Federated Asymptotics: a model to compare federated learning algorithms

Gary Cheng[⋆]
Stanford University

Karan Chadha[⋆]
Stanford University

John Duchi
Stanford University

Abstract

We develop an asymptotic framework to compare the test performance of (personalized) federated learning algorithms whose purpose is to move beyond algorithmic convergence arguments. To that end, we study a high-dimensional linear regression model to elucidate the statistical properties (per client test error) of loss minimizers. Our techniques and model allow precise predictions about the benefits of personalization and information sharing in federated scenarios, including that Federated Averaging with simple client fine-tuning achieves identical asymptotic risk to more intricate meta-learning approaches and outperforms naive Federated Averaging. We evaluate and corroborate these theoretical predictions on federated versions of the EMNIST, CIFAR-100, Shakespeare, and Stack Overflow datasets.

1 Introduction

In federated learning (FL), a collection of client devices collect data and coordinate with a central server to fit machine-learned models, where communication and availability constraints add challenges Kairouz et al. (2019). A natural formulation is to assume that among m clients, each client i has distribution P_i, draws observations $Z \sim P_i$, and wishes to fit a model—a parameter $\theta \in \Theta$—to minimize the expected loss $L_i(\theta) := \mathbb{E}_{P_i}[\ell(\theta; Z)]$, where $\ell(\theta; z)$ measures the performance of θ on example z. Federated learning therefore seeks to solve the multi-criterion problem

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

Problem (1) is challenging as no device has sufficient data to individually minimize L_i. Consequently, FL methods typically depart from the multicriterion objective (1) to provide more tractable problems. Many approaches seek a single θ that does well across all data, minimizing a (weighted) average loss

$$\sum_{i=1}^{m} p_i L_i(\theta) \quad \text{over } \theta \in \Theta, \tag{2}$$

where $p \in \mathbb{R}_+^m$ satisfies $p^T \mathbf{1} = 1$. This "zero personalization" approach has the advantage that data is plentiful and has led to a literature that develops communication-efficient optimization methods (Haddadpour and Mahdavi, 2019; Reddi et al., 2021; McMahan et al., 2017; Karimireddy et al., 2020; Mohri et al., 2019; Li et al., 2020). To address that distributions P_i across individual devices typically differ, however, it is important to more closely target problem (1). One natural assumption is that optimal client parameters are "close" and so must be near the minimizer of the zero-personalization objective (2). Personalized FL approaches leverage this assumption (Dinh et al., 2020; Smith et al., 2017; Wang et al., 2019; Fallah et al., 2020), often by designing new loss functions to regularize client parameters towards a global parameter.

While a growing literature (e.g. Fallah et al., 2020; Dinh et al., 2020; McMahan et al., 2017) appeals to optimization-based convergence rates as theoretical justification for their methods, such approaches neglect the impact of the choice of objective: minimizers of different loss functions perform differently. In standard statistical learning settings, the choice of loss and its attendant statistical properties are central, with researchers characterizing consistency properties of surrogate losses (Steinwart (2007); Zhang (2004a,b); Bartlett et al. (2006); Nowak-Vila et al. (2020)) or asymptotic risk in high-dimensional linear models (e.g. Hastie et al., 2019). In federated scenarios, in contrast, the surrogate losses (e.g. (2)) recent papers use to tackle the actual problem (1) lack such characterizations. By developing theoretical tools for this analysis, this paper addresses this gap through three main contributions:

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.

[⋆] Equal contribution, author order random.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Precise risk characterization (Sec. 3.1 & Sec. 4): We use our stylized model to evaluate the asymptotic test loss of several (existing) procedures. These include fine-tuned variants of Federated Averaging McMahan et al. (2017), where one learns an average global model (2) and updates once using local data; meta-learning FL variants; and proximal-regularized personalization.

Precise predictions and experiments (Sec. 5): Our theory makes several predictions, including that fairly naive methods (fine-tuning variants) should perform as well as more sophisticated methods, and giving conditions under which they improve upon zero-personalization (2) or zero-collaboration methods. To test these predicted behaviors, we experiment on federated versions of the EMNIST, CIFAR-100, Shakespeare, and Stack Overflow datasets; the results are quite consistent with the behavior the theory predicts.

Our choice to study linear models in the high-dimensional asymptotic setting 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. 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. For instance, Hastie et al. (2019) show that "double-descent" phenomena, where (neural-network) models show decreasing test loss as model size grows, exists even in linear regression. In adversarial learning, Carmon et al. (2019) use a simple 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 state-of-the-art. Feldman (2020) 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 high-dimensional asymptotics to develop statistical insights underpinning FL. This allows direct comparison between different FL methods—not between upper bounds, but actual losses—serving as a basic framework to motivate new methodologies in FL.

Related Work Fine tuning (Howard and Ruder (2018)) is a popular method of adapting to distribution shifts, which in FL corresponds to fine-tuning a global model on a user's local data (Wang et al. (2019); Yu et al. (2020); Li et al. (2021)). While fine-tuning's simplicity and practical efficacy recommend it, we know of little theoretical analysis.

A major direction in FL is toward personalization-incentivizing objectives. Smith et al. (2017) use transfer learning ideas (Caruana (1997); Pan and Yang (2009)) to formulate a multi-task learning objective, treating each machine as an individual task; this and other papers (Fallah et al. (2020); Mansour et al. (2020); Dinh et al. (2020)) 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 the conditions 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 are unclear. Related meta-learning approaches (Finn et al. (2017); Fallah et al. (2020); Jiang et al. (2019)) seek a global model that can "adapt" to local distributions P_i, typically by using a few gradient steps. This generally yields a complicated non-convex objective, making even heuristic guarantees hard to give and leading authors to emphasize convergence rates to stationary points.

Alternative personalization methods exist. For example, the papers Mansour et al. (2020); Zec et al. (2020) combine global and local models to incorporate personalized features. Chen et al. (2021), as we do, evaluate 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 Linear Model of (Personalized) Federated Learning

We propose a high-dimensional asymptotic model of Federated Learning, where clients solve related linear regression problems, and each client $i \in [m]$ has a local dataset of size n_i smaller than (but comparable to) the dimension d of the problem. This choice models that, empirically, 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). Each client's data follows an (overparameterized) linear regression model, where client i has n_i i.i.d. observations $(\mathbf{x}_{i,k}, y_{i,k}) \in \mathbb{R}^d \times \mathbb{R}$, $k = 1, \ldots, n_i$, following

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

We make the routine assumptions that the features are centered, with $\mathbb{E}[\mathbf{x}_{i,k}] = 0$ and $\mathrm{Cov}(\mathbf{x}_{i,k}) = \Sigma_i$, and that the is mean-zero with $\mathbb{E}[\xi_{i,k}] = 0$ and $\mathrm{Var}(\xi_{i,k}) = \sigma_i^2$. 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, defining the full data matrix $X := [X_1^T \cdots X_m^T]^T$. We also let $N := \sum_{j=1}^m n_j$.

