Journal of Machine Learning Research 24 (2023) 1-48 Submitted 2/22; Published 5/23

A Non-parametric View of FedAvg and FedProx: Beyond Stationary Points

Lili Su L.SU@NORTHEASTERN.EDU
Electrical and Computer Engineering
Northeastern University

Jiaming Xu JIAMING.XU868@DUKE.EDU
The Fuqua School of Business
Duke University

Pengkun Yang* YANGPENGKUN@TSINGHUA.EDU.CN
Center for Statistical Science
Tsinghua University

Editor: Po-Ling Loh

Abstract

Federated Learning (FL) is a promising decentralized learning framework and has great potentials in privacy preservation and in lowering the computation load at the cloud. Recent work showed that FedAvg and FedProx – the two widely-adopted FL algorithms – fail to reach the stationary points of the global optimization objective even for homogeneous linear regression problems. Further, it is concerned that the common model learned might not generalize well locally at all in the presence of heterogeneity.

In this paper, we analyze the convergence and statistical efficiency of FedAvg and FedProx, addressing the above two concerns. Our analysis is based on the standard non-parametric regression in a reproducing kernel Hilbert space (RKHS), and allows for heterogeneous local data distributions and unbalanced local datasets. We prove that the estimation errors, measured in either the empirical norm or the RKHS norm, decay with a rate of $1/t$ in general and exponentially for finite-rank kernels. In certain heterogeneous settings, these upper bounds also imply that both FedAvg and FedProx achieve the optimal error rate. To further analytically quantify the impact of the heterogeneity at each client, we propose and characterize a novel notion-federation gain, defined as the reduction of the estimation error for a client to join the FL. We discover that when the data heterogeneity is moderate, a client with limited local data can benefit from a common model with a large federation gain. Two new insights introduced by considering the statistical aspect are: (1) requiring the standard bounded dissimilarity is pessimistic for the convergence analysis of FedAvg and FedProx; (2) despite inconsistency of stationary points, their limiting points are unbiased estimators of the underlying truth. Numerical experiments further corroborate our theoretical findings.

Keywords: Federated Learning, FedAvg and FedProx, non-parametric regression, convergence, statistical efficiency

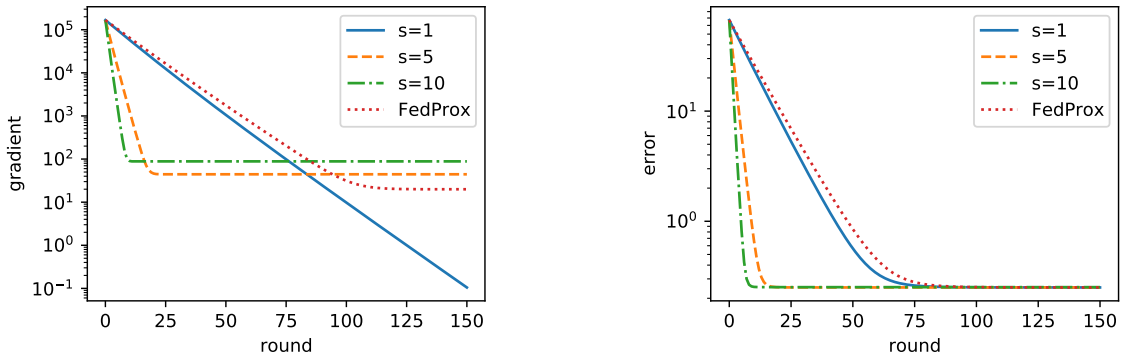
*. Correspondence author.

1. Introduction

Federated Learning (FL) is a rapidly developing decentralized learning framework in which a parameter server (PS) coordinates with a massive collection of end devices in executing machine learning tasks (Konečnỳ et al., 2016b,a; McMahan et al., 2017; Kairouz et al., 2021). Instead of uploading data to the PS, the end devices work at the front line in processing their own data and periodically report updates to the PS. On the one hand, FL has great potentials in privacy-preservation and in lowering the computation load at the cloud, both of which are crucial for modern machine learning applications. On the other hand, the defining characters of FL, i.e., costly communication, massively-distributed system architectures, highly unbalanced and heterogeneous data across devices, make it extremely challenging to understand the theoretical foundations of popular FL algorithms.

FedAvg and FedProx are two widely-adopted FL algorithms (McMahan et al., 2017; Li et al., 2020). They center around minimizing a global objective function $\ell(f) \triangleq \sum_{i=1}^{M} w_i \ell_i(f)$, where $\ell_i(f)$ is the local empirical risk of model f evaluated at client i's local data (Kairouz et al., 2021; McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020) and w_i is the weight assigned to client i. Specifically, in each round t, starting from f_{t-1} each client i computes its local model update $f_{i,t}$, which is then aggregated by the PS to produce f_t. To save communication, FedAvg only aggregates the local updates every s-th step of local gradient descent, where $s \geq 1$; when $s = 1$, FedAvg reduces to the standard stochastic gradient descent (SGD) algorithm. FedProx is a proximal-variant of FedAvg where the local gradient descent is replaced by a proximal operator.

Despite ample recent effort and some progress, the convergence and statistical efficiency of these two FL algorithms remain elusive (Kairouz et al., 2021; Pathak and Wainwright, 2020). In particular, existing attempts often impose restrictive assumptions such as balanced local data (Li et al., 2019), bounded gradients dissimilarity (i.e., $\nabla \ell_i \approx \nabla \ell$ for all i) (Li et al., 2020; Karimireddy et al., 2020; Stich, 2019; Zinkevich et al., 2010), and fresh data (Kairouz et al., 2021), and mostly ignore the impact of the model dimension (see Section 2 for detailed discussions). More concerningly, recent work (Pathak and Wainwright, 2020; Karimireddy et al., 2020; Zhao et al., 2018) showed, both experimentally and theoretically, that both FedAvg and FedProx fail to reach the stationary point of $\ell(f)$ even for the simple homogeneous linear regression problems. This observation is also illustrated in Fig. 1a, wherein we plot the trajectories of the gradient magnitudes $\|\nabla \ell(f_t)\|_2$ versus the communication rounds t under FedProx and FedAvg with aggregation period s being $1, 5, 10$, respectively. While the gradient magnitude of FedAvg with $s = 1$ quickly drops to 0, the gradient magnitudes under FedProx and FedAvg with $s = 5, 10$ stay well above 0. Does the failure of reaching stationary points lead to unsuccessful learning? We plot the evolution of the estimation error illustrated in Fig. 1b. Surprisingly, both FedAvg with $s = 5, 10$ and FedProx quickly converge to almost the same estimation error as FedAvg with $s = 1$ (i.e. the standard SGD). Moreover, the convergence time of FedAvg with $s = 5, 10$ shrinks roughly by a factor of s compared to $s = 1$, indicating that FedAvg enjoys significant saving in communication cost. *Why can FedAvg and FedProx achieve low estimation errors despite the failure of reaching stationary points?* The current paper aims to demystify this paradox. In particular, through statistical lens we discover that even though the limiting points of FedAvg and FedProx can be far from being stationary points of the global risk



(a) Plots of the gradient magnitudes versus the communication rounds

(b) Plots of the estimation errors versus the communication rounds

Figure 1: Plots of linear regression under FedProx and FedAvg. Experiment specifications: 25 clients, covariate dimension is 100, local sample size is 500, and observation noise follows $\mathcal{N}(0, 0.25I)$. Detailed specifications can be found in Section 7.1.

function, they are unbiased estimators and their variance can be effectively bounded by aggregating the local perturbations, thereby achieving low estimation errors.

