i'(-\lambda)$.

Variance:

We let $\xi_i = [\xi_{i,1}, \ldots, \xi_{i,n}]$ denote the vector of noise. Substituting in the variance formula and using $\mathbb{E}[\xi_i \xi_j^T] = 0$ and $\mathbb{E}[\xi_i \xi_i^T] = \sigma^2 I$, we get

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_i^R(\lambda)|X) &= \mathbb{E}\left[\left\|(\hat{\Sigma}_i + \lambda I)^{-1}\left(\frac{1}{n_i}X_i^T\xi_i + \lambda\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\sum_{j=1}^m p_j\frac{X_j^T\xi_j}{n_j}\right)\right\|_{\Sigma_i}^2 \Bigg| X\right] \\
&= \underbrace{\sum_{j=1}^m \frac{\lambda^2 p_j^2}{n_j}\,\mathrm{tr}\left((\hat{\Sigma}_i + \lambda I)^{-1}\Sigma(\hat{\Sigma}_i + \lambda I)^{-1}\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\hat{\Sigma}_j\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right)\sigma_j^2}_{(i)} \\
&\quad + \underbrace{2\lambda p_i\,\mathrm{tr}\left(\frac{X_i^TX_i}{n_i^2}\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}(\hat{\Sigma}_i + \lambda I)^{-1}\Sigma(\hat{\Sigma}_i + \lambda I)^{-1}\right)\sigma_i^2}_{(ii)} + \underbrace{\mathrm{tr}\left((\hat{\Sigma}_i + \lambda I)^{-1}\Sigma(\hat{\Sigma}_i + \lambda I)^{-1}\hat{\Sigma}_i\right)\frac{\sigma_i^2}{n_i}}_{(iii)}
\end{aligned}$$

We now study the asymptotic behavior of each of the terms $(i), (ii)$ and (iii) separately.

(i) Using the Cauchy Schwartz inequality on Schatten $p-$norms, we get

$$\begin{aligned}
&\sum_{j=1}^m \frac{p_j^2\lambda^2\sigma_j^2}{n_j}\,\mathrm{tr}\left((\hat{\Sigma}_i + \lambda I)^{-1}\Sigma(\hat{\Sigma}_i + \lambda I)^{-1}\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\hat{\Sigma}_j\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right) \\
&\le \sum_{j=1}^m \frac{\lambda^2 p_j^2\sigma_j^2}{n_j}\|(\hat{\Sigma}_i + \lambda I)^{-1}\|_1\,\|\Sigma\|_{\mathrm{op}}\left\|(\hat{\Sigma}_i + \lambda I)^{-1}\right\|_{\mathrm{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\mathrm{op}}\left\|\hat{\Sigma}_j\right\|_{\mathrm{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\mathrm{op}} \\
&\le C_5\sigma_{\max}^2\,\gamma_{\max}\left(\sum_{j=1}^m p_j^2\left\|\hat{\Sigma}_j\right\|_{\mathrm{op}}\right),
\end{aligned}$$

where the last inequality holds with probability going to 1 for some constant C_5 because $\mathbb{P}(B_t) \to 0$. Note that this expression is same as (20) and hence the rest of the analysis for this term is same as the one in the proof of FTFA (Section 6.4).

(ii) Using the Cauchy Schwartz inequality on Schatten $p-$norms, we get

$$\begin{aligned}
&\frac{2p_i\lambda\sigma_i^2}{n_i}\,\mathrm{tr}\left((\hat{\Sigma}_i + \lambda I)^{-1}\hat{\Sigma}_i\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}(\hat{\Sigma}_i + \lambda I)^{-1}\Sigma\right) \\
&\le \frac{2p_i\lambda\sigma_i^2}{n_i}\|(\hat{\Sigma}_i + \lambda I)^{-1}\|_1\left\|\hat{\Sigma}_i\right\|_{\mathrm{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}\left\|(\hat{\Sigma}_i + \lambda I)^{-1}\right\|_{\mathrm{op}} \\
&\le \frac{C_6\sigma_i^2 dp_i}{\lambda n_i},
\end{aligned}$$

where C_6 is an absolute constant which captures an upper bound on the operator norm of the sample covariance matrix $\hat{\Sigma}_i$ using Bai Yin Theorem [BY93], and an upper bound on the operator norm of

$\left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1}$, which follows from $\mathbb{P}(B_t) \to 0$. Since $p_i \to 0$, we have
$\frac{2p_i\lambda\sigma_i^2}{n_i} \operatorname{tr}\left((\hat{\Sigma}_i + \lambda I)^{-1}\hat{\Sigma}_i \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1} (\hat{\Sigma}_i + \lambda I)^{-1}\Sigma\right) \xrightarrow{p} 0$

(iii) Using Theorem 3 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\operatorname{tr}((\hat{\Sigma}_i + \lambda I)^{-2}\hat{\Sigma}_i\Sigma_i)\frac{\sigma_i^2}{n_i}$ is given by (18) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Theorem 1 of [HMRT19] we have $V_i(\hat{\theta}_i^R(\lambda)|X) = \frac{\sigma_i^2}{n_i}\operatorname{tr}((\hat{\Sigma}_i + \lambda I)^{-2}\hat{\Sigma}_i^\dagger\Sigma_i) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}$.

6.6 Proof of Theorem 3

On solving (12) and (13), the closed form of the estimators $\hat{\theta}_0^M(\alpha)$ and $\hat{\theta}_i^M(\alpha)$ is given by

$$\begin{aligned}
\hat{\theta}_0^M(\alpha) &:= \operatorname*{argmin}_\theta \sum_{j=1}^m \frac{p_j}{2n_j}\left\|X_j\left[\theta - \frac{\alpha}{n_j}X_j^T(X_j\theta - y_j)\right] - y_j\right\|_2^2 \\
&= \operatorname*{argmin}_\theta \sum_{j=1}^m \frac{p_j}{2n_j}\left\|\left(I_n - \frac{\alpha}{n}X_jX_j^T\right)(X_j\theta - y_j)\right\|_2^2 \\
&= \left(\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2y_j
\end{aligned}$$

where $W_j := I - \frac{\alpha}{n_j}X_jX_j^T$ and

$$\begin{aligned}
\hat{\theta}_i^M(\alpha) &:= \operatorname*{argmin}_\theta \left\|\hat{\theta}_0^M(\alpha) - \theta\right\|_2 \quad s.t. \quad X_i\theta = y_i \\
&= (I - \hat{\Sigma}_i^\dagger\hat{\Sigma}_i)\hat{\theta}_0^M(\alpha) + \hat{\Sigma}_i^\dagger\hat{\Sigma}_i\theta_i^\star + \frac{1}{n_i}\hat{\Sigma}_i^\dagger X_i^T\xi_i
\end{aligned}$$

We now calculate the risk by splitting it into two parts as in (3), and then calculate the asymptotic bias and variance.

Bias:

$$\begin{aligned}
B(\hat{\theta}_i^M(\alpha)|X) &:= \left\|\mathbb{E}[\hat{\theta}_i^M(\alpha)|X] - \theta_i^\star\right\|_{\Sigma_i}^2 = \left\|\Pi_i\left[\left(\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j(\theta_j^\star - \theta_i^\star)\right]\right\|_2^2 \\
&= \left\|\Pi_i\left[\theta_0^\star - \theta_i^\star + \left(\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j(\theta_j^\star - \theta_0^\star)\right]\right\|_{\Sigma_i}^2
\end{aligned}$$

For simplicity, we let $\Delta_j := \theta_j^\star - \theta_0^\star$, and we define the events:

$$B_t := \left\{\left\|\left(\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1} - \mathbb{E}\left[\sum_{j=1}^m p_j\frac{1}{n_j}X_j^TW_j^2X_j\right]^{-1}\right\|_{\text{op}} > t\right\} \tag{21}$$

$$A_t := \left\{\left\|\left(\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\Delta_j\right\|_{\Sigma_i} > t\right\} \tag{22}$$

The proof proceeds in the following steps:

Bias Proof Outline

Step 1. We first show for any $t > 0$, the $\mathbb{P}(B_t) \to 0$ as $d \to \infty$

Step 2. Then, we show for any $t > 0$, the $\mathbb{P}(A_t) \to 0$ as $d \to \infty$

Step 3. We show that for any $t \in (0, 1]$[1] on event A_t^c, $B(\hat{\theta}_i^M(\alpha)|X) \leq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + ct$ and $B(\hat{\theta}_i^M(\alpha)|X) \geq \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 - ct$

Step 4. Show that $\lim_{d\to\infty} \mathbb{P}(|B(\hat{\theta}_i^M(\alpha)|X) - \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2| \leq \varepsilon) = 1$

Step 5. Finally, using the asymptotic limit of $\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2$ from Theorem 1 of [HMRT19], we get the result.