To facilitate analysis, we take a Bayesian perspective. A prior P_θ^i on the parameter $\theta_i^\star$ relates tasks on each client, where we assume a global parameter $\theta_0^\star$ such that each $\theta_i^\star$ has support $r_i \mathbb{S}^{d-1} + \theta_0^\star$, the sphere of radius r_i centered at $\theta_0^\star$, with $\mathbb{E}[\theta_i^\star] = \theta_0^\star$. Differences in r_i (label shift) and Σ_i (covariate shift) capture variation between clients, while the shared center $\theta_0^\star$ captures similarity. 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.

Client i seeks to minimize its local loss—the squared prediction error on an independent new sample $\mathbf{x}_{i,0}$—conditioned on X. For sample loss $\ell(\theta; (\mathbf{x}, y)) = (\mathbf{x}^T\theta - y)^2 - \sigma_i^2$, client i's test loss is then

$$L_i(\hat{\theta}_i \mid X) := \mathbb{E}[(\mathbf{x}_{i,0}^T\hat{\theta}_i - \mathbf{x}_{i,0}^T\theta_i^\star)^2 \mid X] \\ = \mathbb{E}\left[\left\|\hat{\theta}_i - \theta_i^\star\right\|_{\Sigma_i}^2 \mid X\right],$$

where the expectation is 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 goal is to improve performance on each client (1). For analysis purposes, we often consider the equivalent bias-variance decomposition

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

Our main asymptotic assumption, which captures the high-dimensional (d large) and many-device (m large) 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. Hastie et al., 2019, 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 $\mu(s; \Sigma) := \frac{1}{d}\sum_{j=1}^d \mathbf{1}\{s \ge s_i\}$, where $s_1 \ge s_2 \ge \cdots \ge s_d$ are the eigenvalues of Σ.*

Assumption A2. *For each user i, data $\mathbf{x} = \Sigma_i^{\frac{1}{2}}\mathbf{z}$, where 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) $s_1 = \|\Sigma_i\|_{\mathrm{op}} \le M$, $s_d = \lambda_{\min}(\Sigma_i) \ge 1/M$, and $\int s^{-1} d\mu(s; \Sigma_i) < M$.

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

Assumption A.2 is fairly general. Assumption A.2.a is a higher moment bound on the entries. This is strictly weaker than standard assumptions like subgaussianity or boundedness. Without A.2.a (or something instead of it), we shouldn't expect any concentration at all. Assumption A.2.b prevents any degenerate settings where covariance matrices are low rank. Low rank covariance matrices imply the covariates lie within a low rank subspace, which is uncommon in real data sets. For example, MNIST training data has full rank covariance (c.f., Paquette et al., 2022). Finally, A.2.c ensures that the spectrum of the covariance matrices converge weakly (to some arbitrary probability distribution on $\mathbb{R}$, which we term ν_i); such a condition is standard for asymptotic statistics (c.f., Hastie et al., 2019). Without such a condition, the covariance structure of our data can vary wildly, and in turn, we wouldn't expect any concentration behavior.

3 Convergence guarantees for fine tuning with federated averaging

In this section, we use the high-dimensional asymptotic model above to 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 Federated Averaging (FTFA)

Fine-tuned Federated Averaging (FTFA) (Wang et al. (2019)) approximates minimizing the multi-criterion loss (1) using the two-step procedure in Algorithm 1 (see Appendix C for fuller pseudocode). Let $\mathcal{S}_i$ denote client i's sample. The idea is to replace the local risks L_i in (2) with their empirical counterparts

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

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 personalization algorithms. Each client i can run this local training phase independently of 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 $\sum_{j=1}^m \frac{p_j}{2n_j}\|X_j\theta - \mathbf{y}_j\|_2^2$. As the local regression problem to minimize $\|X_i\theta - \mathbf{y}_i\|_2^2$ is overparameterized, first-order methods on it solve minimum ℓ_2-norm interpolation problems (Gunasekar et al., 2018, Thm. 1). Thus, when converged, FTFA is equivalent to the two step

Algorithm 1 Fine-tuned and Ridge-Tuned Federated Averaging (FTFA & RTFA)

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

$$\hat{\theta}_0^{FA} = \operatorname*{argmin}_\theta \sum_{j=1}^m p_j \widehat{L}_j(\theta), \tag{4}$$

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

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

(b) **RTFA:** Client i minimizes a regularized empirical risk to return model

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

procedure

$$\hat{\theta}_0^{FA} = \operatorname*{argmin}_\theta \sum_{j=1}^m p_j \frac{1}{2n_j} \|X_j\theta - \mathbf{y}_j\|_2^2 \tag{5}$$

$$\hat{\theta}_i^{FA} = \operatorname*{argmin}_\theta \left\{ \left\|\hat{\theta}_0^{FA} - \theta\right\|_2 \text{ s.t. } X_i\theta = \mathbf{y}_i \right\}. \tag{6}$$

We presently present a convergence guarantee for the asymptotic bias and variance of FTFA, but to give the result, we require a technical assumption on the number of clients and problem dimension.

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 jth client when finding the global model (5), and we think of it as requiring that the number of machines grows with the average client sample size. As one case of interest, when each client has equal weight so that $p_j = 1/m$, A3 is equivalent to $\frac{N}{m}\frac{\log^{cq} d}{m^{q/2}} \to 0$, that is, that $m^{q/2}$ grows faster than the average client sample size N/m.

Theorem 1. *Consider the estimator $\hat{\theta}_i^{FA}$ in (6). Let Assumption A1 hold, and let Assumptions A2 and A3 hold with $c = 2$ and $q > 2$. Additionally, assume that for each m and $j \in [m]$, $\|\mathbb{E}[\hat{\Sigma}_j^2]\|_{\text{op}} \le \tau_2 < \infty$. Then for client i, the asymptotic bias and variance (3) 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 $\|z\|_{\Sigma_i}^2 := z^T \Sigma_i z$. For the special case when $\Sigma_i = I$, the limits are

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

The first part of the theorem provides expressions for the asymptotic bias and variance. When Σ_i are non-identity, Theorem 1's limits are unwieldy; see Appendix B.4 for exact expressions. One can use these expressions to numerically evaluate the asymptotic risk for arbitrary covariance matrices (Hastie et al., 2019, Section 4.1) relevant for a problem of interest. For the isotropic setting, these limits have closed-form expressions. As problems become more over parameterized, the bias dominates the asymptotic risk, while the asymptotic variance tends to zero.

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 (5) and second step

$$\hat{\theta}_i^R(\lambda) = \operatorname*{argmin}_\theta \frac{1}{2n_i}\|X_i\theta - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\left\|\hat{\theta}_0^{FA} - \theta\right\|_2^2, \tag{7}$$

where RTFA (Li et al. (2021)) 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*

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

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

For the special case when $\operatorname{Cov}(\mathbf{x}_{i,k}) = \Sigma_i = I$, 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 $m_i(z) = (1 - \gamma_i - z - \sqrt{(1-\gamma_i - z)^2 - 4\gamma_i z})/(2\gamma_i z)$. For each client $i \in [m]$, when λ takes the minimizing value $\lambda_i^\star = \sigma_i^2\gamma_i / r_i^2$, $L_i(\hat{\theta}_i^R(\lambda_i^\star)|X) \to \sigma_i^2 \gamma_i m_i(-\lambda_i^\star)$.

As with Theorem 1, see Appendix B.5 for exact limit expressions when Σ_i are general.

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 (6).