Our study is further motivated by the concern on the lack of model personalization. Under both FedAvg and FedProx, a common model is trained but is used to serve all the clients without further tailoring to their local datasets. The tension between such standardized model and the data heterogeneity leads to ever-increasing concern on the generalization performance of the common model at different clients. In fact, on highly skewed heterogeneous data, evidence has been found suggesting that a common model could be problematic (Zhao et al., 2018; Fallah et al., 2020; Deng et al., 2020; Dinh et al., 2020). *Under what scenarios can a client benefit from a common model in the presence of heterogeneity?* The current paper seeks to address this question by quantifying the benefits and the impact of heterogeneity via a novel notion – *federation gain*.

Contributions In this paper, we analyze the convergence and statistical efficiency of FedAvg and FedProx by combining the optimization and statistical perspectives. Specifically, we assume that each client i has n_i local data points $\{x_{ij}, y_{ij}\}_{j=1}^{n_i}$ such that $y_{ij} = f_i^*(x_{ij}) + \xi_{ij}$, where f_i^* is the true model and ξ_{ij} is the noise. We allow n_i, x_{ij}, f_i^*, and ξ_{ij} to vary across different clients i, capturing the unbalanced data partition, covariate heterogeneity, and model heterogeneity, which are three most important types of heterogeneity (Kairouz et al., 2021). We base our analysis on the standard non-parametric regression setup and assume that f_i^* belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (Wainwright, 2019). Two new insights introduced by considering the statistical aspect are: (1) requiring the standard bounded dissimilarity is pessimistic for the convergence analysis of FedAvg and FedProx, (2) and, despite inconsistency of stationary points, their limiting points are unbiased estimators of the underlying truth.

We first show in Section 6.1 that the existence of heterogeneity does not prevent the convergence of f_t to a good common model f under FedAvg and FedProx. Specifically,

we show in Theorem 8 that with a proper early stopping rule, the estimation error decays with a rate of $1/t$. This further implies that: (i) in the presence of only unbalanced data partition and covariate heterogeneity where $f_i^* \equiv f^*$, f_t converges to f^*; (ii) with additional model heterogeneity, f_t approaches a common f that balances the model discrepancy up to a residual estimation error. For finite-rank kernel matrices, we further improve the convergence rate to be exponential without early stopping in Theorem 11. High probability bounds are derived in Theorem 10 for both light-tailed and heavy-tailed noises.

We show in Section 6.2 that the finite-rankness of the kernel also enables us to derive an explicit expression of the common $\bar{f}$ that perfectly balances out the heterogeneity across clients. In fact, in Theorem 12 we establish the convergence in RKHS norm - a strictly stronger notion of convergence. In particular, we show that the estimation error $\|f_t - \bar{f}\|_{\mathcal{H}}$ decays exponentially fast to $O(\sqrt{d/N})$ for $N = \sum_i n_i$, provided that the sample covariance matrix is well-conditioned. This error rate coincides with the minimax-optimal rate in the centralized setting. We further present two exemplary settings where the well-conditionedness assumption is shown to hold with high probability.

Moreover, we bound the difference $\|\bar{f} - f_j^*\|_{\mathcal{H}}$, showing that when the model heterogeneity is moderate, a client with limited local data can still benefit from a common model. To formally study the benefits of joining FL, in Section 6.3 we propose and characterize the *federation gain*, defined as the reduction of the estimation error for a client to join the FL. We establish a threshold on the heterogeneity in terms of model dimensions and local data sizes under which the federation gain exceeds one. Our characterization of federation gain serves as a guidance in encouraging end devices to make their participation decisions.

Finally, using numerical experiments, we corroborate our theoretical findings. Specifically, in Section 7.1 we demonstrate that both FedAvg and FedProx can achieve low estimation errors despite the failure of reaching stationary points. The same phenomenon is found to still persist when minibatches are used in local updates. In Sections 7.2 and 7.3, we adapt the experiment setup to allow for unbalanced local data partition, covariate heterogeneity, and model heterogeneity. For both FedAvg and FedProx, we empirically observe that the federation gains are large when a client has a small local data size and the data heterogeneity is moderate, matching our theoretical predictions. In Section 7.4, we fit nonlinear models and confirm that both FedAvg and FedProx continue to attain nearly optimal estimation rates.

2. Related work

2.1 On Convergence of FedAvg and FedProx

There is vast literature on the convergence of FedAvg/FedProx and related FL algorithms. Here we only hope to cover a fraction of them that are closest to our work. FedAvg has emerged as the algorithm of choice for FL (Kairouz et al., 2021; Karimireddy et al., 2020). Both empirical successes and failures of convergence have been reported (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020); however, the theoretical characterization of its convergence (for general s) turns out to be notoriously difficult.

Convergence in the homogeneous settings. With *i.i.d.* data, convergence is shown in Zinkevich et al. (2010); Stich (2019) under the name local SGD. In particular, Zinkevich

et al. (2010) proves the asymptotic convergence. Convergence in the non-asymptotic regime is derived in Stich (2019) under assumptions of strong convexity and bounded gradients.

Convergence under bounded dissimilarity conditions. The proof techniques of Zinkevich et al. (2010); Stich (2019) are adapted to data heterogeneity setting by postulating the variances of gradients are bounded or the dissimilarities of gradients/Hessians are uniformly bounded (Karimireddy et al., 2020; Li et al., 2020). Since most existing analyses focus on the optimization respective only, the scaling of the dissimilarity bounds in the local data volume n_i, the model dimension d, and the total number of data N is often overlooked. When taking such scaling into account (which is the common practice of statistical learning), those assumptions turn out to be stringent, especially under the popular federated learning scenario that the client only has a small local dataset despite the collectively large global dataset. For example, Karimireddy et al. (2020) adopted a popular (B, G)-bounded gradient dissimilarity condition, i.e., $\frac{1}{M}\sum_{i=1}^{M}\|\nabla\ell_i(\theta)\|^2 \leq B^2\|\nabla\ell(\theta)\|^2 + G^2$, $\forall\theta$, where $B \geq 0$ and $G \geq 0$. Consider the linear regression problem: each client only contributes a single data point (x_i, y_i) for $i = 1, \ldots, M$, where $x_i \sim N(0, I_d)$ and $y_i = x_i^\top\theta^* + \sigma\zeta_i$; the local objective function is $\ell_i(\theta) = \frac{1}{2}\|x_i^\top\theta - y_i\|^2$. In the noiseless case, $\nabla\ell_i(\theta) = x_i(x_i^\top\theta - y_i) = x_i x_i^\top(\theta - \theta^*)$ and $\nabla\ell(\theta) = \frac{1}{M}\sum_{i=1}^{M} x_i x_i^\top(\theta - \theta^*)$. For the (B, G)-bounded gradient dissimilarity to hold, we have $B = \Omega(\sqrt{d})$ with high probability. Therefore, Theorem 1 of Karimireddy et al. (2020) implies that to achieve estimation error of ϵ, the number of communication rounds grows linearly with d.

In contrast, by leveraging the underlying statistical structure we obtain faster rates that are independent of B and G. One new insight is that existing bounded dissimilarity conditions are unnecessary or too pessimistic for the validity of FedAvg and FedProx. We carefully characterize the aggregate effect of the local deviations across all clients. In particular, our analysis shows that even when each individual client may deviate greatly from the global model, the collective deviation over the whole system remains small.