We now give the detailed proof:

Step 1 Since $\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n_j} X_j^T W_j^2 X_j\right]\right) \geq \lambda_0$, it suffices to show by Lemma 6.4 that the probability of

$$C_t := \left\{ \left\| \sum_{j=1}^m p_j \left(\frac{1}{n_j} X_j^T W_j^2 X_j - \mathbb{E}\left[\frac{1}{n_j} X_j^T W_j^2 X_j\right] \right) \right\|_{\text{op}} > t \right\}$$

goes to 0 as $d, m \to \infty$ under Assumption A1.

$$\begin{aligned}
\mathbb{P}(C_t) &= \mathbb{P}\left(\left\| \sum_{j=1}^m p_j \left[\left[\hat{\Sigma}_j - 2\alpha^2 \hat{\Sigma}_j^2 + \alpha^2 \hat{\Sigma}_j^3 \right] - \mathbb{E}\left[\frac{1}{n_j} X_j^T W_j^2 X_j\right] \right] \right\|_{\text{op}} > t \right) \\
&\leq \mathbb{P}\left(\left\| \sum_{j=1}^m p_j \left(\hat{\Sigma}_j - \mu_{1,j} \right) \right\|_{\text{op}} > t/3 \right) \\
&+ \mathbb{P}\left(\left\| 2\alpha \sum_{j=1}^m p_j \left(\hat{\Sigma}_j^2 - \mu_{2,j} \right) \right\|_{\text{op}} > t/3 \right) + \mathbb{P}\left(\left\| \alpha^2 \sum_{j=1}^m p_j \left(\hat{\Sigma}_j^3 - \mu_{3,j} \right) \right\|_{\text{op}} > t/3 \right), \qquad (23)
\end{aligned}$$

where $\mu_{p,j} := \mathbb{E}[\hat{\Sigma}_j^p]$. We repeatedly apply Lemma 6.3 for $p = 1, 2, 3$ to bound each of these three terms. It is clear that if $\mathbb{E}[\|\mathbf{x}_{j,k}\|_2^{6q}]^{1/(6q)} \lesssim \sqrt{d}$, $\left\|\mathbb{E}[\hat{\Sigma}_j^6]\right\|_{\text{op}} \leq C_3$ for some constant C_3, $(\log d)^{4q} \sum_{j=1}^m p_j^q n_j \to 0$, and $(\log d)^{q/2} \sum_{j=1}^m p_j^{q/2+1} \to 0$, then (23) goes to 0.

Step 2

$$\begin{aligned}
\mathbb{P}(A_t) &\leq \mathbb{P}(A_t \cap B_{c_1}^c) + \mathbb{P}(B_{c_1}) \\
&\leq \mathbb{P}\left(M\left(c_1 + \frac{1}{\lambda_0}\right) \left\| \sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \Delta_j \right\|_2 > t \right) + \mathbb{P}(B_{c_1})
\end{aligned}$$

From Step 1, we know that $\lim_{n\to\infty}\mathbb{P}(B_{c_1})=0$. The second inequality comes from the fact that $\|Ax\|_2 \le |\!|\!|A|\!|\!|_{\text{op}}\|x\|_2$. To handle the first term, we use Markov's inequality.

$$\begin{aligned}
\mathbb{P}\left(c_2\left\|\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\Delta_j\right\|_2 > t\right) &\le \frac{c_2^q}{t^q}\mathbb{E}\left[\left\|\sum_{j=1}^m \frac{p_j}{n_j}X_j^TW_j^2X_j\Delta_j\right\|_2^q\right]\\
&\le \frac{(2c_2\sqrt{q})^q}{t^q}\mathbb{E}\left[\left(\sum_{j=1}^m\left\|\frac{p_j}{n_j}X_j^TW_j^2X_j\Delta_j\right\|_2^2\right)^{q/2}\right]\\
&\le \frac{(2c_2\sqrt{q})^q}{t^q}\sum_{j=1}^m p_j\mathbb{E}\left[\left(p_j\left\|\frac{1}{n_j}X_j^TW_j^2X_j\Delta_j\right\|_2^2\right)^{q/2}\right]\\
&= \frac{(2c_2\sqrt{q})^q}{t^q}\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^2\Delta_j\right\|_2^q\right]\\
&\le \frac{(8c_2\sqrt{q})^q}{2t^q}\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[|\!|\!|\hat{\Sigma}_j|\!|\!|_{\text{op}}^q+\alpha^2|\!|\!|\hat{\Sigma}_j|\!|\!|_{\text{op}}^{3q}\right]\mathbb{E}[\|\Delta_j\|_2^q]\\
&\le \frac{(8c_2\sqrt{q})^q}{2t^q}\sum_{j=1}^m p_j^{q/2+1}\left[\alpha^2K(e\log d)^{3q}n_j\right]r_j^q,
\end{aligned}$$

where the last step follows from Lemma 6.2 and the final expression goes to 0 since $(\log d)^{3q}\sum_{j=1}^m p_j^{q/2+1}n_jr_j^q\to 0$.

Step 3, 4 and 5 are same as the bias calculation of proof of Theorem 1.

Variance:

We let $\xi_i=[\xi_{i,1},\dots,\xi_{i,n}]$ denote the vector of noise.

$$\begin{aligned}
V_i(\hat{\theta}_i^M(\alpha);\theta_i^\star|X) &= \text{tr}(\text{Cov}(\hat{\theta}_i^M(\alpha)|X)\Sigma) = \mathbb{E}[\left\|\hat{\theta}_i^M(\alpha)-\mathbb{E}[\hat{\theta}_i^M(\alpha)|X]\right\|_{\Sigma_i}^2|X]\\
&= \mathbb{E}[\left\|\Pi_i\left[\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2\xi_j\right]+\frac{1}{n_i}\hat{\Sigma}_i^\dagger X_i^T\xi_i\right\|_{\Sigma_i}^2|X]\\
&= \underbrace{\sum_{j=1}^m\frac{p_j^2}{n_j^2}\,\text{tr}\left(\Pi_i\Sigma_i\Pi_i\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}X_j^TW_j^4X_j\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\right)\sigma_j^2}_{(i)}+\\
&\underbrace{2\,\text{tr}\left(\hat{\Sigma}_i^\dagger\frac{X_i^TW_j^2X_i}{n_i^2}\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\Pi_i\Sigma_i\right)\sigma_i^2}_{(ii)}+\underbrace{\frac{1}{n_i^2}\,\text{tr}(\hat{\Sigma}_i^\dagger X_i^TX_i\hat{\Sigma}_i^\dagger\Sigma_i)\sigma_i^2}_{(iii)}
\end{aligned}$$

We now study the asymptotic behavior of each of the terms $(i),(ii)$ and (iii) separately.

(i) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the fact that the nuclear norm of a projection matrix is at most d, we get

$$\sum_{j=1}^{m} \frac{p_j^2\sigma_j^2}{n_j} \operatorname{tr}\left(\Pi_i\Sigma_i\Pi_i \left(\sum_{j=1}^{m} \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1} \hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4 \left(\sum_{j=1}^{m} \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1}\right)$$

$$\leq \sum_{j=1}^{m} \frac{p_j^2\sigma_j^2}{n_j} \|\Pi_i\|_1 \|\Sigma_i\|_{\text{op}} \|\Pi_i\|_{\text{op}} \left\|\left(\sum_{j=1}^{m} \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1}\right\|_{\text{op}} \left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}} \left\|\left(\sum_{j=1}^{m} \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1}\right\|_{\text{op}}$$

$$\leq C_7\sigma_{\max}^2\gamma_{\max}\left(\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}}\right),$$

where the last inequality holds with probability going to 1 for some constant C_7 because $\mathbb{P}(C_t) \to 0$. Lastly, we show that $\mathbb{P}\left(\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}} > t\right) \to 0$. Using Markov's and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j(I-\hat{\Sigma}_j)^4\right\|_{\text{op}} > t\right) &\leq \frac{\mathbb{E}\left[\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}}\right]^q}{t^q} \\
&\leq \frac{\sum_{j=1}^{m} p_j^{q+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}}^q\right]}{t^q} \\
&\leq \frac{\sum_{j=1}^{m} p_j^{q+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}} \left\|(I-\alpha\hat{\Sigma}_j)\right\|_{\text{op}}^{4q}\right]}{t^q} \\
&\leq \frac{\sum_{j=1}^{m} p_j^{q+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}} \left(\|I\|_{\text{op}} + \alpha\left\|\hat{\Sigma}_j\right\|_{\text{op}}\right)^{4q}\right]}{t^q} \\
&\leq 2^{4q-1}\frac{\sum_{j=1}^{m} p_j^{q+1}\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}} + \alpha^4\left\|\hat{\Sigma}_j\right\|_{\text{op}}^{5q}\right]}{t^q}
\end{aligned}$$

Using Lemma 6.2 and Markov's inequality, we have

$$\mathbb{P}\left(\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} > t\right) \leq 2^{4q-1}\left(K_1\frac{\sum_{j=1}^{m} p_j^{q+1}(e\log d)n_j}{t^q} + K_2\alpha^4\frac{\sum_{j=1}^{m} p_j^{q+1}(e\log d)^{5q}n_j}{t^q}\right).$$

Finally, since we know that $\sum_{j=1}^{m} p_j^{q+1}(\log d)^{5q}n_j \to 0$, we have $\mathbb{P}\left(\sum_{j=1}^{m} p_j^2 \left\|\hat{\Sigma}_j(I-\alpha\hat{\Sigma}_j)^4\right\|_{\text{op}} > t\right) \to 0$.