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

$$\hat{\theta}_i^N = \operatorname*{argmin}_\theta \|\theta\|_2 \quad s.t. \quad X_i\theta = \mathbf{y}_i, \tag{8}$$

and the ridge-penalized, zero-collaboration estimator

$$\hat{\theta}_i^N(\lambda) = \operatorname*{argmin}_\theta \frac{1}{2n_i}\|X_i\theta - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2. \tag{9}$$

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. (8). 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. (9). Then under the preceding conditions,

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

Taking $\lambda = \lambda_i^\star = \sigma_i^2\gamma_i/\rho_i^2$, 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

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

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, as $\rho_i \ge r_i$, it is 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 Applying our model to Meta learning and Proximal Regularization

The fine-tuning procedures in the previous section provide a (perhaps) naive baseline, so we consider two 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 (Fallah et al. (2020)) and the second using a proximal-regularized approach (Dinh et al. (2020)). 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) (Finn et al. (2017)) learns models that adapt to related tasks by minimizing an empirical loss evaluated, not at a given parameter θ, but at a "one-step-updated" parameter $\theta - \alpha\nabla L(\theta)$, representing one-shot learning. Fallah et al. (2020), 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 Appendix C for detailed pseudocode). Algorithm 2 has two variants (Fallah et al. (2020)); one ignores the Hessian term, and the other approximates the Hessian using finite differences. Fallah et al. (2020) show that these these algorithms converge to a stationary point of eq. (10) (with $p_j = 1/m$) for general non-convex smooth functions.

In our linear model, for appropriately chosen hyperparameters and as the optimization method converges, this per-

Algorithm 2 Model-Agnostic Meta-Learning for Federated Learning (MAML-FL)

1. Server and clients coordinate to (approximately) solve

$$\hat{\theta}_0^M(\alpha) = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^m p_j \widehat{L}_j(\theta - \alpha \nabla \widehat{L}_j(\theta)), \quad (10)$$

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

2. Client i fits model $\hat{\theta}_i^M(\alpha)$ minimizing $\widehat{L}_i(\cdot)$ using first-order method initialized at $\hat{\theta}_0^M(\alpha)$.

sonalization method corresponds to the following two step procedure:

$$\hat{\theta}_0^M(\alpha) = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^m \frac{p_j}{2n_j} \left\| X_j \left[\theta - \frac{\alpha}{n_j} X_j^T (X_j \theta - \mathbf{y}_j) \right] - \mathbf{y}_j \right\|_2^2 \quad (11)$$

$$\hat{\theta}_i^M(\alpha) = \underset{\theta}{\operatorname{argmin}} \left\| \hat{\theta}_0^M(\alpha) - \theta \right\|_2 \quad s.t. \quad X_i \theta = \mathbf{y}_i \quad (12)$$

As in Section 3.1, any fully converged model in step 2 of Alg. 2 must be a minimum norm interpolant (12). The representations (11) and (12) 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 (12). 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 $\|\mathbb{E}[\hat{\Sigma}_j^6]\|_{\text{op}} \leq \tau_6 < \infty$. Then for client i, the asymptotic errors of MAML-FL satisfy*

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

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

where $\Pi_i := I - \hat{\Sigma}_i^\dagger \hat{\Sigma}_i$ and $\|z\|_{\Sigma_i}^2 := z^T \Sigma_i z$. For the special case that $\Sigma_i = I$, the limits are

$$B_i(\hat{\theta}_i^M(\alpha)|X) \xrightarrow{p} r_i^2 \left(1 - \frac{1}{\gamma_i}\right)$$

$$V_i(\hat{\theta}_i^M(\alpha)|X) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}.$$

(As usual, see Appendix B.6 for expressions for general Σ.)

In short, the asymptotic test risk of MAML-FL matches that of FTFA (Theorem 1). In general, the MAML-FL objective (10) 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 (Nesterov, 2004) upon which convergence analyses (and algorithms) repose fail to hold. Additionally, computing gradients of the MAML-FL objective (10) 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 Fallah et al. (2020) propose performs a single stochastic gradient step for personalization, which is distinct from the analyzed procedure (12), which is equivalent to running SGM until convergence from the initialization $\hat{\theta}_0^M(\alpha)$ (see step 2 of Algorithm 2). We find two main justifications for this choice: first, experimental work (Jiang et al. (2019)), in addition to our own experiments (see Figures 1e and 1f), suggests that the more (stochastic gradient) steps of personalization, the better performance. Further, performing personalization SGM steps locally, in parallel, and asynchronously is no more expensive than running the first step of Algorithm 2. This assumption of convergence also presents a fair point of comparison between the algorithms we consider.

4.2 Proximal-Regularized Approach

Instead of sequential fine-tuning, an alternative approach to personalization is to jointly optimize global and local parameters. In this vein, Dinh et al. (2020) propose the pFedMe algorithm to solve the a regularized and coupled optimization problem that in our linear model simplifies to

$$\left(\hat{\theta}_0^P(\lambda), \hat{\theta}_1^P(\lambda), \ldots, \hat{\theta}_m^P(\lambda)\right) = \underset{\theta_0, \theta_1, \ldots, \theta_m}{\operatorname{argmin}} \sum_{j=1}^m p_j \left(\frac{1}{2n_j} \|X_j \theta_j - \mathbf{y}_j\|_2^2 + \frac{\lambda}{2} \|\theta_j - \theta_0\|_2^2 \right), \quad (13)$$

where $\hat{\theta}_0^P(\lambda)$ is the global model and $\hat{\theta}_i^P(\lambda)$ denote the local models. Here, the proximal penalty $\|\theta_j - \theta_0\|_2$ encourages the local models θ_i to be close to one another.

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 (13). Let Assumption A1 hold, and let Assumptions A3 and A2 hold with $c = 2$ and the*

same $q > 2$. Additionally, assume that $\mathbb{E}[\|\hat{\Sigma}_j^2\|_{\rm op}] \le \tau_3 < \infty$. Further suppose that there exists $R \ge M$ such that $\limsup_{m\to\infty} \sup_{j\in[m]} \mathbb{P}(\lambda_{\max}(\hat{\Sigma}_j) > R) \le \frac{1}{16M^2\tau_3}$. Then for client i, the asymptotic prediction bias and variance of pFedMe are

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

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

For the special case that $\Sigma_i = I$, the limits are

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

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 $L_i(\hat{\theta}_i^P(\lambda_i^\star)|X) \to \sigma_i^2\gamma_i m_i(-\lambda_i^\star)$.

(See Appendix B.7 for exact expressions of these limits for general Σ.)

The asymptotic test loss of the proximal-regularized approach is thus identical to the locally ridge-regularized (RTFA) solution; see Theorem 2. Dinh et al. (2020)'s algorithm to optimize eq. (13) 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.