Convergence to the surrogate minimizer. Instead of imposing the bounded dissimilarity assumptions, a different line of work shows that FedAvg and its variants converge to the minimizer of a suitably defined *surrogate* function. In particular, they unroll the local updates and characterize the effect of multiple local updates via some *distortion matrix* Q_i (see (Charles and Konečnỳ, 2020, eq. (8))). They introduce a local surrogate function $\widetilde{\ell}_i$ for each client i in terms of Q_i. The multiple local updates can be equivalently viewed as one step of gradient descent on the surrogate function $\widetilde{\ell}_i$. Therefore, the global update reduces to performing one step of SGD on the global surrogate function $\widetilde{\ell}$, which is an average of $\widetilde{\ell}_i$ across all clients i. In this way, the global convergence analysis reduces to the standard analysis of SGD on the surrogate function $\widetilde{\ell}$. They further characterize the discrepancy between the surrogation minimizer and the true minimizer of the original global risk function ℓ. However, crucially Charles and Konečnỳ (2021) and Charles and Konečnỳ (2020) require the local objective functions ℓ_i at all clients to be strongly convex. However, even in the linear regression setting, if the number of local data points is smaller than the model dimension at some client i, then the local objective function ℓ_i is convex but not strongly convex. In this case, the strong-convexity parameter μ defined in (Charles and Konečnỳ, 2020, Assumption 2)) is equal to 0; as a result, the convergence bounds obtained in (Charles and Konečnỳ, 2020, Theorem 26, 27) become vacuous, so do the bounds on the distance

between the surrogate minimizer and true risk minimizer derived in (Charles and Konečnỳ, 2020, Theorem 16, 17). Closely inspecting their analysis in (Charles and Konečnỳ, 2020, Section 5), we can find that their analysis is limited and loose because they only characterize the effect of multiple local steps separately for each client i and fails to account for the *aggregate* effect combined over all clients i.

In sharp contrast, our analysis (Proposition 3) carefully characterizes the *aggregate* effect of the multiple local steps over all client i, which enables us to relax the strong convexity requirements on each local loss function. Note that Proposition 3, albeit simple to state, is by no means easy to obtain. Its derivation crucially relies on the novel matrix identities that we derive in Lemma 2. It allows us to prove the global convergence in terms of the eigenvalues of $N \times N$ matrix $K_{\mathbf{x}}P$ instead of the $n_i \times n_i$ local matrices $K_{\mathbf{x}_i}P_{ii}$. In particular, using Proposition 3, Proposition 7 and Theorem 12 yield the convergence of the prediction error and the model parameters, even when some local clients have only a few data points and their local loss functions are not strongly convex.

Convergence via "client-drift" corrections. Recently, several novel FL algorithms based on "client-drift" corrections were proposed, and were shown to converge to the correct stationary point (Karimireddy et al., 2020; Mitra et al., 2021; Gorbunov et al., 2021). In particular, by using variance-reduction or gradient-tracking techniques to correct the "client-drift" effect, the desired stationary points become fixed points of the local updates. Therefore, the global convergence are obtained even without the bounded dissimilarity conditions on gradients/Hessians. Despite the elegancy of these results, they focus exclusively on the optimization perspective, characterizing the training errors only. In passing, we further remark that they often assume fresh data is drawn in each update for technical convenience (Kairouz et al., 2021). Both the randomness in the design matrix (which is harder to handle) and the impacts of the covariate dimension are mostly neglected.

Instead of forcing the algorithms to converge to the correct stationary points, by introducing the statistical perspective, we obtain the insight that even though the limiting points of FedAvg and FedProx can be far from any stationary points of the global risk function, they are unbiased estimators of the underlying truth f^*.

2.2 Personalization

In the context of Model Agnostic Meta Learning (MAML), personalized Federated Learning is investigated both experimentally (Chen et al., 2018; Jiang et al., 2019) and theoretically (Fallah et al., 2020; Lin et al., 2020). MAML-type personalized FL finds a *shared initial model* that a participating device can quickly get personalized by running a few updates on its local data. Adaptive Personalized Federated Learning (APFL) is proposed in Deng et al. (2020) under which each end device trains its local model while contributing to the global model. A personalized model is then learned as a mixture of optimal local and global models. Other personalization techniques include model division, contextualization, and multi-task learning. Due to space limitation, readers are referred to Kairouz et al. (2021) for details. In this paper, we show that without introducing additional personalization techniques, an end device can still benefit from joining FL under certain mild conditions.

3. Problem Formulation

System model. A federated learning system consists of a parameter server (PS) and M clients. Each client $i \in \{1, \cdots, M\} \triangleq [M]$ locally keeps its personal data $\mathcal{S}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$, where $n_i = |\mathcal{S}_i|$ is referred to as local data volume of client i. Let $N \triangleq \sum_{i=1}^{M} n_i$. It is possible that $n_i \neq n_j$ for some $i \neq j$, i.e., the local data volume at different clients could be highly unbalanced. The magnitude of n_i varies with different real-world applications: when $\mathcal{S}_i$ are records of recently browsed websites, n_i is typically moderate; when $\mathcal{S}_i$ are records of recent places visited by walk in pandemic, the volume of $\mathcal{S}_i$ is low. Observing this, in this work, we consider a wide range of n_i which covers both the small and moderate n_i regions as special cases.

Data heterogeneity. We consider both covariate heterogeneity (a.k.a. covariate shift) and response heterogeneity (a.k.a. concept shift) (Kairouz et al., 2021). Formally, at each client i,

$$y_{ij} = f_i^*(x_{ij}) + \xi_{ij}, \quad 1 \leq j \leq n_i, \tag{1}$$

where f_i^* is the underlying mechanism governing the true responses, $x_{ij} \in \mathcal{X}$ is the covariate, and ξ_{ij} is the observation noise. We impose the mild assumptions that $\xi_{i1}, \cdots, \xi_{in_i}$ are independent yet possibly non-identically distributed, zero-mean, and have variance up to σ^2.

Non-parametric regression. We base our analysis on the standard non-parametric regression setup and assume that f_i^* belongs to a reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}$ with a defining positive semidefinite kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. For completeness, we present the relevant fundamentals of RKHS (see (Wainwright, 2019, Chapter 12) for an in-depth exposition). Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denote the inner product of the RKHS $\mathcal{H}$. At any $x \in \mathcal{X}$, $k(\cdot, x)$ acts as the representer of evaluation at x, i.e.,

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x), \qquad \forall f \in \mathcal{H}. \tag{2}$$

Let $\|g\|_{\mathcal{H}} = \sqrt{\langle g, g \rangle_{\mathcal{H}}}$ denote the norm of function g in $\mathcal{H}$. For a given distribution $\mathbb{P}$ on $\mathcal{X}$, let $\|g\|_2 = \left(\int_{\mathcal{X}} g(x)^2 d\mathbb{P}(x)\right)^{1/2}$ denote the norm in $L^2(\mathbb{P})$. In this paper, we take the following minimal assumptions that are common in literature (Wainwright, 2019). We assume that $\mathcal{X}$ is compact, k is continuous, $\sup_{x \in \mathcal{X}} k(x, x) < \infty$, and that $\int_{\mathcal{X} \times \mathcal{X}} k^2(x, z) d\mathbb{P}(x) d\mathbb{P}(z) < \infty$. Mercer's theorem shows that such kernel k admits an expansion

$$k(x, z) = \sum_{\ell=1}^{\infty} \mu_\ell \varphi_\ell(x) \varphi_\ell(z), \tag{3}$$

where $\{\varphi_\ell\}$ forms an orthonormal basis of $L^2(\mathbb{P})$, and $\{\mu_\ell\}$ are the non-negative eigenvalues. Notably, $\langle \varphi_\ell, \varphi_{\ell'} \rangle_{\mathcal{H}} = 0$ for $\ell \neq \ell'$ and $\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{H}} = \frac{1}{\mu_\ell}$ for all ℓ such that $\mu_\ell \neq 0$ (Wainwright, 2019, Corollary 12.26). Define the feature mapping $\phi : \mathcal{X} \mapsto \ell^2(\mathbb{N})$ as $\phi(x) = \left[\sqrt{\mu_1}\varphi_1(x), \sqrt{\mu_2}\varphi_2(x), \ldots\right]$, where $\ell^2(\mathbb{N})$ denotes the space of square-summable sequences. Then for any $f \in \mathcal{H}$ with $f(x) = \sum_{\ell=1}^{\infty} \beta_\ell \varphi_\ell(x)$ such that $\sum_{\ell=1}^{\infty} \frac{\beta_\ell^2}{\mu_\ell} < \infty$[1], we have

1. With a little abuse of notation, $\sum_{\ell=1}^{\infty} \frac{\beta_\ell^2}{\mu_\ell}$ sums over all ℓ such that $\mu_\ell > 0$.