(ii) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the fact that the nuclear norm of a projection matrix is $d-n$, we get

$$
\begin{aligned}
&2\,\mathrm{tr}\left(\hat{\Sigma}_i^\dagger \frac{X_i^T W_i^2 X_i}{n_i^2}\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1} \Pi_i \Sigma_i\right)\sigma_i^2 \\
&= 2\,\mathrm{tr}\left(\Pi_i \Sigma_i \hat{\Sigma}_i^\dagger \frac{\hat{\Sigma}_i (I-\hat{\Sigma}_i)^2}{n_i}\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1}\right)\sigma_i^2 \\
&\le \frac{2p_i\sigma^2}{n_i} \|\Pi_i\|_1 \|\Sigma_i\|_{\mathrm{op}} \left\|\hat{\Sigma}_i^\dagger \hat{\Sigma}_i\right\|_{\mathrm{op}} \left\|(I-\hat{\Sigma}_i)^2\right\|_{\mathrm{op}} \left\|\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1}\right\|_{\mathrm{op}} \\
&\le C_4 p_i,
\end{aligned}
$$

where the last inequality holds with probability going to 1 for some constant C_4 because $\mathbb{P}(B_t) \to 0$ and using Assumption A2. Since $p_i \to 0$, we have

$$
2\,\mathrm{tr}\left(\hat{\Sigma}_i^\dagger \frac{X_i^T W_i^2 X_i}{n_i^2}\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1} \Pi_i \Sigma_i\right)\sigma_i^2 \to 0
$$

(iii)

$$
\frac{1}{n_i^2}\,\mathrm{tr}(\hat{\Sigma}_i^\dagger X_i^T X_i \hat{\Sigma}_i^\dagger \Sigma_i)\sigma_i^2 = \frac{1}{n_i}\,\mathrm{tr}(\hat{\Sigma}_i^\dagger \Sigma_i)\sigma_i^2
$$

Using Theorem 3 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\frac{\sigma_i^2}{n_i}\,\mathrm{tr}(\hat{\Sigma}_i^\dagger \Sigma_i)$ is given by (16) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Theorem 1 of [HMRT19] we have $V_i(\hat{\theta}_i^M(\alpha)|X) = \frac{\sigma_i^2}{n_i}\,\mathrm{tr}(\hat{\Sigma}_i^\dagger) \to \frac{\sigma_i^2}{\gamma_i - 1}$.

6.7 Proof of Theorem 4

The solution to this minimization problem in (14) is given by

$$
\hat{\theta}_0^P(\lambda) = \theta_0^\star + Q^{-1}\left(\sum_{j=1}^m p_j T_j^{-1} \hat{\Sigma}_j \Delta_j + \sum_{j=1}^m p_j T_j^{-1} \frac{1}{n_j} X_j^T \xi_j\right),
$$

where $\Delta_j = \theta_j^\star - \theta_0^\star$, $T_j = \hat{\Sigma}_j + \lambda I$ and $Q = I - \lambda \sum_{j=1}^m p_j T_j^{-1}$. The personalized solutions are then given by

$$
\hat{\theta}_i^P(\lambda) = T_i^{-1}\left(\lambda \hat{\theta}_0^P(\lambda) + \hat{\Sigma}_i \theta_i^\star + \frac{1}{n_i} X_i^T \xi_i\right)
$$

We now calculate the risk by splitting it into two parts as in (3), and then calculate the asymptotic bias and variance.

Bias:

Let $\Delta_j := \theta_j^\star - \theta_0^\star$, then we have

$$
B(\hat{\theta}_i^P(\lambda)|X) := \left\| T_i^{-1}\left(\lambda\theta_0^\star - \lambda\theta_i^\star + \lambda Q^{-1}\left[\sum_{j=1}^m p_j T_j^{-1} \hat{\Sigma}_j \Delta_j\right]\right)\right\|_{\Sigma_i}^2
$$

The idea is to show that the second term goes to 0 and use results from [HMRT19] to find the asymptotic bias. To do this, we first define the events:

$$C_t := \left\{ \left\| \sum_{j=1}^{m} p_j (T_j^{-1} - \mathbb{E}[T_j^{-1}]) \right\|_{\text{op}} > t \right\}$$
$$A_t := \left\{ \left\| Q^{-1} \left[\sum_{j=1}^{m} p_j T_j^{-1} \hat{\Sigma}_j \Delta_j \right] \right\|_{\Sigma_i} > t \right\}$$

The proof proceeds in the following steps:

Bias Proof Outline

Step 1. We first show for any $t > 0$, the $\mathbb{P}(C_t) \to 0$ as $d \to \infty$

Step 2. Then, we show for any $t > 0$, the $\mathbb{P}(A_t) \to 0$ as $d \to \infty$.

Step 3. We show that for any $t \in (0, 1]$, $B(\hat{\theta}_i^P(\lambda)|X) \leq \left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star]\right\|_2^2 + ct$ and $B(\theta, X) \geq \left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star]\right\|_2^2 - ct$

Step 4. Show that $\lim_{d\to\infty} \mathbb{P}(|B(\hat{\theta}_i^P(\lambda)|X) - \left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star]\right\|_2^2| \leq \varepsilon) = 1$

Step 5. Finally, using the asymptotic limit of $\left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star]\right\|_2^2$ from Corollary 5 of [HMRT19], we get the result.

We now give the detailed proof:

Step 1

$$\mathbb{P}(C_t) = \mathbb{P}\left(\left\| \sum_{j=1}^{m} p_j (T_j^{-1} - \mathbb{E}[T_j^{-1}]) \right\|_{\text{op}} > t \right) \leq \frac{2^q \mathbb{E}\left[\left\| \sum_{j=1}^{m} \xi_j p_j T_j^{-1} \right\|_{\text{op}}^q \right]}{t^q},$$

We use Theorem A.1 from [CGT12] to bound this object.

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{j=1}^{m} \xi_j T_j^{-1} \right\|_{\text{op}}^q \right] &\leq \left[\sqrt{e \log d} \left\| \left(\sum_{j=1}^{m} p_j^2 \mathbb{E}[T_j^{-1}] \right)^{1/2} \right\|_{\text{op}} + (e \log d)(\mathbb{E} \max_j \left\| p_j T_j^{-1} \right\|_{\text{op}}^q)^{1/q} \right]^q \\
&\leq 2^{q-1} \left(\sqrt{e \log d}^q \left\| \left(\sum_{j=1}^{m} p_j^2 \mathbb{E}[(T_j^{-1})^2] \right)^{1/2} \right\|_{\text{op}}^q + (e \log d)^q (\mathbb{E} \max_j \left\| p_j T_j^{-1} \right\|_{\text{op}}^q) \right) \\
&\leq 2^{q-1} \left(\sqrt{e \log d}^q \left\| \sum_{j=1}^{m} p_j^2 \mathbb{E}[(T_j^{-1})^2] \right\|_{\text{op}}^{q/2} + \frac{(e \log d)^q \max_j p_j^q}{\lambda^q} \right) \\
&\leq 2^{q-1} \left(\sqrt{e \log d}^q \sum_{j=1}^{m} p_j^{q/2+1} \left\| (\mathbb{E}[(T_j^{-1})^2])^{q/2} \right\|_{\text{op}} + \frac{1}{\lambda^q} (e \log d)^q \sum_{j=1}^{m} p_j^q \right) \\
&\leq \frac{2^{q-1}}{\lambda^q} \left((e \log d)^{q/2} \sum_{j=1}^{m} p_j^{q/2+1} + (e \log d)^q \sum_{j=1}^{m} p_j^q \right),
\end{aligned}$$

where we use the fact that $\left\|\left\|T_j^{-1}\right\|\right\|_{\text{op}} = \left\|\left\|(\hat{\Sigma}_j + \lambda I)^{-1}\right\|\right\|_{\text{op}} \leq \frac{1}{\lambda}$ since $\hat{\Sigma}_j$ is always positive semidefinite. Since $(\log d)^{q/2} \sum_{j=1}^m p_j^{q/2+1}$ and $(\log d)^q \sum_{j=1}^m p_j^q$, we get that $\mathbb{P}(C_t) \to 0$ for all $t > 0$.