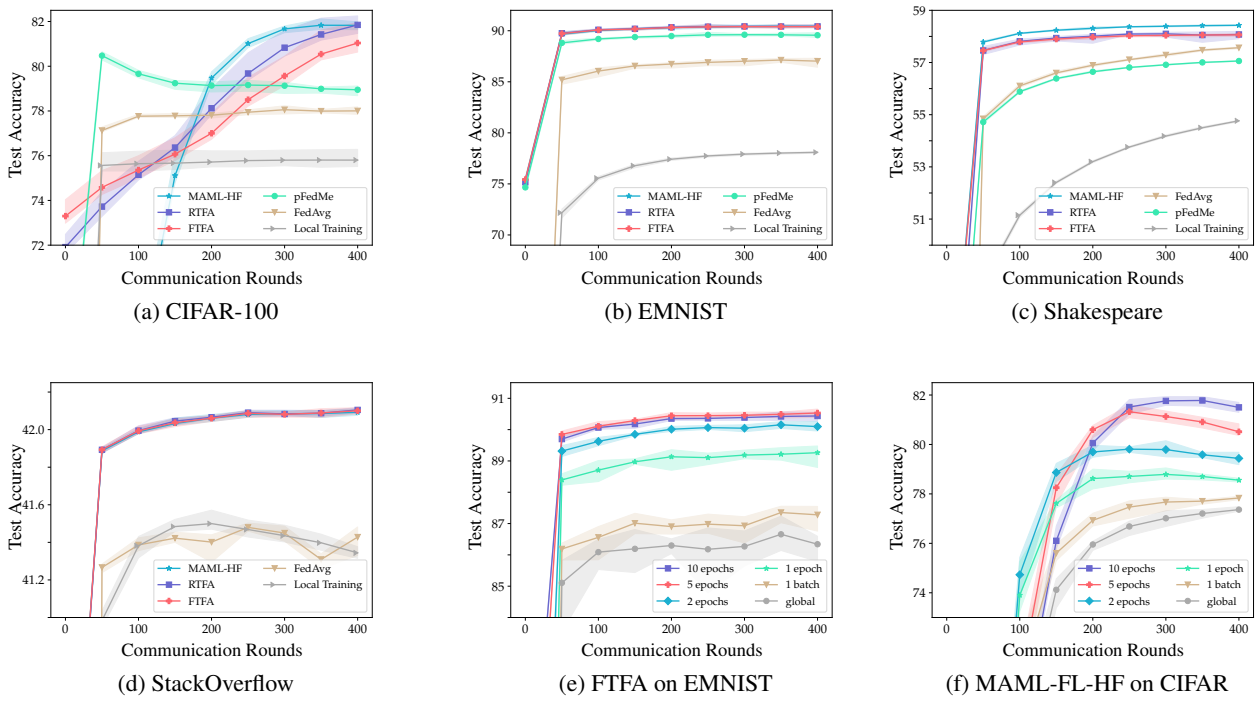
5 Experiments

We have presented a theoretical model which makes predictions about the asymptotic test risk of FTFA, RTFA, MAML-FL, and pFedMe under stylized assumptions on the data generating distribution. Two natural questions arise: 1. How well do the asymptotic results predict finite sample behavior? and 2. Do the predictions from our theoretical model accurately predict performance on benchmark federated learning datasets, where the assumptions on the data generating distribution may not hold? We choose to use this section to focus on the second question because we believe it is relatively more important. We answer the first question in the affirmative in Appendix A. Code for our experiments can be found at `https://github.com/garyxcheng/personalized-federated-learning`.

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 (McMahan et al. (2017)), CIFAR-100 (Krizhevsky and Hinton (2009)), EMNIST (Cohen et al. (2017)), and Stack Overflow (McMahan et al. (2019)) datasets; dataset statistics and details of how we divide the data to make effective "users" are in Appendix D. For each dataset, we compare the performance of the following algorithms: Zero Communication (Local Training), Zero Personalization (FedAvg), FTFA, RTFA, MAML-FL, and pFedMe (Dinh et al. (2020)). 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 D 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



(a) CIFAR-100 (b) EMNIST (c) Shakespeare

(d) StackOverflow (e) FTFA on EMNIST (f) MAML-FL-HF on CIFAR

Figure 1: (a) & (d) plot the best-average-worst intervals created from different train-val splits. (b) & (c) plot the best-average-worst intervals created from different random seeds. (e) & (f) show the effect of the number of personalization steps on test accuracy for FTFA and MAML-FL-HF respectively.

performance over a single centralized, pre-trained model (see, e.g., Reddi et al. (2021, 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 (Strubell et al. (2019)). 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 experiments, doing the experimental training and validation on an independent dataset.

Results Figures 1a to 1d 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 1a to 1c is worse than that of FTFA, RTFA and MAML-FL.

In Figures 1e and 1f, 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.

Acknowledgements

This work was supported by Office of Naval Research N00014-22-1-2669, NSF Robust Intelligence 2006777 and DAWN Consortium. The authors would also like to thank Daniel Levy and Hilal Asi for their helpful feedback on the presentation of the paper. Gary Cheng acknowledges support from the Professor Michael J. Flynn Stanford Graduate Fellowship.

References

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. (2006). Con-

vexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. (2019). Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems 32*.

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

Chen, R., Gittens, A., and Tropp, J. A. (2012). The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, to appear.

Chen, S., Zheng, Q., Long, Q., and Su, W. J. (2021). A theorem of the alternative for personalized federated learning. *arXiv:2103.01901 [stat.ML]*.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

de la Peña, V. H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Springer.

Dinh, C., Tran, N., and Nguyen, T. D. (2020). Personalized federated learning with moreau envelopes. *arXiv:2006.08848 [cs.LG]*.

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems 33*.

Feldman, V. (2020). Does learning require memorization? A short tale about a long tail. In *Proceedings of the Fifty-Second Annual ACM Symposium on the Theory of Computing*, pages 954–959.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*.

Haddadpour, F. and Mahdavi, M. (2019). On the convergence of local descent methods in federated learning. *arXiv:1910.14425 [cs.LG]*.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. (2019). Surprises in high-dimensional ridgeless linear least squares interpolation. *arXiv:1903.08560 [math.ST]*.

He, C., Li, S., So, J., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Shen, L., Zhao, P., Kang, Y., Liu, Y., Raskar, R., Yang, Q., Annavaram, M., and Avestimehr, S. (2020). FedML: A research library and benchmark for federated machine learning. *arXiv:2007.13518 [cs.LG]*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv:1801.06146 [cs.LG]*.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. (2019). Improving federated learning personalization via model agnostic meta learning. *arXiv:1909.12488 [cs.LG]*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv:1912.04977 [cs.LG]*.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Li, T., Hu, S., Beirami, A., and Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In *ICML*.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450.

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

McMahan, B., Rush, K., Reneer, M., Garrett, Z., and TensorFlow Federated Team (2019). Tensorflow federated stack overflow dataset. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers.

Nowak-Vila, A., Bach, F., and Rudi, A. (2020). Consistent structured prediction with max-min margin markov networks. In *Proceedings of the 37th International Conference on Machine Learning*.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. (2022). Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. (2021). Adaptive federated optimization. In *Proceedings of the Ninth International Conference on Learning Representations*.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. (2017). Federated multi-task learning. In *Advances in Neural Information Processing Systems 30*.

Steinwart, I. (2007). How to compare different loss functions. *Constructive Approximation*, 26:225–287.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory*. Academic Press.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. (2019). Federated evaluation of on-device personalization. *arXiv:1910.10252 [cs.LG]*.

weiaicunzai (2020). Pytorch-cifar100. https://github.com/weiaicunzai/pytorch-cifar100.

Yu, T., Bagdasaryan, E., and Shmatikov, V. (2020). Salvaging federated learning by local adaptation. *arXiv:2002.04758 [cs.LG]*.