$f(x) = \sum_{\ell=1}^{\infty} \theta_\ell \phi_\ell(x)$, where $\theta_\ell = \frac{\beta_\ell}{\sqrt{\mu_\ell}}$. Hence,

$$\|f\|_2^2 = \sum_{\ell=1}^{\infty} \beta_\ell^2, \quad \text{and} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \theta_\ell^2. \tag{4}$$

The above non-parametric setting can be used to approximate more sophisticated settings. In particular, it is applicable to random kernels by using the corresponding eigenvalues and thus covers the neural tagent kernels (NTKs) to approximate the NNs in certain regimes. For instance, the NTK for two-layer NNs is $k(x,y) = \mathbb{E}_{w\sim\mathcal{N}(0,I)}\left[h'(w^\top x)h'(w^\top y)\right]$, where $h(x)$ is the activation function. The setup can be further extended to the non-realizable case that $f_i^* \notin \mathcal{H}$. Specifically, considering the best approximation of f_i^* by functions in $\mathcal{H}$, we get an additional term on the misspecification error in (1). Our analysis in the paper can then be extended by incorporating this misspecification error as the observation bias.

Additional notation Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i}) \in \mathcal{X}^{n_i}$ denote the covariate of the local data at client i; all data covariate $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_M) \in \mathcal{X}^N$. Similarly, let $y_i \in \mathbb{R}^{n_i}$ and $y \in \mathbb{R}^N$ be the vectors that stack the responses of the local data at client i and the total data, respectively. For $a, b \in \mathbb{R}^d$, let $a \cdot b \triangleq \sum_{i=1}^d a_i b_i$. Given a multivariate function $f : \mathcal{X} \mapsto \mathbb{R}^d$, we use f_j to denote the j-th components of f; for $a \in \mathbb{R}^d$, define $a \cdot f : \mathcal{X} \mapsto \mathbb{R}$ as $(a \cdot f)(x) = a \cdot f(x)$; for $A \in \mathbb{R}^{n\times d}$, define $Af : \mathcal{X} \mapsto \mathbb{R}^n$ as $(Af)_i(x) \triangleq \sum_{j=1}^d A_{ij} f_j(x)$ for $i \in [n]$. For $x \in \mathcal{X}$, let $k_x \triangleq k(\cdot, x) : \mathcal{X} \mapsto \mathbb{R}$; for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $k_{\mathbf{x}} \triangleq (k_{x_1}, \ldots, k_{x_n}) : \mathcal{X} \mapsto \mathbb{R}^n$, and $K_{\mathbf{x}}$ be the normalized Gram matrix of size $n \times n$ with $(K_{\mathbf{x}})_{ij} = \frac{1}{n} k(x_i, x_j)$. Given a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $f(\mathbf{x}) \triangleq (f(x_1), \ldots, f(x_n))$; in particular, when $\mathcal{Y} = \mathbb{R}^d$, let $f(\mathbf{x})$ be a matrix of size $n \times d$ that stacks $f(x_i)$ in rows. For an operator $\mathcal{L} : \mathcal{H} \mapsto \mathcal{H}$ and $f = (f_1, \ldots, f_n) \in \mathcal{H}^n$, let $\mathcal{L}f \triangleq (\mathcal{L}f_1, \ldots, \mathcal{L}f_n)$. Let $\|v\|_2$ and $\|V\|_2$ denote the ℓ^2 norm of a vector v and the spectral norm of matrix V, respectively. The operator norm is denoted by $\|\cdot\|_{\text{op}}$. For a positive definite matrix A, let $A^{1/2}$ denote the unique square root of A. Throughout this paper, we use $c, c_1, ...$ to denote absolute constants. For ease of exposition, the specific values of these absolute constants might vary across different concrete contexts in this paper.

4. FedAvg and FedProx

FedAvg can be viewed as a communication-light implementation of the standard SGD. Different from the standard SGD, wherein the updates at different clients are aggregated right after *every* local step, in FedAvg the local updates are only aggregated after *every s-th* local step, where $s \geq 1$ is an algorithm parameter. FedProx is a distributed proximal algorithm wherein a round-varying proximal term is introduced to control the deviation of the local updates from the most recent global model.

Recall from Section 1 that $\ell_i(f) = \frac{1}{2n_i} \sum_{j=1}^{n_i} \left(f(x_{ij}) - y_{ij}\right)^2$ is the local empirical risk function for each $f \in \mathcal{H}$. Let f_t denote the global model at the end of the t-th communication round, and let f_0 denote the initial global model. At the beginning of each round $t \geq 1$, the PS broadcasts f_{t-1} to each of the M clients. At the end of round t, upon receiving the

local updates $f_{i,t}$ from each client i, the PS updates the global model as

$$f_t = \sum_{i=1}^{M} w_i f_{i,t}\,, \tag{5}$$

where $w_i = \frac{n_i}{N}$ – recalling that N is the number of all the data tuples in the FL system. The local updates $f_{i,t}$ under FedAvg and FedProx are obtained as follows.

FedAvg From f_{t-1} each client i runs s local gradient descent steps on $\ell_i(f)$, and reports its updated model to the PS. Concretely, we denote the mapping of one-step local gradient descent by $\mathcal{G}_i(f) = f - \eta\nabla\ell_i(f)$, where $\eta > 0$ is the stepsize. After s local steps, the locally updated model at client i is given by

$$f_{i,t} = \mathcal{G}_i^s(f_{t-1}).$$

FedProx From f_{t-1}, each client i locally updates the model as

$$f_{i,t} \;=\; \arg\min_{f\in\mathcal{H}}\; \ell_i(f) + \frac{1}{2\eta}\|f - f_{t-1}\|_{\mathcal{H}}^2. \tag{6}$$

Notably, $\eta > 0$ controls the regularization and can be interpreted as a step size in the FedProx: As η increases, the penalty for moving away from f_{t-1} decreases and hence the local update $f_{i,t}$ will be farther way from f_{t-1}. In practice, the local optimization problem in (6) might not be solved exactly in each round. We would like to study the impacts of inexactness of solving (6) in future work.

5. Recursive Dynamics of FedAvg and FedProx

In this section, we derive expressions for the recursive dynamics of f_t in (5) under FedAvg and FedProx, respectively. All missing proofs of this section can be found in Appendix A. We first introduce two local linear operators. Within iteration t of FedAvg, the one-step local gradient descent on client i is given by an affine mapping

$$\mathcal{G}_i(f_{t-1}) = f_{t-1} - \frac{\eta}{n_i}\sum_{j=1}^{n_i}(f_{t-1}(x_{ij}) - y_{ij})k_{x_{ij}} = \mathcal{L}_i f_{t-1} + \frac{\eta}{n_i}\sum_{j=1}^{n_i} y_{ij}k_{x_{ij}}, \tag{7}$$

where $\mathcal{L}_i$ denotes the local operator

$$\mathcal{L}_i f \;\triangleq\; f - \frac{\eta}{n_i}\sum_{j=1}^{n_i} f(x_{ij})k_{x_{ij}}. \tag{8}$$

For FedProx, the global model dynamics involves the inverse of local operator $\widetilde{\mathcal{L}}_i$ where

$$\widetilde{\mathcal{L}}_i f \triangleq f + \frac{\eta}{n_i}\sum_{j=1}^{n_i} f(x_{ij})k_{x_{ij}}.$$

Recall that $w_i = \frac{n_i}{N}$. The following proposition characterizes the dynamics of f_t.