Step 2 To prove this step, we will first use a helpful lemma,

Lemma 6.5. *Suppose that $\Sigma = \mathbb{E}[\hat{\Sigma}] \in \mathbb{R}^{d,d}$ has a spectrum supported on $[a, b]$ where $0 < a < b < \infty$. Further suppose that $\mathbb{E}\left[\left\|\left\|\hat{\Sigma}^2\right\|\right\|_{\text{op}}\right] \leq \tau$ and there exists an $R \geq b$ such that $\mathbb{P}(\lambda_{\max}(\hat{\Sigma}) > R) \leq \frac{a^2}{8\tau}$, then*

$$\left\|\left\|\mathbb{E}[(\hat{\Sigma} + \lambda I)^{-1}]\right\|\right\|_{\text{op}} \leq \frac{1}{\lambda}\left(1 - \frac{a^3}{16\tau(R+\lambda)}\right) \leq \frac{1}{\lambda}$$

Proof Fix an arbitrary vector $u \in \mathbb{R}^d$ with unit ℓ_2 norm. We fix $\delta = a/2 > 0$, we define the event $A \coloneqq \{u^T\hat{\Sigma}u \geq \delta\}$ and $B \coloneqq \{\lambda_{\max}(\hat{\Sigma}) \leq R\}$

$$u^T\mathbb{E}[(\hat{\Sigma} + \lambda I)^{-1}]u \leq \mathbb{E}[\mathbf{1}\{A \cap B\}\, u^T(\hat{\Sigma} + \lambda I)^{-1}u] + \frac{1}{\lambda}(1 - \mathbb{P}(A \cap B))$$

Let σ_i^2 and v_i denote the ith eigenvalue and eigenvector of $\hat{\Sigma}$ respectively sorted in descending order with respect to eigenvalue ($\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2$). On the event A, we have that $u^T(\hat{\Sigma} + \lambda I)^{-1}u$ has value no larger than

$$\begin{aligned}
&\max_{\alpha \in \mathbb{R}^d} \sum_{i=1}^d \frac{1}{\sigma_i^2 + \lambda}\alpha_i \\
&\text{s.t. } \alpha \geq 0 \\
&\quad \mathbf{1}^T\alpha = 1 \\
&\quad \sum_{i=1}^d \sigma_i^2\alpha_i \geq \delta
\end{aligned}$$

The dual of this problem is

$$\begin{aligned}
&\min_{\theta}\ \max_{j \in [d]} \left\{\theta\sigma_j^2 + \frac{1}{\sigma_j^2 + \lambda}\right\} - \theta\delta \\
&\text{s.t. } \theta \geq 0
\end{aligned}$$

It suffices to demonstrate that there exists a θ which satisfies the constraints of the dual and has objective value less than $\frac{1}{\lambda}$. We can verify that selecting $\theta = \frac{1}{\lambda(\sigma_1^2 + \lambda)}$ has an objective value of

$$\frac{1}{\lambda} - \frac{\delta}{\lambda(\sigma_1^2 + \lambda)}$$

which is less than the desired $\frac{1}{\lambda}$. All that remains is to lower bound $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$. We know by Paley-Zygmund

$$\mathbb{P}(A) \geq \mathbb{P}\left(u^T\hat{\Sigma}u \geq \frac{\delta}{a}u^T\Sigma u\right) \geq \mathbb{P}\left(u^T\hat{\Sigma}u \geq \frac{1}{2}u^T\Sigma u\right) \geq \frac{(u^T\Sigma u)^2}{4\mathbb{E}[(u^T\hat{\Sigma}u)^2]} \geq \frac{a^2}{4\tau}$$

Note that $a^2/4\tau < 1$ because the second moment of a random variable is no smaller than the first moment squared of the random variable. Moreover, by construction, R is large enough such that $\mathbb{P}(B^c) \leq \mathbb{P}(A)/2$, thus,

$$\begin{aligned}
u^T\mathbb{E}[(\hat{\Sigma} + \lambda I)^{-1}]u &\leq \frac{1}{\lambda}\left(1 - \frac{a}{2(R+\lambda)}\right)\frac{a^2}{8\tau} + \frac{1}{\lambda}\left(1 - \frac{a^2}{8\tau}\right) \\
&= \frac{1}{\lambda}\left(1 - \frac{a^3}{16\tau(R+\lambda)}\right)
\end{aligned}$$

□

Recall that we have the assumptions that for sufficiently large m, for all $j \in [m]$ we have Σ_j has a spectrum supported on $[a, b]$ where $a = 1/M$ and $b = M$ and $\mathbb{E}\left[\left\|\hat{\Sigma}_j^2\right\|_{\text{op}}\right] \leq \tau_3$. Moreover, since we have the assumption that there exists an $R \geq b$ such that $\limsup_{m\to\infty} \sup_{j\in[m]} \mathbb{P}(\lambda_{\max}(\hat{\Sigma}_j) > R) \leq \frac{a^2}{16\tau_3}$, by Lemma 6.5 there exists and $1 > \varepsilon > 0$ such that for sufficiently large m, for all $j \in [m]$, $\left\|\mathbb{E}[(\hat{\Sigma}_j + \lambda I)^{-1}]\right\|_{\text{op}} \leq \frac{1-\varepsilon}{\lambda}$.

$$\mathbb{P}(A_t) \leq \mathbb{P}(A_t \cap C_{c_1}^c) + \mathbb{P}(C_{c_1})$$

Since we know $\mathbb{P}(C_{c_1}) \to 0$, it suffices to bound the first term.

$$\begin{aligned}
\mathbb{P}(A_t \cap C_{c_1}^c) &\leq \mathbb{P}\left(\sqrt{M}\left\|Q^{-1}\right\|_{\text{op}}\left\|\left[\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right]\right\|_2 > t \cap C_{c_1}^c\right) \\
&= \mathbb{P}\left(\sqrt{M}\left(1 - \left\|\lambda\sum_{j=1}^m p_j T_j^{-1}\right\|_{\text{op}}\right)^{-1}\left\|\left[\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right]\right\|_2 > t \cap C_{c_1}^c\right) \\
&\leq \mathbb{P}\left(\sqrt{M}\left(1 - \left\|\lambda\sum_{j=1}^m p_j \mathbb{E}[T_j^{-1}] + E_{c_1}\right\|_{\text{op}}\right)^{-1}\left\|\left[\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right]\right\|_2 > t\right) \\
&\leq \mathbb{P}\left(\sqrt{M}\left(1 - \lambda\sum_{j=1}^m p_j \left\|\mathbb{E}[T_j^{-1}]\right\|_{\text{op}} - c_1\right)^{-1}\left\|\left[\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right]\right\|_2 > t\right)
\end{aligned}$$

where we used Jensen's inequality in the last step. E_{c_1} is a matrix error term which on the event $C_{c_1}^c$ has operator norm bounded by c_1. As discussed, we have that $\sum_{j=1}^m p_j \left\|\mathbb{E}[T_j^{-1}]\right\|_{\text{op}}$ is less than $\frac{1-\varepsilon}{\lambda}$, which shows there exists a constant c_2, such that $\sqrt{M}(1 - \lambda\sum_{j=1}^m p_j \left\|\mathbb{E}[T_j^{-1}]\right\|_{\text{op}} - c_1)^{-1} < c_2$. Now, we have, using Lemma 6.1,

$$\mathbb{P}\left(c_2\left\|\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2 > t\right) \leq \frac{c_2^q\mathbb{E}\left[\left\|\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2^q\right]}{t^q} \leq \frac{(2c_2\sqrt{q})^q\mathbb{E}\left[\left(\sum_{j=1}^m \left\|p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q}$$

Using Jensen's inequality and the definition of operator norm, we have

$$\begin{aligned}
\frac{(2c_2\sqrt{q})^q\mathbb{E}\left[\left(\sum_{j=1}^m p_j^2\left\|T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q} &= \frac{(2c_2\sqrt{q})^q\mathbb{E}\left[\left(\sum_{j=1}^m p_j \cdot p_j\left\|T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2^2\right)^{q/2}\right]}{t^q} \\
&\leq \frac{(2c_2\sqrt{q})^q\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[\left\|T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2^q\right]}{t^q} \\
&\leq \frac{(2c_2\sqrt{q})^q\sum_{j=1}^m p_j^{q/2+1}\mathbb{E}\left[\left\|T_j^{-1}\right\|_{\text{op}}^q\right]\mathbb{E}\left[\left\|\hat{\Sigma}_j\right\|_{\text{op}}^q\right]\mathbb{E}\left[\|\Delta_j\|_2^q\right]}{t^q}
\end{aligned}$$

Lastly, we can bound this using Lemma 6.2 as follows and using the fact that $\left\|T_j^{-1}\right\|_{\text{op}} \leq \frac{1}{\lambda}$.

$$\mathbb{P}\left(c_2\left\|\sum_{j=1}^m p_j T_j^{-1}\hat{\Sigma}_j\Delta_j\right\|_2 > t\right) \leq \frac{(2c_2\sqrt{q})^q\sum_{j=1}^m p_j^{q/2+1}Kn_j(e\log d)^q r_j^q}{\lambda^q t^q} \to 0,$$

using $(\log d)^q \sum_{j=1}^m n_j p_j^{q/2+1} \to 0$

Step 3 For any $t \in (0,1]$, on the event A_t^c, we have that

$$B(\hat{\theta}_i^P(\lambda)|X) = \left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star + E]\right\|_{\Sigma_i}^2$$

for some vector E where we know $\|E\|_{\Sigma_i} \le t$ (which means $\|E\|_2 \le t\sqrt{M}$).