Zec, E. L., Mogren, O., Martinsson, J., Sütfeld, L. R., and Gillblad, D. (2020). Specialized federated learning using a mixture of experts. *arXiv:2010.02056 [cs.LG]*.

Zhang, T. (2004a). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251.

Zhang, T. (2004b). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85.

A Synthetic Experiment

Our theoretical results are asymptotic in nature. A natural question is how many samples (i.e., how large does n need to be) before the finite sample behavior of the client risks start behaving like the asymptotic predictions we make. To answer this question, we run FTFA on synthetic data and vary n—the number of samples given each client—and plot the corresponding risk of the solution returned by FTFA. In particular, for each of the $i \in [m]$ clients where $m = 100$, for varying γ_i, and for varying n, we generate n sample inputs drawn from $\mathsf{N}(0, I_d)$, where $d = \gamma_i n$ to form a matrix X_i. We generate an optimal client parameter β_i by drawing it from $\mathsf{N}(0, I_d)$. We generate labels $y_i = X_i\beta_i + \xi$ where $\xi \sim \mathsf{N}(0, \sigma^2 I_n)$. We plot the risk of client i against γ_i in Figure 2. We see our theory is predictive of finite sample behavior even when n is small.

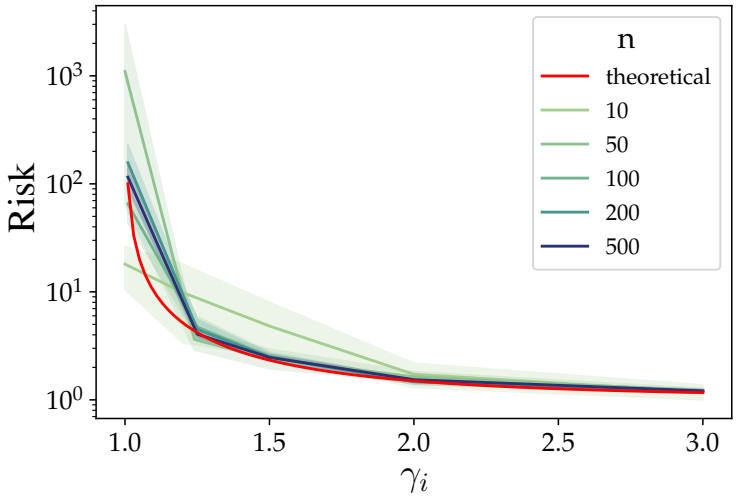


Figure 2: Test risk of client i against $\gamma_i = \frac{d}{n_i}$, the ratio of dimensions to number of samples for client i.

B Proofs

B.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.

B.2 Useful Lemmas

Lemma B.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 de la Peña and Giné (1999), 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 B.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 Chen et al. (2012), 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 B.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 Chen et al. (2012) 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 B.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,

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

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 B.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 $\left\|V^{-1} - U^{-1}\right\|_{\text{op}} \xrightarrow{p} 0$.*

Proof For any $t > 0$, we have by Theorem 2.5 (from Section III) of Stewart and Sun (1990)

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

We know this quantity goes to 0 by assumption. □

B.3 Some useful definitions from previous work

In this section, we recall some definitions from Hastie et al. (2019) 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 Hastie et al. (2019). 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 B.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

$$\mathscr{B}(\widehat{H}_n, \widehat{G}_n, \gamma) \coloneqq \|\theta^\star\|_2^2 \left\{1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)} d\widehat{H}_n(s)}\right\} \cdot \int \frac{s}{(1 + c_0\gamma s)} d\widehat{G}_n(s), \tag{14}$$

$$\mathscr{V}(\widehat{H}_n, \gamma) \coloneqq \sigma^2\gamma \frac{\int \frac{s^2}{(1+c_0\gamma s)} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)} d\widehat{H}_n(s)}. \tag{15}$$

Definition B.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

$$\mathscr{B}(\lambda; \widehat{H}_n, \widehat{G}_n, \gamma) := \lambda^2 \left\|\theta^\star\right\|_2 (1 + \gamma m_{n,1}(-\lambda)) \int \frac{s}{[\lambda + (1 - \gamma + \gamma\lambda m_n(-\lambda))s]^2} d\widehat{G}_n(s), \tag{16}$$

$$\mathscr{V}(\lambda; \widehat{H}_n, \gamma) := \sigma^2\gamma \int \frac{s^2((1 - \gamma + \gamma\lambda m_n'(-\lambda)))}{[\lambda + (1 - \gamma + \gamma\lambda m_n(-\lambda))s]^2} d\widehat{H}_n(s). \tag{17}$$

B.4 Proof of Theorem 1

On solving (5) and (6), the closed form of the estimators $\hat{\theta}_0^{FA}$ and $\hat{\theta}_i^{FA}$ is given by

$$\begin{aligned}\hat{\theta}_0^{FA} &= \operatorname*{argmin}_\theta \sum_{j=1}^m p_j \frac{1}{2n_j} \left\|X_j\theta - \mathbf{y}_j\right\|_2^2 = \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j \frac{X_j^T \mathbf{y}_j}{n_j} \\ &= \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j \hat{\Sigma}_j \theta_j^\star + \left(\sum_{j=1}^m p_j \hat{\Sigma}_j\right)^{-1} \sum_{j=1}^m p_j \frac{X_j^T \xi_j}{n_j}\end{aligned} \tag{18}$$

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:

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

The idea is to show that the second term goes to 0 and use results from Hastie et al. (2019) 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 Hastie et al. (2019), 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 B.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 B.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 $t > 0$,

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

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 B.1). We have that

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

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

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

Lastly, we can bound this using Lemma B.2 as follows.

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

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 $\|E\|_2 \le t$(which means $\|E\|_2 \le t\sqrt{M}$).

Thus, we have

$$\begin{aligned}\|\Pi_i[\theta_0^\star - \theta_i^\star + E]\|_{\Sigma_i}^2 &\le \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + \|\Pi_i E\|_{\Sigma_i}^2 + 2\|\Pi_i E\|_{\Sigma_i}\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i} \\ &\le \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + M^2t^2 + 2tM^{3/2}r_i^2 \\ \|\Pi_i[\theta_0^\star - \theta_i^\star + E]\|_{\Sigma_i}^2 &\ge \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + \|\Pi_i E\|_{\Sigma_i}^2 - 2\|\Pi_i E\|_2\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i} \\ &\ge \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 - 2tM^2r_i^2\end{aligned}$$

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

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

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

Step 5 Using Theorem 3 of Hastie et al. (2019), 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 (14) 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 Hastie et al. (2019) 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.

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

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

$$
\begin{aligned}
&\sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}\operatorname{tr}\left(\Pi_i\Sigma_i\Pi_i\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\hat{\Sigma}_j\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right) \\
&\leq \sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}\|\Pi_i\|_1\|\Sigma_i\|_{\text{op}}\|\Pi_i\|_{\text{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\text{op}}\left\|\hat{\Sigma}_j\right\|_{\text{op}}\left\|\left(\sum_{j=1}^m p_j\hat{\Sigma}_j\right)^{-1}\right\|_{\text{op}} \\
&\leq C_3\sigma_{\max}^2\gamma_{\max}\left(\sum_{j=1}^m p_j^2\left\|\hat{\Sigma}_j\right\|_{\text{op}}\right), \qquad (19)
\end{aligned}
$$

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 B.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

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

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 Hastie et al. (2019), 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 (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 Hastie et al. (2019) we have $V_i(\hat{\theta}_i^{FA}|X)=\frac{\sigma_i^2}{n_i}\operatorname{tr}(\hat{\Sigma}_i^\dagger)\to\frac{\sigma_i^2}{\gamma_i-1}$.