Proposition 1 *The global model f_t satisfies the following recursion:*

$$f_t = \mathcal{L} f_{t-1} + y \cdot \Psi\,, \tag{9}$$

where $\Psi_i : \mathcal{X} \mapsto \mathbb{R}^{n_i}$ for $i \in [M]$, $\Psi = (w_1\Psi_1, \ldots, w_M\Psi_M) : \mathcal{X} \mapsto \mathbb{R}^N$, and

$$\mathcal{L} = \begin{cases} \sum_{i=1}^M w_i \mathcal{L}_i^s \\ \sum_{i=1}^M w_i \widetilde{\mathcal{L}}_i^{-1} \end{cases} \quad \text{and} \quad \Psi_i = \begin{cases} \frac{\eta}{n_i} \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau k_{\mathbf{x}_i} & \text{for FedAvg,} \\ \frac{\eta}{n_i} \widetilde{\mathcal{L}}_i^{-1} k_{\mathbf{x}_i} & \text{for FedProx.} \end{cases} \tag{10}$$

FedAvg with $s = 1$ coincides with the standard distributed gradient descent, which naturally fuses a global model as $\sum_i w_i \mathcal{G}_i$ effectively aggregates all local data. However, for $s > 1$, analyzing the dynamics in Proposition 1 directly is challenging as the local updates involve high-order operators $\mathcal{G}_i^s$. This makes the global model fusion more difficult because $\sum_i w_i \mathcal{G}_i^s$ aggregates local progress in a nontrivial manner and further drives f_t away from the stationary points of the global objective function $\ell(f)$. Similar challenges also appear in FedProx due to the inverse of $\widetilde{\mathcal{L}}_i$.

Fortunately, from Proposition 1 we can derive compact expressions of the evolution of the in-sample prediction values under FedAvg and FedProx, which serve as the foundation for the convergence analysis in Section 6. Under FedAvg with $s = 1$, it is well-known in the literature of *kernel methods* (Hastie et al., 2009, Chapter 12) (and also follows from (7)) that

$$f_t(\mathbf{x}) = (I - \eta K_{\mathbf{x}}) f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} y, \tag{11}$$

For $s > 1$, there is no immediate extension of (11) to $s > 1$ and to FedProx. The key step in our derivation is a set of identities for $\mathcal{L} f_{t-1}(\mathbf{x})$ and $\Psi(\mathbf{x})$, which are stated in the next lemma. Those identities are also used in our convergence proofs, and could be of independent interest to a broader audience.

Lemma 2 *For any $f \in \mathcal{H}$, the following identities are true:*

$$\begin{aligned} &\Psi = \frac{\eta}{N} P k_{\mathbf{x}}, && \Psi(\mathbf{x}) = \eta K_{\mathbf{x}} P, \\ &f(\mathbf{x}) \cdot \Psi = f - \mathcal{L} f, && \mathcal{L} f(\mathbf{x}) = (I - \eta K_{\mathbf{x}} P) f(\mathbf{x}), \end{aligned}$$

where $P \in \mathbb{R}^{N \times N}$ is a block diagonal matrix whose i-th diagonal block of size $n_i \times n_i$ is

$$P_{ii} = \begin{cases} \sum_{\tau=0}^{s-1} [I - \eta K_{\mathbf{x}_i}]^\tau & \text{for FedAvg,} \\ [I + \eta K_{\mathbf{x}_i}]^{-1} & \text{for FedProx.} \end{cases} \tag{12}$$

Proposition 3 *The prediction value satisfies the following recursion:*

$$f_t(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y. \tag{13}$$

Proof By Proposition 1,

$$f_t(x_{ij}) = \mathcal{L} f_{t-1}(x_{ij}) + \Psi(x_{ij}) \cdot y.$$

Consequently, applying Lemma 2 yields that

$$f_t(\mathbf{x}) = \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y.$$

■

The dynamics of the model f_t in (9) and the corresponding in-sample prediction values $f_t(\mathbf{x})$ in (13) are both governed by linear time invariant (LTI) systems with $y \in \mathbb{R}^N$ as the constant system input. Those autoregressions converge if all eigenvalues of $\mathcal{L}$ and $I - \eta K_{\mathbf{x}} P$ are less than one in absolute value, and locations of the eigenvalues such as the distance to the unit circle have important implications for the model evolution (Brockwell and Davis, 2009, 2016). Although it is challenging to characterize the eigenvalues of $\mathcal{L}$ due to the insufficiency of local data, system heterogeneity, and the involved aggregation of high-order or inverse operators, the eigenvalues of $I - \eta K_{\mathbf{x}} P$ in the evolution of prediction values are more tractable.

Compared with the classical kernel gradient descent, here the crucial difference is the effect of matrix P, which arises from multiple local updates of FedAvg and the proximal term in the local update of FedProx. In particular, it is essential to characterize the spectrum of $K_{\mathbf{x}} P$. When P is positive definite, analagous to the normalized graph Laplacians (see e.g. (Von Luxburg, 2007, Section 3.2)), the eigenvalues of $K_{\mathbf{x}} P$ coincide with those of the symmetric matrix $P^{1/2} K_{\mathbf{x}} P^{1/2}$, and hence must be real and non-negative. It follows that the eigenvalues of $I - \eta K_{\mathbf{x}} P$ are no more than 1. Define

$$\gamma \triangleq \eta \max_{i \in [M]} \|K_{\mathbf{x}_i}\|_2 .$$

By the block diagonal structure of P, $\gamma < 1$ guarantees that $P \succ 0$, and furthermore both $\mathcal{L}$ and $I - \eta K_{\mathbf{x}} P$ have non-negative eigenvalues only.

Lemma 4 *If $\gamma < 1$, then all eigenvalues of $\mathcal{L}$ and $I - \eta K_{\mathbf{x}} P$ are within $[0, 1]$.*

Throughout this paper, we assume $\gamma < 1$[2]. The local update and the global aggregation are stable if P is well-conditioned, e.g., $P = I$ for the gradient descent. In general, we have the following upper bound on the condition number of P.

Lemma 5 *Suppose that $\gamma < 1$.*

$$\|P\|_2 \left\|P^{-1}\right\|_2 \leq \kappa \triangleq \begin{cases} \frac{\gamma s}{1-(1-\gamma)^s} & \text{for FedAvg,} \\ 1 + \gamma & \text{for FedProx.} \end{cases} \tag{14}$$

Moreover, we have

$$\Lambda_i \in \begin{cases} [\lambda_i s/\kappa, \lambda_i s] & \text{for FedAvg,} \\ [\lambda_i/\kappa, \lambda_i] & \text{for FedProx,} \end{cases} \tag{15}$$

where $\lambda_i \geq 0$ and $\Lambda_i \geq 0$ are the i-th largest eigenvalue of $K_{\mathbf{x}}$ and $K_{\mathbf{x}} P$, respectively.

2. For FedProx, our results continue to hold without any assumption on γ. In particular, the matrix P is always positive definite regardless of γ. In a sense, FedProx is more stable than FedAvg. Yet, the conditioning of P degrades with γ.

From Lemma 5 and the definition of γ, κ – the upper bound to the conditioning number of P – approaches 1 with properly chosen small learning rate η and small number of local steps s in FedAvg. Larger η and s accelerate the optimization and reduce the communication rounds at the expense of worsening the conditioning of P and incurring a larger statistical error; this tradeoff will be quantified in Section 6.

Remark 6 *When k is a neural tangent kernel (NTK) (Du et al., 2018, 2019), the kernel matrix $K_{\mathbf{x}}$ is positive definite provided that the input training data is non-parallel. Therefore, the series of the gradient descent* (11) *given by*

$$f_t(\mathbf{x}) = (I - \eta K_{\mathbf{x}})^t f_0(\mathbf{x}) + (I - (I - \eta K_{\mathbf{x}})^t) y$$

converge to y and thus attain zero training error for a properly small learning rate η. It immediately follows from (13) *that both FedAvg and FedProx attain zero training error for NTKs.*

6. Convergence Results

In this section we present our results on the convergence of FedAvg and FedProx in terms of both the global model f_t and the model coefficients θ_t – recalling that $f_t = \langle \phi, \theta_t \rangle$, where ϕ is the feature mapping. For ease of exposition, we state our results for FedAvg and FedProx in a unified and compact form with s as one characterizing parameter. Recall that s is the algorithm parameter of FedAvg only. To recover the formal statements and involved quantities for FedProx, we only need to set $s = 1$.