We can form the bounds

$$\begin{aligned}
\left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star + E]\right\|_{\Sigma_i}^2 &\le \lambda^2 \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + \left\|T_i^{-1}E\right\|_{\Sigma_i}^2 + 2\lambda \left\|T_i^{-1}E\right\|_{\Sigma_i} \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i} \\
&\le \lambda^2 \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + \lambda^{-2}t^2M^2 + 2t\lambda^{-1}r_i^2M^{3/2} \\
\left\|T_i^{-1}[\lambda\theta_0^\star - \lambda\theta_i^\star + E]\right\|_{\Sigma_i}^2 &\ge \lambda^2 \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 + \left\|T_i^{-1}E\right\|_{\Sigma_i}^2 - 2\lambda \left\|T_i^{-1}E\right\|_{\Sigma_i} \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i} \\
&\ge \lambda^2 \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2 - 2t\lambda^{-1}r_i^2M^{3/2}
\end{aligned}$$

Since $t \in (0,1]$, we have that $t^2 \le t$ and thus we can choose $c = \lambda^{-2}M^2 + 2r_i^2\lambda^{-1}M^{3/2}$

Step 4 Reparameterizing $\varepsilon := ct$, we have that for any $\varepsilon > 0$

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(|B(\hat{\theta}_i^P(\lambda), X) - \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2| \le \varepsilon) &\ge \lim_{n\to\infty} \mathbb{P}(|B(\hat{\theta}_i^P(\lambda)|X) - \left\|T_i^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2| \le \varepsilon \wedge c) \\
&\ge \lim_{n\to\infty} \mathbb{P}(A^c_{\frac{\varepsilon}{c}\wedge 1}) = 1
\end{aligned}$$

Step 5 Using Theorem 6 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\lambda^2 \left\|(\hat{\Sigma}_i + \lambda I)^{-1}[\theta_0^\star - \theta_i^\star]\right\|_{\Sigma_i}^2$ is given by (17) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Corollary 5 of [HMRT19] we have $B_i(\hat{\theta}_i^P(\lambda)|X) = \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_2^2 \to r_i^2\lambda^2 m_i'(-\lambda)$.

Variance

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_i^P(\lambda)|X) &= \mathbb{E}\left[\left\|T_i^{-1}\left(\lambda Q^{-1}\left[\sum_{j=1}^m p_j T_j^{-1}\frac{1}{n}X_j^T\xi_j\right] + \frac{1}{n_i}X_i^T\xi_i\right)\right\|_{\Sigma_i}^2\right] \\
&= \underbrace{\sum_{j=1}^m \frac{\lambda^2 p_j^2}{n_j}\,\mathrm{tr}\left(T_i^{-1}\Sigma_i T_i^{-1}Q^{-1}T_j\hat{\Sigma}_j T_j Q^{-1}\right)\sigma_j^2}_{(i)} \\
&\quad + \underbrace{\frac{2\lambda\sigma_i^2 p_i}{n_i}2\,\mathrm{tr}\left(\Sigma_i T_i^{-1}Q^{-1}T_i^{-1}\hat{\Sigma}_i T_i^{-1}\right)}_{(ii)} \\
&\quad + \underbrace{\mathrm{tr}\left(T_i^{-1}\Sigma_i T_i^{-1}\hat{\Sigma}_i\right)\frac{\sigma_i^2}{n_i}}_{(iii)}
\end{aligned}$$

We now study the asymptotic behavior of each of the terms $(i), (ii)$ and (iii) separately. In these steps, we will have to bound $\left\|Q^{-1}\right\|_{\mathrm{op}}$. To do this, we observe that there exists a sufficiently large constant t

such that the following statement is true.

$$\begin{aligned}
\mathbb{P}(\left\|\left\|Q^{-1}\right\|\right\|_{\text{op}} > t) &= \mathbb{P}(\left\|\left\|Q^{-1}\right\|\right\|_{\text{op}} > t \cap C_{c_1}^c) + \mathbb{P}(C_{c_1}) \\
&= \mathbb{P}\left(\left(1 - \left\|\left\|\lambda \sum_{j=1}^m p_j T_j^{-1}\right\|\right\|_{\text{op}}\right)^{-1} > t \cap C_{c_1}^c\right) + o(1) \\
&\leq \mathbb{P}\left(\left(1 - \left\|\left\|\lambda \sum_{j=1}^m p_j \mathbb{E}[T_j^{-1}] + E_{c_1}\right\|\right\|_{\text{op}}\right)^{-1} > t\right) + o(1) \\
&\leq \mathbb{P}\left(\left(1 - \lambda \sum_{j=1}^m p_j \left\|\left\|\mathbb{E}[T_j^{-1}]\right\|\right\|_{\text{op}} - c_1\right)^{-1} > t\right) + o(1) \\
&\leq o(1).
\end{aligned}$$

This is true because of Lemma 6.5.

(i) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the high probability bounds from the bias proof, we get that for some constant C_8, the following holds with probability going to 1.

$$\begin{aligned}
&\sum_{j=1}^m \frac{\lambda^2 p_j^2}{n_j} \operatorname{tr}\left(T_i^{-1}\Sigma_i T_i^{-1} Q^{-1} T_j \hat{\Sigma}_j T_j Q^{-1}\right) \sigma_j^2 \\
&\leq \sum_{j=1}^m \frac{\sigma_j^2 \lambda^2 p_j^2}{n_j} \|T_i^{-1}\|_1 \left\|\left\|T_i^{-1}\right\|\right\|_{\text{op}} \|\Sigma_i\|_{\text{op}} \left\|\left\|Q^{-1}\right\|\right\|_{\text{op}} \|T_j\|_{\text{op}} \left\|\left\|\hat{\Sigma}_j\right\|\right\|_{\text{op}} \|T_j\|_{\text{op}} \left\|\left\|Q^{-1}\right\|\right\|_{\text{op}} \\
&\leq \sum_{j=1}^m \frac{M C_8 \sigma_j^2 \lambda^2 p_j^2 d}{n_j} \left\|\left\|\hat{\Sigma}_j\right\|\right\|_{\text{op}} \\
&\leq \gamma_{\max} M C_8 \sigma_{\max}^2 \lambda^2 \sum_{j=1}^m p_j^2 \left\|\left\|\hat{\Sigma}_j\right\|\right\|_{\text{op}}
\end{aligned}$$

$\left\|\left\|Q^{-1}\right\|\right\|_{\text{op}}$ is upper bounded by some constant as shown above. We use the same technique as in proof of the variance of Theorem 1 calculation from Section 6.4 to show that $\sum_{j=1}^m p_j^2 \left\|\left\|\hat{\Sigma}_j\right\|\right\|_{\text{op}} \xrightarrow{p} 0$.

(ii) Using the Cauchy Schwartz inequality on Schatten $p-$norms and using the high probability bounds from the bias proof, we get that for some constant C_9, the following holds with probability going to 1.

$$\begin{aligned}
\frac{2\lambda\sigma_i^2 p_i}{n_i} 2 \operatorname{tr}\left(\Sigma_i T_i^{-1} Q^{-1} T_i^{-1} \hat{\Sigma}_i T_i^{-1}\right) &\leq \frac{2 p_i \lambda \sigma_i^2}{n_i} \|T_i^{-1}\|_1 \left\|\left\|T_i^{-1}\right\|\right\|_{\text{op}} \|\Sigma_i\|_{\text{op}} \left\|\left\|Q^{-1}\right\|\right\|_{\text{op}} \left\|\left\|T_i^{-1}\right\|\right\|_{\text{op}} \left\|\left\|\hat{\Sigma}_i\right\|\right\|_{\text{op}} \\
&\leq \frac{M C_9 \sigma_i^2 d p_i}{\lambda n_i}.
\end{aligned}$$

$\left\|\left\|Q^{-1}\right\|\right\|_{\text{op}}$ is upper bounded by some constant as shown above. Moreover, since $p_i \to 0$, we have $\frac{2\lambda\sigma^2 p_i}{n_i} 2 \operatorname{tr}\left(T_i^{-1} Q^{-1} T_i^{-1} \hat{\Sigma}_i T_i^{-1}\right) \xrightarrow{p} 0$

(iii) Using Theorem 3 of [HMRT19], as $d \to \infty$, such that $\frac{d}{n_i} \to \gamma_i > 1$, we know that the limit of $\operatorname{tr}((\hat{\Sigma}_i + \lambda I)^{-2} \hat{\Sigma}_i \Sigma_i) \frac{\sigma_i^2}{n_i}$ is given by (18) with $\gamma = \gamma_i$ and $\widehat{H}_n, \widehat{G}_n$ be the empirical spectral distribution and weighted empirical spectral distribution of Σ_i respectively.

In the case when $\Sigma_i = I$, using Theorem 1 of [HMRT19] we have $V_i(\hat{\theta}_i^P(\lambda)|X) = \frac{\sigma_i^2}{n_i} \operatorname{tr}((\hat{\Sigma}_i + \lambda I)^{-2} \hat{\Sigma}_i^\dagger \Sigma_i) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}$.