B.5 Proof of Theorem 2

We use the global model from (5) and the personalized model from (7). The closed form of the estimators $\hat{\theta}_0^{FA}$ and $\hat{\theta}_i^R(\lambda)$ is given by

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

and

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

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

Bias:

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

The idea is to show that the second term goes to 0 and use results from Hastie et al. (2019) 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. 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 Hastie et al. (2019), 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

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

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 $\varepsilon > 0$

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

Step 5 Using Theorem 6 of Hastie et al. (2019), 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 (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 Corollary 5 of Hastie et al. (2019) 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

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

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

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

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 (19) and hence the rest of the analysis for this term is same as the one in the proof of FTFA (Appendix B.4).

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

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

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 Bai and Yin (1993), 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 Hastie et al. (2019), 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 (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 Theorem 1 of Hastie et al. (2019) we have $V_i(\hat{\theta}_i^R(\lambda)|X)=\frac{\sigma_i^2}{n_i}\operatorname{tr}((\hat{\Sigma}_i+\lambda I)^{-2}\hat{\Sigma}_i^\dagger\Sigma_i)\xrightarrow{p}\frac{\sigma_i^2}{\gamma_i-1}$.

B.6 Proof of Theorem 3

On solving (11) and (12), 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:

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

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

$$B_t := \left\{\left\|\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1} - \mathbb{E}\left[\sum_{j=1}^m p_j \frac{1}{n_j} X_j^T W_j^2 X_j\right]^{-1}\right\|_{\text{op}} > t\right\} \tag{20}$$

$$A_t := \left\{\left\|\left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j\right)^{-1} \sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \Delta_j\right\|_{\Sigma_i} > t\right\} \tag{21}$$

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, $B(\hat{\theta}_i^M(\alpha)|X) \le \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 + ct$ and $B(\hat{\theta}_i^M(\alpha)|X) \ge \|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2 - ct$

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

Step 5. Finally, using the asymptotic limit of $\|\Pi_i[\theta_0^\star - \theta_i^\star]\|_{\Sigma_i}^2$ from Theorem 1 of Hastie et al. (2019), we get the result.

We now give the detailed proof:

Step 1 Since $\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n_j} X_j^T W_j^2 X_j\right]\right) \ge \lambda_0$, it suffices to show by Lemma B.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.

$$\mathbb{P}(C_t) = \mathbb{P}\left(\left\|\sum_{j=1}^m p_j \left[\left[\hat{\Sigma}_j - 2\alpha^2\hat{\Sigma}_j^2 + \alpha^2\hat{\Sigma}_j^3\right] - \mathbb{E}\left[\frac{1}{n_j} X_j^T W_j^2 X_j\right]\right]\right\|_{\text{op}} > t\right)$$
$$\le \mathbb{P}\left(\left\|\sum_{j=1}^m p_j \left(\hat{\Sigma}_j - \mu_{1,j}\right)\right\|_{\text{op}} > t/3\right)$$
$$+ \mathbb{P}\left(\left\|2\alpha \sum_{j=1}^m p_j \left(\hat{\Sigma}_j^2 - \mu_{2,j}\right)\right\|_{\text{op}} > t/3\right) + \mathbb{P}\left(\left\|\alpha^2 \sum_{j=1}^m p_j \left(\hat{\Sigma}_j^3 - \mu_{3,j}\right)\right\|_{\text{op}} > t/3\right), \tag{22}$$

where $\mu_{p,j} := \mathbb{E}[\hat{\Sigma}_j^p]$. We repeatedly apply Lemma B.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 (22) 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 $\|Ax\|_2 \leq \|A\|_{\text{op}}\|x\|_2$. To handle the first term, we use Markov's inequality.

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

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

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

Variance:

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

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

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

$$
\begin{aligned}
&\sum_{j=1}^m \frac{p_j^2 \sigma_j^2}{n_j} \operatorname{tr}\left(\Pi_i \Sigma_i \Pi_i \left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \right)^{-1} \hat{\Sigma}_j (I - \alpha \hat{\Sigma}_j)^4 \left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \right)^{-1} \right) \\
&\leq \sum_{j=1}^m \frac{p_j^2 \sigma_j^2}{n_j} |\!|\!| \Pi_i |\!|\!|_1 \, |\!|\!| \Sigma_i |\!|\!|_{\mathrm{op}} \, |\!|\!| \Pi_i |\!|\!|_{\mathrm{op}} \left\| \left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \right)^{-1} \right\|_{\mathrm{op}} \left\| \hat{\Sigma}_j (I - \alpha \hat{\Sigma}_j)^4 \right\|_{\mathrm{op}} \left\| \left(\sum_{j=1}^m \frac{p_j}{n_j} X_j^T W_j^2 X_j \right)^{-1} \right\|_{\mathrm{op}} \\
&\leq C_7 \sigma_{\max}^2 \gamma_{\max} \left(\sum_{j=1}^m p_j^2 \left\| \hat{\Sigma}_j (I - \alpha \hat{\Sigma}_j)^4 \right\|_{\mathrm{op}} \right),
\end{aligned}
$$

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

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

Using Lemma B.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

$$\begin{aligned}
&2\,\text{tr}\left(\hat{\Sigma}_i^\dagger\frac{X_i^TW_i^2X_i}{n_i^2}\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\Pi_i\Sigma_i\right)\sigma_i^2 \\
&= 2\,\text{tr}\left(\Pi_i\Sigma_i\hat{\Sigma}_i^\dagger\frac{\hat{\Sigma}_i(I-\hat{\Sigma}_i)^2}{n_i}\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\right)\sigma_i^2 \\
&\leq \frac{2p_i\sigma^2}{n_i}\|\Pi_i\|_1\|\Sigma_i\|_{\text{op}}\left\|\hat{\Sigma}_i^\dagger\hat{\Sigma}_i\right\|_{\text{op}}\left\|(I-\hat{\Sigma}_i)^2\right\|_{\text{op}}\left\|\left(\sum_{j=1}^m\frac{p_j}{n_j}X_j^TW_j^2X_j\right)^{-1}\right\|_{\text{op}} \\
&\leq C_4p_i,
\end{aligned}$$

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

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

(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 Hastie et al. (2019), 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 (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 Hastie et al. (2019) we have $V_i(\hat{\theta}_i^M(\alpha)|X) = \frac{\sigma_i^2}{n_i}\operatorname{tr}(\hat{\Sigma}_i^\dagger) \to \frac{\sigma_i^2}{\gamma_i - 1}$.