- To study the convergence of f_t, we compare f_t with any given function $f \in \mathcal{H}$ at the *observed covariates*. In particular, we study the prediction error, as measured in the (empirical) $L^2(\mathbb{P}_N)$ norm, that is

$$\|f_t - f\|_N^2 \triangleq \frac{1}{N} \|f_t(\mathbf{x}) - f(\mathbf{x})\|_2^2 = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{n_i} (f_t(x_{ij}) - f(x_{ij}))^2 . \tag{16}$$

  Note that the $L^2(\mathbb{P}_N)$ norm is a commonly adopted performance metric in regression (See e.g. (Wainwright, 2019, Sections 7.4 and 13.2)). Different from the training error $(1/N)\|f_t(\mathbf{x}) - y\|_2^2$, the prediction error under (16) is able to reflect the over-fitting phenomenon. Concretely, when an algorithm is over-fitting noises, the training error could approach 0 whereas the prediction error under (16) would stay large.

- When the RKHS $\mathcal{H}$ is of finite dimension, we further study the convergence of f_t in the RKHS $\mathcal{H}$ norm. This is equivalent to the convergence of the model coefficient θ_t in the L^2 norm in view of (4). One can readily check that the convergence of the model f_t in the $\mathcal{H}$ norm is stronger than that in the $L^2(\mathbb{P}_N)$ norm.[3]

3. By the reproducing property of kernels (i.e., the identity (2)) and the Cauchy-Schwarz inequality, we have

$$(f_t(x) - f(x))^2 = \langle f_t - f, k_x \rangle_{\mathcal{H}}^2 \le \|f_t - f\|_{\mathcal{H}}^2 \|k_x\|_{\mathcal{H}}^2 = \|f_t - f\|_{\mathcal{H}}^2 k(x,x), \quad \text{for any } x \in \mathcal{X}. \tag{17}$$

Since $\sup_{x \in \mathcal{X}} k(x,x) < \infty$, the convergence of f_t to f in $\mathcal{H}$ norm implies the convergence in $L^2(\mathbb{P}_N)$ norm.

6.1 Convergence of prediction error

The following proposition bounds the expected prediction error in terms of the eigenvalues of $K_{\mathbf{x}}P$, denoted as $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_N \geq 0$ as per Lemma 5.

Proposition 7 *Suppose that $\gamma < 1$. For any $f \in \mathcal{H}$, it holds that for all $t \geq 1$*

$$\mathbb{E}_\xi \left[\|f_t - f\|_N^2 \right] \leq 3\kappa \left(\delta_1(t) \|f_0 - f\|_{\mathcal{H}}^2 + \delta_2(t)\sigma^2 + \frac{1}{N} \|\Delta_f\|_2^2 \right), \tag{18}$$

where

$$\delta_1(t) = \frac{1}{s} \max_{1 \leq i \leq N} (1 - \eta\Lambda_i)^{2t} \Lambda_i \leq \frac{1}{2e\eta t s}, \tag{19}$$

$$\delta_2(t) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - (1 - \eta\Lambda_i)^t\right)^2 \leq \frac{1}{N} \sum_{i=1}^{N} \min\{1, \eta t \Lambda_i\}, \tag{20}$$

$$\Delta_f = (f_1^*(\mathbf{x}_1), f_2^*(\mathbf{x}_2), \cdots, f_M^*(\mathbf{x}_M)) - f(\mathbf{x}). \tag{21}$$

The expectation in Proposition 7 is only taken over the observation noise ξ which has zero mean and bounded variance. The above result nicely separates the impact of bias, variance, and heterogeneity on the error dynamics.

- In (18), the first term on the right hand side is of the order $\delta_1(t) \|f_0 - f\|_{\mathcal{H}}^2$ and is related to the bias in estimation. As indicated in (19), $\delta_1(t)$ decreases to 0 as iterations proceed. The upper bound of $\delta_1(t)$ in (19), which decreases at a rate c/t, for a constant c independent of N and the kernel function k. When the kernel matrix $K_{\mathbf{x}}$ is of rank d, the convergence rate can be improved to be $\exp(-c_d t)$ for a constant c_d independent of N.

- The second term on the right hand side of (18) is of the order $\delta_2(t)\sigma^2$ and characterizes the variance in estimation. Note that $\delta_2(t)$ is capped at 1 and is increasing in t. Specifically, it converges to 1 as $t \to \infty$, capturing the phenomenon of over-fitting to noises.

- The third term on the right hand side of (18) is of the order $\|\Delta_f\|_2^2 / N$ and quantifies the impact of the heterogeneity with respect to f. In the presence of only unbalanced data partition and covariate heterogeneity, we have $f_i^* = f^*$ for all i and naturally $\Delta_{f^*} = 0$. Somewhat surprisingly, even under additional model heterogeneity that $f_i^* \neq f_j^*$, with assumptions such as invertibility of $\mathcal{I} - \mathcal{L}$, there exists a choice of f under which $\Delta_f = 0$ (cf. (27)). We caution the reader that while several new FL algorithms including Karimireddy et al. (2020); Mitra et al. (2021); Gorbunov et al. (2021) are shown to converge to the correct stationary point of the original global objective function in the parametric settings, they focus exclusively on the optimization errors and still suffer from prediction errors in the presence of model heterogeneity.

To prevent over-fitting, i.e., to control $\delta_2(t)$, we can terminate the algorithms at some time T before they enter the over-fitting phase. The stopping time T needs to be carefully

chosen to balance the bias and variance (Raskutti et al., 2014). Note that $\delta_1(t) \leq \frac{1}{2e\eta ts}$. To further control $\delta_2(t)$, we need to introduce the empirical Rademacher complexity (Bartlett et al., 2005) defined as

$$\mathcal{R}(\epsilon) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \min\{\lambda_i, \epsilon^2\}}, \tag{22}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are the eigenvalues of kernel matrix $K_{\mathbf{x}}$ as per Lemma 5. Intuitively, $\mathcal{R}(\epsilon)$ is a data-dependent complexity measure of the underlying RKHS and decreases with faster eigenvalue decay and smoother kernels. Recall from (15) that $\Lambda_i \leq \lambda_i s$. Hence it follows from (20) that

$$\delta_2(t) \leq \eta ts\mathcal{R}^2\left(1/\sqrt{\eta ts}\right).$$

Therefore, we can set T as follows:

$$T \triangleq \max\left\{t \in \mathbb{N} : \mathcal{R}\left(1/\sqrt{\eta ts}\right) \leq \frac{1}{\sqrt{2e}\sigma\eta ts}\right\}. \tag{23}$$

That is, we choose T to be the largest time index t so that roughly the bias $\frac{1}{\eta ts}$ dominates the variance $\eta ts\mathcal{R}^2\left(1/\sqrt{\eta ts}\right)\sigma^2$.

With early-stopping, we can specialize the general convergence in Proposition 7 as follows.

Theorem 8 (With early-stopping) *Suppose that $\gamma < 1$. For any $f \in \mathcal{H}$, it holds that for all $1 \leq t \leq T$,*

$$\mathbb{E}_\xi\left[\|f_t - f\|_N^2\right] \leq \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\|\Delta_f\|_2^2.$$

Our result in Theorem 8 shows that the average prediction error decays at a rate of $O(1/t)$ and eventually saturates at the heterogeneity term $\frac{3\kappa}{N}\|\Delta_f\|_2^2$. This encompasses as a special case the existing convergence result of the centralized gradient descent for non-parametric regression (Raskutti et al., 2014) wherein similar early stopping is adopted with the specification $s = 1$ and $f_i^* = f^*$ for all $i \in [M]$. For instance, if the eigenvalues exhibit a polynomial decay with exponent β, that is, $\lambda_i \lesssim i^{-2\beta}$ for $\beta > 1/2$, then direct computation yields that $\eta Ts \asymp (\sigma^2/N)^{-2\beta/(2\beta+1)}$ (see Lemma 26 and Corollary 24). Therefore, Theorem 8 implies that the prediction error converges at a rate of $(\sigma^2/N)^{2\beta/(2\beta+1)}$. Theorem 8 also reassures the common folklore and confirms our empirical observation in Fig. 1b on FedAvg. Specifically, with multiple local steps s up to a certain threshold, the convergence rate increases proportionally to s while the final convergence error stays almost the same, i.e., we can recoup the accuracy loss while enjoying the saving of the communication cost. We cannot set s to be arbitrarily large because as s gets larger, the prediction error increases by a factor of κ, which is an increasing function of s.