6.8 Proof of Corollary 3.1

We first prove that $\rho_i \geq r_i$. If we let $\omega \in \Omega$ be the probability space associated with $\theta_i^\star$, we claim that $\langle \theta_0^\star, \theta_i^\star(\omega) - \theta_0^\star \rangle = 0$ for all (not just a.s.) $\omega \in \Omega$. For the sake of contradiction, suppose that there exists ω' such that $\langle \theta_0^\star, \theta_i^\star(\omega') - \theta_0^\star \rangle > 0$. If there exists $\omega'' \neq \omega'$ such that $\langle \theta_0^\star, \theta_i^\star(\omega'') - \theta_0^\star \rangle \neq \langle \theta_0^\star, \theta_i^\star(\omega') - \theta_0^\star \rangle$, then observe that the following equalities are true:

$$\begin{aligned}\|\theta_i^\star(\omega') - \theta_0^\star\|_2^2 + \|\theta_0^\star\|_2^2 + 2\langle \theta_0^\star, \theta_i^\star(\omega') - \theta_0^\star \rangle = \|\theta_i^\star(\omega')\|_2^2 = \rho_i^2 = \|\theta_i^\star(\omega'')\|_2^2 \\ = \|\theta_i^\star(\omega'') - \theta_0^\star\|_2^2 + \|\theta_0^\star\|_2^2 + 2\langle \theta_0^\star, \theta_i^\star(\omega'') - \theta_0^\star \rangle\end{aligned}$$

In looking at the first and last term, we see that this implies $\langle \theta_0^\star, \theta_i^\star(\omega') - \theta_0^\star \rangle = \langle \theta_0^\star, \theta_i^\star(\omega'') - \theta_0^\star \rangle$, which is a contradiction. Consider the other possibility that for all $\omega'' \in \Omega$, $\langle \theta_0^\star, \theta_i^\star(\omega'') - \theta_0^\star \rangle = \langle \theta_0^\star, \theta_i^\star(\omega') - \theta_0^\star \rangle > 0$. This would imply that $\mathbb{E}[\langle \theta_0^\star, \theta_i^\star(\omega) - \theta_0^\star \rangle] > 0$ which is a contradiction, since $\mathbb{E}[\theta_i^\star(\omega)] = \theta_0^\star$. Now since $\langle \theta_0^\star, \theta_i^\star(\omega) - \theta_0^\star \rangle = 0$, we have that $\rho_i^2 = r_i^2 + \|\theta_0^\star\|_2^2 \geq r_i^2$.

To show the result for $\hat{\theta}_0^{FA}$, recall eq. (19). The bias associated with this estimator is

$$B_i(\hat{\theta}_0^{FA}|X) = \left\| \theta_0^\star - \theta_i^\star + (\sum_{j=1}^m p_j \hat{\Sigma}_j)^{-1} \sum_{j=1}^m p_j \hat{\Sigma}_j (\theta_j^\star - \theta_0^\star) \right\|_2^2$$

Using the bias proof of Theorem 1, treating Π_i as the identity, we know that this quantity converges in probability to r_i^2. Showing the variance of $\hat{\theta}_0^{FA}(0)$ of goes to 0 follows directly from part (i) of the variance proof of Theorem 1, again treating Π_i as the identity.

The result regarding the estimator $\hat{\theta}_i^N$ is a direct consequence of Theorem 1 from [HMRT19]. The result regarding the estimator $\hat{\theta}_i^N(\lambda)$ is a direct consequence of Corollary 5 from [HMRT19].

6.9 Proof that RTFA has lower risk than FedAvg

We show that RTFA with optimal hyperparameter has lower risk than FedAvg by using the fact that $(1 - 1/\gamma)^2 \leq (1 + 1/\gamma)^2$ for $\gamma \geq 1$ and completing the square:

$$\begin{aligned}L_i(\hat{\theta}_i^R(\lambda^\star); \theta_i^\star|X) &= \frac{1}{2}\left[r_i^2\left(1 - \frac{1}{\gamma}\right) - \sigma_i^2 + \sqrt{r_i^4\left(1 - \frac{1}{\gamma}\right)^2 + \sigma_i^4 + 2\sigma_i^2 r_i^2\left(1 + \frac{1}{\gamma}\right)} \right] \\ &\leq r_i^2 = L_i(\hat{\theta}_0^{FA}; \theta_i^\star|X).\end{aligned}$$

Algorithm 3 Naive local training

Require: m: number of users, K: epochs

for $i \leftarrow 1$ to m **do**
  Each client runs K epochs of SGM with personal stepsize α
end for

Algorithm 4 Federated Averaging [MMR+17]

Require: R: Communication Rounds, D: Number of users sampled each round, K: Number of local update steps, $\hat{\theta}^{FA}_{0,0}$: Initial iterate for global model

for $r \leftarrow 0$ to $R-1$ **do**
  Server samples a subset of clients $\mathcal{S}_r$ uniformly at random such that $|\mathcal{S}_r| = D$
  Server sends $\hat{\theta}^{FA}_{0,r}$ to all clients in $\mathcal{S}_r$
  for $i \in \mathcal{S}_r$ **do**
    Set $\hat{\theta}^{FA}_{i,r+1,0} \leftarrow \hat{\theta}^{FA}_{0,r}$
    for $k \leftarrow 1$ to K **do**
      Sample a batch $\mathcal{D}^i_k$ of size B from user i's data $\mathcal{D}_i$
      Compute Stochastic Gradient $g(\hat{\theta}^{FA}_{i,r+1,k-1}; \mathcal{D}^i_k) = \frac{1}{B}\sum_{S\in\mathcal{D}^i_k} \nabla F(\hat{\theta}^{FA}_{i,r+1,k-1}; S)$
      Set $\hat{\theta}^{FA}_{i,r+1,k} \leftarrow \hat{\theta}^{FA}_{i,r+1,k-1} - \eta g(\hat{\theta}^{FA}_{i,r+1,k-1}; \mathcal{D}^i_k)$
    end for
    Client i sends $\hat{\theta}^{FA}_{i,r+1,K}$ back to the server.
  end for
  Server updates the central model using $\hat{\theta}^{FA}_{0,r+1} = \sum_{j=1}^{D} \frac{n_j}{\sum_{j=1}^{D} n_j} \hat{\theta}^{FA}_{i,r+1,K}$.
end for
return $\hat{\theta}^{FA}_{0,R}$

7 Algorithm implementations

In this section, we give all steps of the exact algorithms used to implement all algorithms in the experiments section.

Algorithm 5 FTFA

Require: P: Personalization iterations

Server sends $\hat{\theta}^{FA}_0 = \hat{\theta}^{FA}_{0,R}$ (using Algorithm 4 with stepsize η) to all clients
for $i \leftarrow 1$ to m **do**
  Run P steps of SGM on $\widehat{L}_i(\cdot)$ using $\hat{\theta}^{FA}_0$ as initial point with learning rate α and output $\hat{\theta}^{FA}_{i,P}$
end for
return $\hat{\theta}^{FA}_{i,P}$

Algorithm 6 RTFA

Require: P: Personalization iterations

Server sends $\hat{\theta}_0^{FA} = \hat{\theta}_{0,R}^{FA}$ (using Algorithm 4 with stepsize η) to all clients
for $i \leftarrow 1$ to m **do**
  Run P steps of SGM on $\widehat{L}_i(\theta) + \frac{\lambda}{2}\left\|\theta - \hat{\theta}_0^{FA}\right\|_2^2$ with learning rate α and output $\hat{\theta}_{i,P}^{FA}$
end for
return $\hat{\theta}_{i,P}^{FA}$

Algorithm 7 MAML-FL-HF [FMO20]

Require: R: Communication Rounds, D: Number of users sampled each round, K: Number of local update steps, $\hat{\theta}_{0,0}^{M}(\alpha)$: Initial iterate for global model

for $r \leftarrow 0$ to $R-1$ **do**
  Server samples a subset of clients $\mathcal{S}_r$ uniformly at random such that $|\mathcal{S}_r| = D$
  Server sends $\hat{\theta}_{0,r}^{M}(\alpha)$ to all clients in $\mathcal{S}_r$
  for $i \in \mathcal{S}_r$ **do**
    Set $\hat{\theta}_{i,r+1,0}^{M}(\alpha) \leftarrow \hat{\theta}_{0,r}^{M}(\alpha)$
    for $k \leftarrow 1$ to K **do**
      Sample a batch $\mathcal{D}_k^i$ of size B from user i's data $\mathcal{D}_i$
      Compute Stochastic Gradient $g(\hat{\theta}_{i,r+1,k-1}^{M}(\alpha); \mathcal{D}_k^i) = \frac{1}{B}\sum_{S\in\mathcal{D}_k^i} \nabla F(\hat{\theta}_{i,r+1,k-1}^{M}(\alpha); S)$
      Set $\hat{\theta}_{i,r+1,k}^{M}(\alpha) \leftarrow \hat{\theta}_{i,r+1,k-1}^{M}(\alpha) - \alpha g(\hat{\theta}_{i,r+1,k-1}^{M}(\alpha); \mathcal{D}_k^i)$
    end for
    Client i sends $\hat{\theta}_{i,r+1,K}^{M}(\alpha)$ back to the server.
  end for
  Server updates the central model using $\hat{\theta}_{0,r+1}^{M}(\alpha) = \sum_{j=1}^{D} \frac{n_j}{\sum_{j=1}^{D} n_j} \hat{\theta}_{i,r+1,K}^{M}(\alpha)$.
end for
Server sends $\hat{\theta}_{0,R}^{M}(\alpha)$ to all clients
for $i \leftarrow 1$ to m **do**
  Run P steps of SGM on $\widehat{L}_i(\cdot)$ using $\hat{\theta}_0^{M}(\alpha)$ as initial point with learning rate α and output $\hat{\theta}_{i,P}^{M}(\alpha)$
end for
return $\hat{\theta}_{0,R}^{M}(\alpha)$

8 Experimental Details

8.1 Dataset Details

In this section, we provide detailed descriptions on datasets and how they were divided into users. We perform experiments on federated versions of the Shakespeare [MMR+17], CIFAR-100 [KH09], EMNIST [CATvS17], and Stack Overflow [MRR+19] datasets. We download all datasets using FedML APIs [HLS+20] which in turn get their datasets from [MRR+19]. For each dataset, for each client, we divide their data into train, validation and test sets with roughly a $80\%, 10\%, 10\%$ split. The information regarding the number of users in each dataset, dimension of the model used, and the division of all samples into train, validation and test sets is given in Table 1.