B.7 Proof of Theorem 4

The solution to this minimization problem in (13) 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 Hastie et al. (2019) 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 Hastie et al. (2019), 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 Chen et al. (2012) 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 $\left\|T_j^{-1}\right\|_{\text{op}} = \left\|(\hat{\Sigma}_j + \lambda I)^{-1}\right\|_{\text{op}} \leq \frac{1}{\lambda}$ 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 B.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 $\mathbb{E}\left[\left\|\hat{\Sigma}^2\right\|_{\text{op}}\right] \leq \tau$ and there exists an $R \geq b$ such that $\mathbb{P}(\lambda_{\max}(\hat{\Sigma}) > R) \leq \frac{a^2}{8\tau}$, then*

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

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 $A := \{u^T\hat{\Sigma}u \geq \delta\}$ 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

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

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 B.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.

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

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 B.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

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

Lastly, we can bound this using Lemma B.2 as follows and using the fact that $\left\|\left\|T_j^{-1}\right\|\right\|_{\text{op}} \leq \frac{1}{\lambda}$.

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

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 $\|E\|_{\Sigma_i} \leq t$ (which means $\|E\|_2 \leq t\sqrt{M}$).

We can form the bounds

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

Since $t \in (0, 1]$, we have that $t^2 \leq t$ 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 $\varepsilon > 0$

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

Step 5 Using Theorem 6 of Hastie et al. (2019), 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 (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 Corollary 5 of Hastie et al. (2019) 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

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

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.

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

This is true because of Lemma B.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.

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

$\left\|Q^{-1}\right\|_{\text{op}}$ 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 Appendix B.4 to show that $\sum_{j=1}^m p_j^2 \left\|\hat{\Sigma}_j\right\|_{\text{op}} \xrightarrow{p} 0$.

(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.

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

$\left\|Q^{-1}\right\|_{\text{op}}$ 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 Hastie et al. (2019), 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 (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 Theorem 1 of Hastie et al. (2019) we have $V_i(\hat{\theta}_i^P(\lambda)|X) = \frac{\sigma_i^2}{n_i} \operatorname{tr}((\hat{\Sigma}_i + \lambda I)^{-2} \hat{\Sigma}_i^\dagger \Sigma_i) \xrightarrow{p} \frac{\sigma_i^2}{\gamma_i - 1}$.

Algorithm 3 Naive local training

Require: m: number of users, K: epochs
1: **for** $i \leftarrow 1$ to m **do**
2: Each client runs K epochs of SGM with personal stepsize α
3: **end for**

B.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. (18). 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 Hastie et al. (2019). The result regarding the estimator $\hat{\theta}_i^N(\lambda)$ is a direct consequence of Corollary 5 from Hastie et al. (2019).

B.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:

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

C Algorithm implementations

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

Algorithm 4 Federated Averaging McMahan et al. (2017)

Require: R: Communication Rounds, D: Number of users sampled each round, K: Number of local update steps, $\hat{\theta}_{0,0}^{FA}$: 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 $\hat{\theta}_{0,r}^{FA}$ to all clients in $\mathcal{S}_r$
  for $i \in \mathcal{S}_r$ **do**
    Set $\hat{\theta}_{i,r+1,0}^{FA} \leftarrow \hat{\theta}_{0,r}^{FA}$
    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 $g(\hat{\theta}_{i,r+1,k-1}^{FA}; \mathcal{D}_k^i) = \frac{1}{B}\sum_{S\in\mathcal{D}_k^i} \nabla F(\hat{\theta}_{i,r+1,k-1}^{FA}; S)$
      Set $\hat{\theta}_{i,r+1,k}^{FA} \leftarrow \hat{\theta}_{i,r+1,k-1}^{FA} - \eta g(\hat{\theta}_{i,r+1,k-1}^{FA}; \mathcal{D}_k^i)$
    end for
    Client i sends $\hat{\theta}_{i,r+1,K}^{FA}$ back to the server.
  end for
  Server updates the central model using $\hat{\theta}_{0,r+1}^{FA} = \sum_{j=1}^{D} \frac{n_j}{\sum_{j=1}^{D} n_j} \hat{\theta}_{i,r+1,K}^{FA}$.
end for
return $\hat{\theta}_{0,R}^{FA}$

Algorithm 5 FTFA

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 $\widehat{L}_i(\cdot)$ using $\hat{\theta}_0^{FA}$ as initial point with learning rate α and output $\hat{\theta}_{i,P}^{FA}$
end for
return $\hat{\theta}_{i,P}^{FA}$

D Experimental Details

D.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 McMahan et al. (2017), CIFAR-100 Krizhevsky and Hinton (2009), EMNIST Cohen et al. (2017), and Stack Overflow McMahan et al. (2019) datasets. We download all datasets using FedML APIs He et al. (2020) which in turn get their datasets from McMahan et al. (2019). 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.

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 Fallah et al. (2020)

Require: R: Communication Rounds, D: Number of users sampled each round, K: Number of local update steps, $\hat{\theta}^M_{0,0}(\alpha)$: 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 $\hat{\theta}^M_{0,r}(\alpha)$ to all clients in $\mathcal{S}_r$
  for $i \in \mathcal{S}_r$ **do**
    Set $\hat{\theta}^M_{i,r+1,0}(\alpha) \leftarrow \hat{\theta}^M_{0,r}(\alpha)$
    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 $g(\hat{\theta}^M_{i,r+1,k-1}(\alpha); \mathcal{D}^i_k) = \frac{1}{B}\sum_{S\in\mathcal{D}^i_k} \nabla F(\hat{\theta}^M_{i,r+1,k-1}(\alpha); S)$
      Set $\hat{\theta}^M_{i,r+1,k}(\alpha) \leftarrow \hat{\theta}^M_{i,r+1,k-1}(\alpha) - \alpha g(\hat{\theta}^M_{i,r+1,k-1}(\alpha); \mathcal{D}^i_k)$
    end for
    Client i sends $\hat{\theta}^M_{i,r+1,K}(\alpha)$ back to the server.
  end for
  Server updates the central model using $\hat{\theta}^M_{0,r+1}(\alpha) = \sum_{j=1}^{D} \frac{n_j}{\sum_{j=1}^{D} n_j} \hat{\theta}^M_{i,r+1,K}(\alpha)$.
end for
Server sends $\hat{\theta}^M_{0,R}(\alpha)$ to all clients
for $i \leftarrow 1$ to m **do**
  Run P steps of SGM on $\widehat{L}_i(\cdot)$ using $\hat{\theta}^M_0(\alpha)$ as initial point with learning rate α and output $\hat{\theta}^M_{i,P}(\alpha)$
end for
return $\hat{\theta}^M_{0,R}(\alpha)$

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

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 Reddi et al. (2021) (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 McMahan et al. (2019) — client division is described in Appendix F of Reddi et al. (2021). Instead of using 500 clients for training and 100 clients for testing as in Reddi et al. (2021), 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 Reddi et al. (2021), we also restrict to the

Algorithm 8 pFedMe Dinh et al. (2020)

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 $\theta_i(\hat{\theta}^P_{i,r+1,k-1}(\lambda)) = \text{argmin}_\theta \frac{1}{B}\sum_{S\in\mathcal{D}^i_k} \nabla F(\theta; S) + \frac{\lambda}{2}\left\|\theta - \hat{\theta}^P_{i,r+1,k-1}(\lambda)\right\|_2^2$
      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 $\hat{\theta}^P_{0,R}(\lambda)$

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 McMahan et al. (2019) 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.