Remark 9 (Convergence in $L^2(\mathbb{P})$ norm) *We can also establish a uniform bound to the RKHS norm of $f_t - f$ up to the early stopping time T, as stated in Lemma 27 in the appendix. Furthermore, when the covariates $x_{ij} \overset{i.i.d.}{\sim} \mathbb{P}$, we can apply the empirical process theory to extend the bounds of (16) to those of the prediction error evaluated at the unseen data, i.e., $\mathbb{E}_{x\sim\mathbb{P}}\left[(f_t(x) - f(x))^2\right]$ (see e.g. Raskutti et al. (2014) and (Wainwright, 2019, Chapter 14)). For many kernels including polynomials and Sobolev classes, this yields the centralized minimax-optimal estimation error rate (Yang and Barron, 1999; Raskutti et al., 2012). See Appendix B.3 for the details. We emphasize that the covariates x_{ij}'s are assumed to be independent and identically distributed according to $\mathbb{P}$. This is needed to apply the empirical process theory to extend the convergence results from $L^2(\mathbb{P}_N)$ to $L^2(\mathbb{P})$. It is an interesting yet challenging open problem to extend the results to non-iid covariates.*

Theorem 8 bounds the prediction error in expectation. In practice, the distributional structures of ξ vary across different applications. High-probability bounds on $\|f_t - f\|_N^2$ can be obtained accordingly.

Theorem 10 (High-probability bounds) *Suppose that $\gamma < 1$. For any $f \in \mathcal{H}$ and any $t < T$, let*

$$\varepsilon_t = \mathbb{P}\left\{\|f_t - f\|_N^2 \geq \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 3\right) + \frac{3\kappa}{N}\|\Delta_f\|_2^2\right\}. \tag{24}$$

- *(Sub-Gaussian noise): Suppose the coordinates of the noise vector ξ are N independent zero-mean and sub-Gaussian variables (with sub-Gaussian norm bounded by σ). There exists a universal constant $c > 0$ such that*

$$\varepsilon_t \leq \exp\left(-cN/(\sigma^2\eta ts)\right).$$

- *(Heavy-tailed noise): Suppose the coordinates of ξ are N independent random variables with $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[\xi_i^2] \leq \sigma^2$, and $\mathbb{E}|\xi_i|^p \leq M_p < \infty$ for $p \geq 4$. There exists a constant c_p that only depends on p such that*

$$\varepsilon_t \leq c_p M_p \left(\frac{\eta ts}{N\sigma^2}\right)^{p/4}.$$

Theorem 10 shows that the failure probability decays to 0 as the sample size N tends to infinity; the decay rate is exponential for sub-Gaussian noise and polynomial for noise with bounded moment.

In general, we cannot hope to get a convergence rate that is strictly better than $O(1/t)$. This is because the minimum eigenvalue λ_N of the kernel matrix is *not* bounded away from 0 and may converge to 0 as N diverges. Fortunately, when the kernel matrix $K_{\mathbf{x}}$ has a finite rank d, the convergence rate can be improved to be exponential.

Theorem 11 (Exponential convergence for finite-rank kernel matrix) *Suppose that $\gamma < 1$ and the kernel matrix $K_{\mathbf{x}}$ has finite rank d. Then*

$$\mathbb{E}_\xi\left[\|f_t - f\|_N^2\right] \leq 3\frac{\kappa}{\eta s}\|f_0 - f\|_{\mathcal{H}}^2 \exp\left(-2\frac{\eta s}{\kappa}\lambda_d t\right) + 3\kappa\sigma^2\frac{d}{N} + \frac{3\kappa}{N}\|\Delta_f\|_2^2, \quad \forall t.$$

Enabled by the finite-rankness of $K_{\mathbf{x}}$, Theorem 11 can be deduced from Proposition 7 via deriving a tighter upper bound on $\delta_1(t)$. Finite-rankness also ensures that the variance term is upper bounded by $\sigma^2 d/N$ – hence no early stopping is needed. The complete proof is deferred to Section B.2.

6.2 Convergence of model coefficients

In this section, we show the convergence of model coefficient θ_t, or equivalently, the convergence of f_t in RKHS norm. As shown in (17), this notion of convergence is strictly stronger than the convergence of f_t in $L^2(\mathbb{P}_N)$ norm. For tractability, we assume that the RKHS is d-dimensional, or equivalently $\phi(x)$ is d-dimensional[4]. This encompasses the popular random feature model which maps the input data to a randomized feature space (Rahimi and Recht, 2007).

Theorem 12 *Suppose that $\gamma < 1$ and $\phi(x)$ is d-dimensional. Then*

$$\mathbb{E}_\xi \left[\left\| \theta_t - \bar{\theta} \right\|_2^2 \right] \leq \left(1 - \frac{s\eta\rho_N}{\kappa}\right)^{2t} \left\| \theta_0 - \bar{\theta} \right\|_2^2 + \sigma^2 \frac{\kappa d}{N\rho_N}, \tag{25}$$

where $\bar{\theta}$ is the model coefficient of $\bar{f} = (\mathcal{I} - \mathcal{L})^{-1} \left((f_1^*(\mathbf{x}_1), \ldots, f_M^*(\mathbf{x}_M)) \cdot \Psi \right)$, and $\rho_N = \frac{\lambda_{\min}(\phi(\mathbf{x})^\top \phi(\mathbf{x}))}{N}$. Moreover, the distance between $\bar{\theta}$ and θ_j^* is upper bounded by

$$\left\| \bar{\theta} - \theta_j^* \right\|_2 \leq \left\| \Delta_{f_j^*} \right\|_2 \sqrt{\frac{\kappa}{N\rho_N}}. \tag{26}$$

High probability bounds, similar to Theorem 10 but for θ_t, can be obtained. A few explanations of Theorem 12 are given as below.

- In view of Lemma 2 and the definition of $\bar{f}$, it holds that

$$\bar{f}(\mathbf{x}) \cdot \Psi = (\mathcal{I} - \mathcal{L})\bar{f} = (f_1^*(\mathbf{x}_1), \ldots, f_M^*(\mathbf{x}_M)) \cdot \Psi, \tag{27}$$

  and hence $\Delta_{\bar{f}} \cdot \Psi = 0$. This turns out to be sufficient to ensure that the global model $\bar{f}$ balances out the impact of covariate and model heterogeneity across all clients.

- From (9), we expect that f_t converges to the limiting point $f_\infty = (\mathcal{I} - \mathcal{L})^{-1}(y \cdot \Psi)$. While f_∞ can be far from being the stationary points of the global objective function $\ell(f)$, it is always an unbiased estimator of $\bar{f}$. Building upon this insight, we further bound the variance of the estimator and obtain Theorem 12. See Section 7.1 for further explanations in the special case of linear regression.

- Note that ρ_N depends on N. When N is sufficiently large, which is often the case as N is the **total** number of data points collectively kept by all the M clients, ρ_N is lower bounded by some positive constant[5]. An example can be found in the analysis of Corollary 13, where it is shown that $\rho_N > \frac{\alpha}{2}$ for a fixed constant $\alpha > 0$ and all sufficiently large N.