Dataset	Users	Dimension	Train	Validation	Test	Total Samples
CIFAR 100	600	51200	48000	6000	6000	60000
Shakespeare	669	23040	33244	4494	5288	43026
EMNIST	3400	31744	595523	76062	77483	749068
Stackoverflow-nwp	300	960384	155702	19341	19736	194779

Table 1: Dataset Information

Algorithm 8 pFedMe [DTN20]

Require: R: Communication Rounds, D: Number of users sampled each round, K: Number of local update steps, $\hat{\theta}^P_{0,0}(\lambda)$: Initial iterate for global model

for $r \leftarrow 0$ to $R-1$ **do**
  Server samples a subset of clients $\mathcal{S}_r$ uniformly at random such that $|\mathcal{S}_r| = D$
  Server sends $\hat{\theta}^P_{0,r}(\lambda)$ to all clients in $\mathcal{S}_r$
  for $i \in \mathcal{S}_r$ **do**
    Set $\hat{\theta}^P_{i,r+1,0}(\lambda) \leftarrow \hat{\theta}^P_{0,r}(\lambda)$
    for $k \leftarrow 1$ to K **do**
      Sample a batch $\mathcal{D}^i_k$ of size B from user i's data $\mathcal{D}_i$
      Compute $\theta_i(\hat{\theta}^P_{i,r+1,k-1}(\lambda)) = \operatorname{argmin}_\theta \frac{1}{B}\sum_{S\in\mathcal{D}^i_k} \nabla F(\theta; S) + \frac{\lambda}{2}\left\|\theta - \hat{\theta}^P_{i,r+1,k-1}(\lambda)\right\|_2^2$
      Set $\hat{\theta}^P_{i,r+1,k}(\lambda) \leftarrow \hat{\theta}^P_{i,r+1,k-1}(\lambda) - \eta\lambda(\hat{\theta}^P_{i,r+1,k-1}(\lambda) - \theta_i(\hat{\theta}^P_{i,r+1,k-1}(\lambda)))$
    end for
    Client i sends $\hat{\theta}^P_{i,r+1,K}(\lambda)$ back to the server.
  end for
  Server updates the central model using $\hat{\theta}^P_{0,r+1}(\lambda) = (1-\beta)\hat{\theta}^P_{0,r}(\lambda) + \beta\sum_{j=1}^D \frac{n_j}{\sum_{j=1}^D n_j}\hat{\theta}^P_{i,r+1,K}(\lambda)$.
end for
return $\hat{\theta}^P_{0,R}(\lambda)$

Shakespeare Shakespeare is a language modeling dataset built using the works of William Shakespeare and the clients correspond to a speaking role with at least two lines. The task here is next character prediction. The way lines are split into sequences of length 80, and the description of the vocabulary size is same as [RCZ+21] (Appendix C.3). Additionally, we filtered out clients with less than 3 sequences of data, so as to have a train-validation-test split for all the clients. This brought the number of clients down to 669. More information on sample sizes can be found in Table 1. The models trained on this dataset are trained on two Tesla P100-PCIE-12GB GPUs.

CIFAR-100 CIFAR-100 is an image classification dataset with 100 classes and each image consisting of 3 channels of 32x32 pixels. We use the clients created in the Tensorflow Federated framework [MRR+19] — client division is described in Appendix F of [RCZ+21]. Instead of using 500 clients for training and 100 clients for testing as in [RCZ+21], we divided each clients' dataset into train, validation and test sets and use all the clients' corresponding data for training, validation and testing respectively. The models trained on this dataset are trained on two Titan Xp GPUs.

EMNIST EMNIST contains images of upper and lower characters of the English language along with images of digits, with total 62 classes. The federated version of EMNIST partitions images by their author providing the dataset natural heterogenity according to the writing style of each person. The task is to classify images into the 62 classes. As in other datasets, we divide each clients' data into train, validation and test sets randomly. The models trained on this dataset are trained on two Tesla P100-PCIE-12GB GPUs.

Stack Overflow Stack Overflow is a language model consisting of questions and answers from the StackOverflow website. The task we focus on is next word prediction. As described in Appendix C.4 of [RCZ+21], we also restrict to the 10000 most frequently used words, and perform padding/truncation to ensure each sentence to have 20 words. Additionally, due to scalability issues, we use only a sample of 300 clients from the original dataset from [MRR+19] and for each client, we divide their data into train, validation and test sets randomly. The models trained on this dataset are trained on two Titan Xp GPUs.

8.2 Hyperparameter Tuning Details

8.2.1 Pretrained Model

We now describe how we obtain our pretrained models. First, we train and hyperparameter tune a neural net classifier on the train and validation sets in a non-federated manner. The details of the hyperparameter sweep are as follows:

Shakespeare For this dataset we use the same neural network architecture as used for Shakespeare in [MMR+17]. It has an embedding layer, an LSTM layer and a fully connected layer. We use the StepLR learning rate scheduler of PyTorch , and we hyperparameter tune over the step size [0.0001, 0.001, 0.01, 0.1, 1] and the learning rate decay gamma [0.1, 0.3, 0.5] for 25 epochs with a batch size of 128.

CIFAR-100 For this dataset we use the Res-Net18 architecture [HZRS16]. We perform the standard preprocessing for CIFAR datasets for train, validation and test data. For training images, we perform a random crop to shape $(32, 32, 3)$ with padding size 4, followed by a horizontal random flip. For all training, validationn and testing images, we normalize each image according to their mean and standard deviation. We use the hyperparameters specified by [wei20] to train our nets for 200 epochs.

EMNIST For this dataset, the architecture we use is similar to that found in [RCZ+21]; the exact architecture can be found in our code. We use the StepLR learning rate scheduler of PyTorch, and we hyperparameter tune over the step size [0.0001, 0.001, 0.01, 0.1, 1] and the learning rate decay gamma [0.1, 0.3, 0.5] for 25 epochs with a batch size of 128.

Stackoverflow For this dataset we use the same neural network architecture as used for Stack Overflow next word prediction task in [RCZ+21]. We use the StepLR learning rate scheduler of PyTorch, and we hyperparameter tune over the step size [0.0001, 0.001, 0.01, 0.1, 1] and the learning rate decay gamma [0.1, 0.3, 0.5] for 25 epochs with a batch size of 128.

8.2.2 Federated Last Layer Training

After selecting the best hyperparameters for each net, we pass our data through said net and store their representations (i.e., output from penultimate layer). These representations are the data we operate on in our federated experiments.

Using these representations, we do multi-class logistic regression with each of the federated learning algorithms we test; we adapt and extend this code base [DTN20] to do our experiments. For all of our algorithms, the number of global iterations R is set to 400, and the number of local iterations K is set to 20. The number of users sampled at global iteration r, D, is set to 20. The batch size per local iteration, B, is 32. The random seed is set to 1. For algorithms FTFA, RTFA, MAML-FL-FO, and MAML-FL-HF, we set the number of personalization epochs P to be 10. We fix some hyperparameters due to computational resource restrictions and to avoid conflating variables; we choose to fix these ones out of precedence, see experimental details of [RCZ+21]. We now describe what parameters we hyperparameter tune over for each algorithm.

Naive Local Training This algorithm is described in Algorithm 3. We hyperparameter tune over the step size α [0.0001, 0.001, 0.01, 0.1, 1, 10].

FedAvg This algorithm is described in Algorithm 4. We hyperparameter tune over the step size η [0.0001, 0.001, 0.01, 0.1, 1, 10].

FTFA This algorithm is described in Algorithm 1. We hyperparameter tune over the step size of FedAvg η [0.0001, 0.001, 0.01, 0.1, 1], and the step size of the personalization SGM steps α [0.0001, 0.001, 0.01, 0.1, 1].

RTFA This algorithm is described in Algorithm 6. We hyperparameter tune over the step size of FedAvg η [0.0001, 0.001, 0.01, 0.1, 1], the step size of the personalization SGM steps α [0.0001, 0.001, 0.01, 0.1, 1], and the ridge parameter λ [0.001, 0.01, 0.1, 1, 10].

MAML-FL-HF This is the hessian free version of the algorithm, i.e., the hessian term is approximated via finite differences (details can be found in [FMO20]). This algorithm is described in Algorithm 7. We hyperparameter tune over the step size η [0.0001, 0.001, 0.01, 0.1, 1], the step size of the personalization SGM steps α [0.0001, 0.001, 0.01, 0.1, 1], and the hessian finite-difference-approximation parameter δ [0.001, 0.00001]. We used only two different values of δ because the results of preliminary experiments suggested little change in accuracy with changing δ.