D.2 Hyperparameter Tuning Details

D.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 McMahan et al. (2017). 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 He et al. (2016). 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 weiaicunzai (2020) to train our nets for 200 epochs.

EMNIST For this dataset, the architecture we use is similar to that found in Reddi et al. (2021); 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 Reddi et al. (2021). 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.

D.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 Dinh et al. (2020) 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 Reddi et al. (2021). 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 Fallah et al. (2020)). 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 Fallah et al. (2020)). 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 Dinh et al. (2020). 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

D.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.

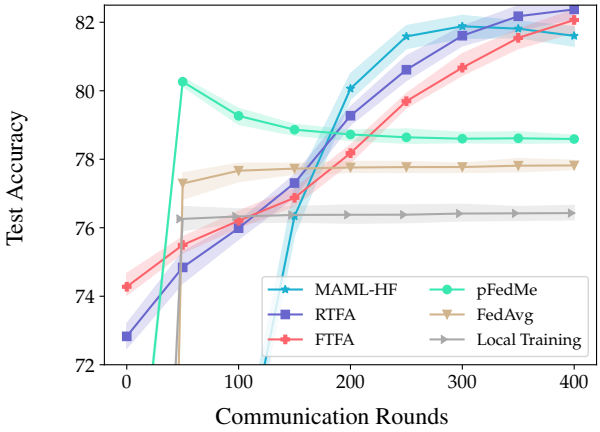


Figure 3: CIFAR-100. Best-average-worst intervals created from different random seeds.

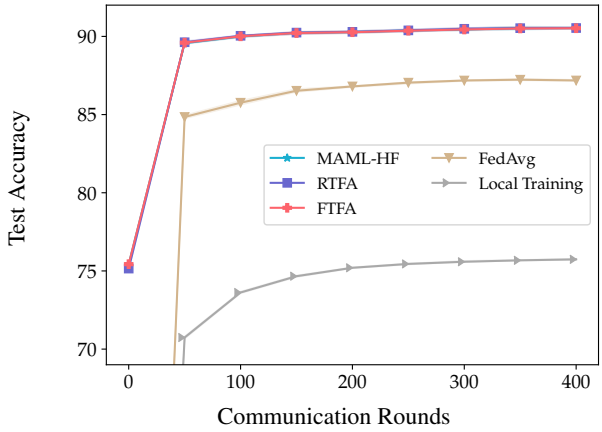


Figure 4: EMNIST. Best-average-worst intervals created from different train-val splits.

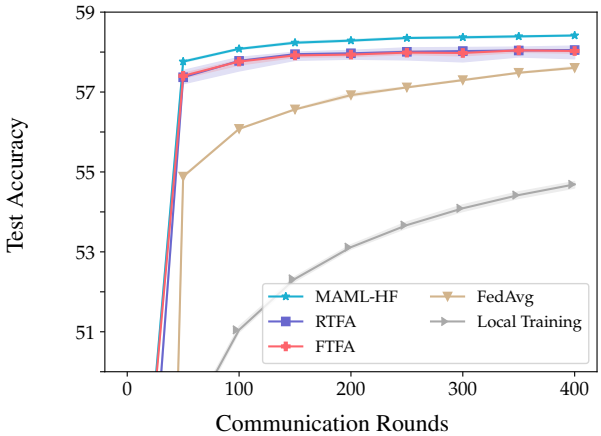


Figure 5: Shakespeare. Best-average-worst intervals created from different random train-val splits.

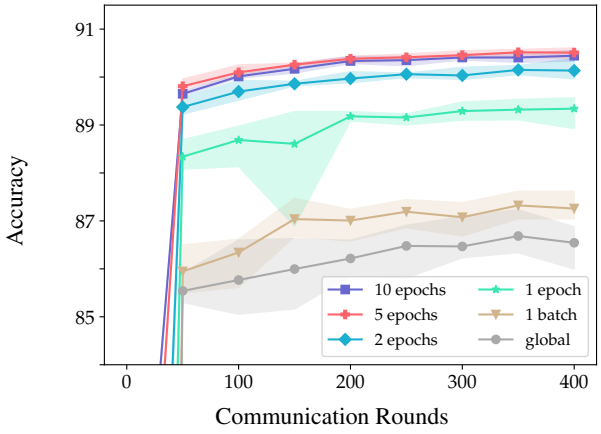


Figure 6: EMNIST. Gains of personalization for MAML-FL-FO

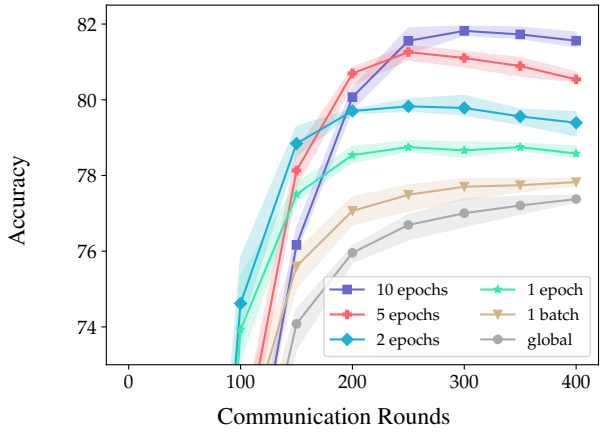


Figure 7: CIFAR. Gains of personalization for MAML-FL-FO

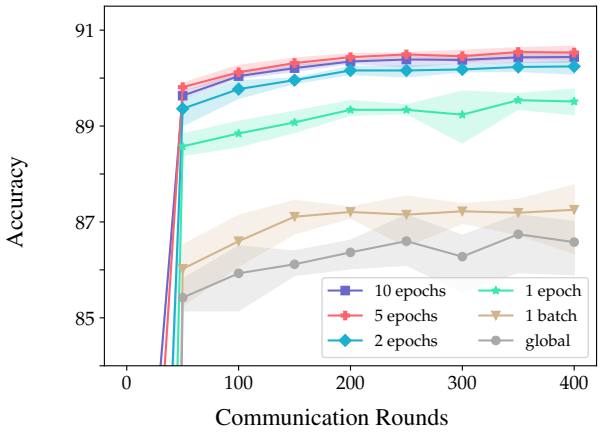


Figure 8: EMNIST. Gains of personalization for MAML-FL-HF

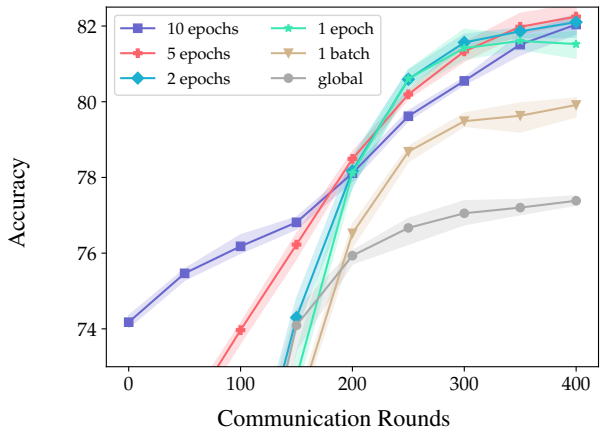


Figure 9: CIFAR. Gains of personalization for FTFA