MAML-FL-FO This is the first order version of the algorithm, i.e., the hessian term is set to 0 (details can be found in [FMO20]). This algorithm is described in Algorithm 7. We hyperparameter tune over the step size η [0.0001, 0.001, 0.01, 0.1, 1], the step size of the personalization SGM steps α [0.0001, 0.001, 0.01, 0.1, 1].

pFedMe This algorithm is described in Algorithm 8. We hyperparameter tune over the step size η [0.0005, 0.005, 0.05], and the weight β [1, 2]. The proximal optimization step size, hyperparameter K, and prox-regularizer λ associated with approximately solving the prox problem is set to 0.05, 5, and 15 respectively. We chose these hyperparameters based on the suggestions from [DTN20]. We were unable to hyperparameter tune pFedMe as much as, for example, RTFA because each run of pFedMe takes significantly longer to run. Additionally, for this same reason, we were unable to run pFedMe on the Stack Overflow dataset. While we do not have wall clock comparisons (due to multiple jobs running on the same gpu), we have observed that pFedMe, with the hyperparameters we specified, takes approximately 20x the compute time to complete relative to FTFA, RTFA, and MAML-FL-FO.

The ideal hyperparameters for each dataset can be found in the tables below:

Algorithm	η	α	λ	δ	β
Naive Local	-	0.1	-	-	-
FedAvg	0.1	-	-	-	-
FTFA	1	0.1	-	-	-
RTFA	1	0.1	0.1	-	-
MAML-FL-HF	1	0.1	-	0.00001	-
MAML-FL-FO	1	0.1	-	-	-
pFedMe	0.05	-	-	-	2

Table 2: Shakespeare Best Hyperparameters

Algorithm	η	α	λ	δ	β
Naive Local	-	0.1	-	-	-
FedAvg	0.01	-	-	-	-
FTFA	0.001	0.1	-	-	-
RTFA	0.001	0.1	0.1	-	-
MAML-FL-HF	0.001	0.01	-	0.001	-
MAML-FL-FO	0.001	0.01	-	-	-
pFedMe	0.05	-	-	-	1

Table 3: CIFAR-100 Best Hyperparameters

Algorithm	η	α	λ	δ	β
Naive Local	-	0.001	-	-	-
FedAvg	0.01	-	-	-	-
FTFA	0.1	0.01	-	-	-
RTFA	0.1	0.01	0.1	-	-
MAML-FL-HF	0.1	0.01	-	0.00001	-
MAML-FL-FO	0.1	0.01	-	-	-
pFedMe	0.05	-	-	-	2

Table 4: EMNIST Best Hyperparameters

Algorithm	η	α	λ	δ	β
Naive Local	-	0.1	-	-	-
FedAvg	1	-	-	-	-
FTFA	1	0.1	-	-	-
RTFA	1	0.1	0.001	-	-
MAML-FL-HF	1	0.1	-	0.00001	-
MAML-FL-FO	1	0.1	-	-	-

Table 5: Stack Overflow Best Hyperparameters

8.3 Additional Results

In this section, we add additional plots from the experiments we conducted, which were omitted from the main paper due to length constraints. In essence, these plots only strengthen the claims made in the experiments section in the main body of the paper.

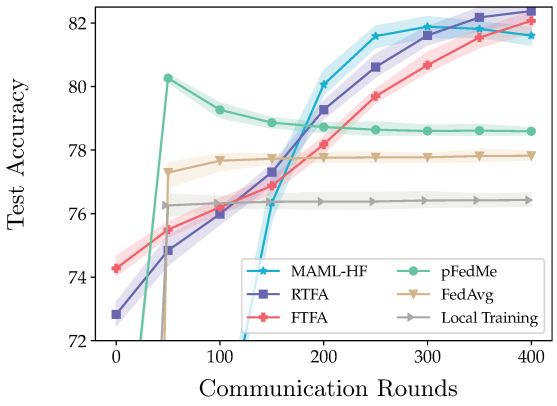

Figure 7. CIFAR-100. Best-average-worst intervals created from different random seeds.

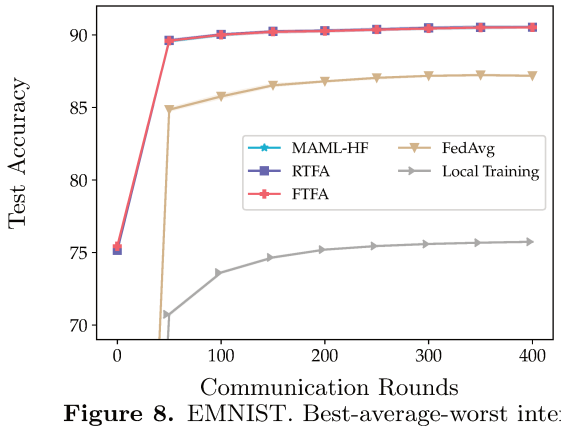

Figure 8. EMNIST. Best-average-worst intervals created from different train-val splits.

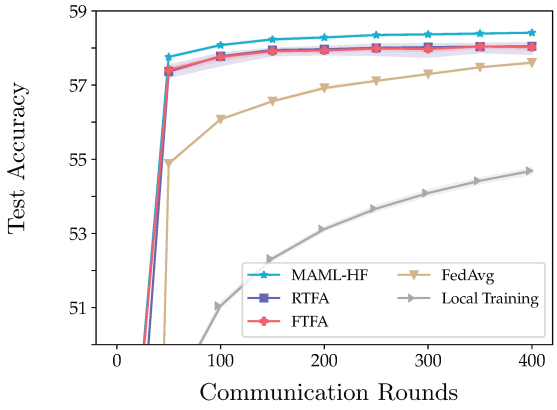

Figure 9: Shakespeare. Best-average-worst intervals created from different random train-val splits.

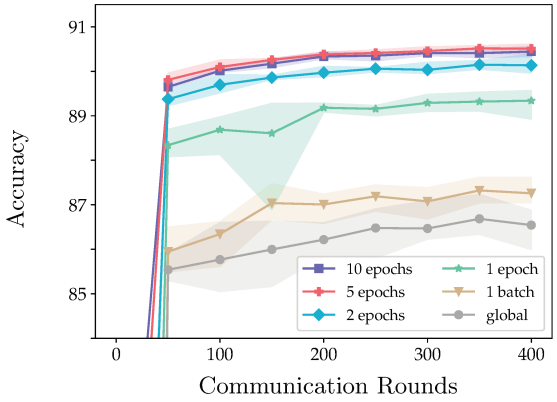

Figure 10. EMNIST. Gains of personalization for FO-MAML-FL

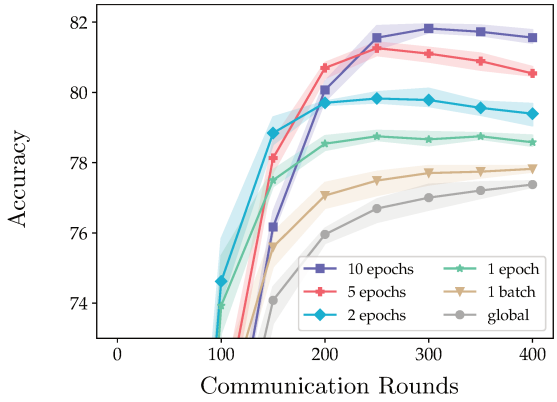

Figure 11. CIFAR. Gains of personalization for FO-MAML-FL

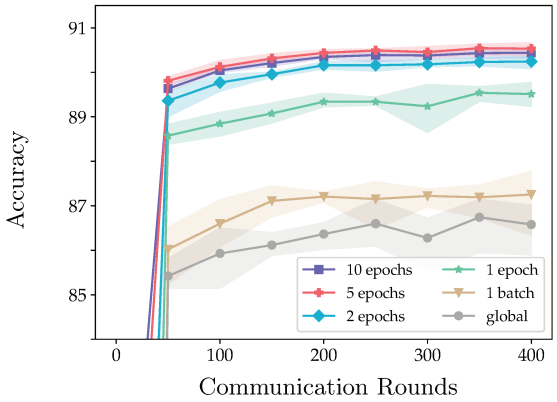

Figure 12. EMNIST. Gains of personalization for HF-MAML-FL

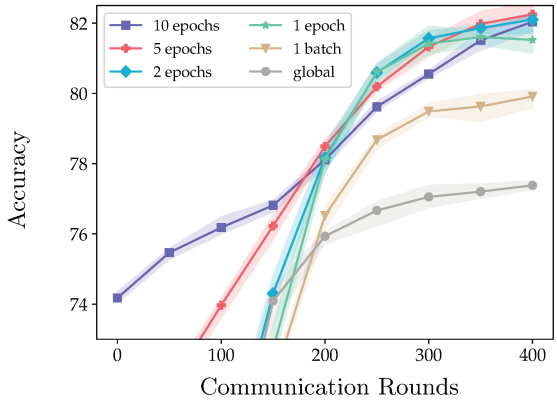

Figure 13. CIFAR. Gains of personalization for FTFA