4. This further implies that $K_{\mathbf{x}}$ is of rank at most d.
5. Note that $\phi(\mathbf{x})^\top \phi(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is the covariance matrix, which is different from the kernel matrix $\phi(\mathbf{x})\phi(\mathbf{x})^\top \in \mathbb{R}^{N \times N}$ whose minimum eigenvalue is 0 when $N > d$.

Theorem 12 casts two key messages, highlighted in italic font below.

Statistical optimality: When ρ_N is lower bounded by a constant independent of N, as $t \to \infty$, the estimation error in (25) converges to $O(d/N)$, which coincides with the minimax-optimal rate for estimating an d-dimensional vector in the centralized setting. Thus, our results immediately imply that when $f_i^* = f_j^*$, even in the presence of covariate heterogeneity, FedAvg and FedProx can achieve statistical optimality by effectively fusing the multi-modal data collected by the clients.

Benefits of Federated Learning: The impact of the model heterogeneity is quantified in (26), which says that $\bar{\theta}$ will stay within a bounded distance to its true local model θ_j^*. In particular, when $n_j \ll d$, though client j cannot learn any meaningful model based on its local dataset, by joining FL it can learn a model which is a reasonable estimation of θ_j^* despite heterogeneity. We formally quantify the benefits of joining FL in depth in Section 6.3.

Depending on the underlying statistical structures of $\phi(\mathbf{x})$, ρ_N and $\|\Delta_{f_j^*}\|_2$ can be further quantified. To cast insights on the magnitudes on ρ_N and $\|\Delta_{f_j^*}\|_2$, next we will present some results on a couple of specific settings.

6.2.1 Covariate heterogeneity with bounded second-moments

Corollary 13 *Suppose that $\gamma < 1$ and $\phi(\mathbf{x})$ is a $N \times d$ matrix whose rows are independent sub-Gaussian with the second-moment matrix $\Sigma_{ij} = \mathbb{E}[\phi(x_{ij})\phi(x_{ij})^\top]$. Assume that $\alpha I \preceq \Sigma_{ij} \preceq \beta I$ for some fixed constants $\alpha, \beta > 0$. There exist constants c_1, c_2 that only depend on α, β such that if $N \geq c_1 d$, then with probability at least $1 - e^{-d}$,*

$$\mathbb{E}_\xi \left[\left\| \theta_t - \bar{\theta} \right\|_2^2 \right] \leq \left(1 - \frac{s\eta}{2\kappa}\right)^{2t} \left\| \theta_0 - \bar{\theta} \right\|_2^2 + \sigma^2 \frac{2\kappa d}{N\alpha}. \tag{28}$$

Moreover, with probability at least $1 - e^{-N}$,

$$\left\| \bar{\theta} - \theta_j^* \right\|_2 \leq c_2 \Gamma \sqrt{\kappa}, \tag{29}$$

where $\Gamma = \max_{i,j} \|f_i^* - f_j^*\|_{\mathcal{H}} = \max_{i,j} \|\theta_i^* - \theta_j^*\|_2$.

Corollary 13 follows from Theorem 12, by showing that with high probability over the randomness of the covariate $\phi(\mathbf{x})$, the matrix $\phi(\mathbf{x})^\top \phi(\mathbf{x})$ is positive definite with $\rho_N \geq \alpha/2$ and moreover $\|\Delta_{f_j^*}\|_2 \lesssim \Gamma\sqrt{N}$.

6.2.2 Covariate heterogeneity with distinct and singular covariance matrices

In this section, we consider distinct covariance matrices, and relax the requirement on the positive-definiteness of $\phi(\mathbf{x})^\top \phi(\mathbf{x})$. In particular, we consider the interesting setting wherein the rows of $\phi(\mathbf{x})$ are drawn from possibly different subspaces of low dimensions. This instance captures a wide range of popular FL applications such as image classification wherein different clients collect different collections of images (McMahan et al., 2017) – some clients may only have images related to airplanes or automobiles while others have images related to cats or dogs.

Suppose the local features on client i lie in a subspace of dimension r_i. Let $\{u_{i1}, \ldots, u_{ir_i}\}$ denote an orthonormal basis of that subspace. The local features $\phi(\mathbf{x}_i)$ can be decomposed as $\phi(\mathbf{x}_i) = \sqrt{d/r_i} F_i U_i^\top$, where $U_i = [u_{i1}, \ldots, u_{ir_i}] \in \mathbb{R}^{d \times r_i}$, and $F_i \in \mathbb{R}^{n_i \times r_i}$ consists of the normalized coefficients. The scaling $\sqrt{d/r_i}$ serves as the normalization factor of the signal-to-noise ratio due to $\|U_i\|_{\mathrm{F}} = r_i$. Furthermore, suppose that the local subspace U_i's are independent with $\mathbb{E}\left[U_i U_i^\top\right] = \frac{r_i}{d} I_d$; for instance, the subspace is uniformly generated at random. Despite the singularity of $\phi(\mathbf{x}_i)$, we show that the statistical accuracy only depends on the conditioning within the local subspace, i.e., the conditioning of F_i.

Corollary 14 *Suppose that $\gamma < 1$, $\lambda_{\min}(F_i^\top F_i/n_i) \geq \alpha$, and $\|F_i^\top F_i/n_i\|_2 \leq \beta$ for $i = 1, \ldots, M$ for some $\alpha, \beta > 0$. There exist a universal constants C such that if $N \geq C\nu d \log d$, where $\nu \triangleq \max_{i \in [M]} n_i/r_i$, then with probability at least $1 - 1/d$:*

$$\mathbb{E}_\xi \left[\left\| \theta_t - \bar{\theta} \right\|_2^2 \right] \leq \left(1 - \frac{s\eta\alpha}{2\kappa}\right)^{2t} \left\| \theta_0 - \bar{\theta} \right\|_2^2 + \sigma^2 \frac{2\kappa d}{N\alpha}, \tag{30}$$

and

$$\left\| \bar{\theta} - \theta_j^* \right\|_2 \leq \Gamma \sqrt{\frac{2\kappa\beta\nu M d}{\alpha N}}. \tag{31}$$

The requirement on $\lambda_{\min}(F_i^\top F_i/n_i)$ is imposed to ensure that the local data at client i contains strong enough signal about θ_i^* on every dimension of the subspace given by U_i. To appreciate the intuition behind this requirement, it is instructive to consider the following two examples:

Example 1 (Orthogonal local dataset) *Suppose that the rows of $\phi(\mathbf{x}_i)$ are orthogonal to each other and each of which has Euclidean norm $\sqrt{d}$. In this case, we have $r_i = n_i$ and $F_i = \sqrt{r_i} I_{r_i}$. Therefore, $\alpha = \beta = \nu = 1$. Then Corollary 14 implies that as long as $N \geq Cd \log d$, θ_t converges exponentially fast to $\bar{\theta}$ up to the optimal mean-squared error rate d/N.*

Example 2 (Gaussian local dataset) *Suppose $\phi(\mathbf{x}_i) = \sqrt{d/r_i} F_i U_i^\top$, where the rows of F_i are i.i.d. $\mathcal{N}(0, I_{r_i})$. In this case, by Gaussian concentration inequality (Vershynin, 2010, Theorem 5.39), with high probability $1 - \delta \leq \alpha \leq \beta \leq 1 + \delta$ for some small constant $\delta > 0$, provided that $n_i \geq C \max\{r_i, \log M\}$ for some sufficiently large constant C. Then Corollary 14 implies that if further $N \geq C\nu d \log d$, θ_t converges exponentially fast to $\bar{\theta}$ up to the optimal mean-squared error rate d/N.*

Note that we pay an extra factor of ν in the sample complexity in Corollary 14. This is necessary in general. To see this, consider the extreme case where $r_i = 1$ and $n_i = n$, i.e., all local data at client i lie on a straight line in $\mathbb{R}^d$. Then by the standard coupon collector's problem, we need $M \geq d \log d$ in order to sample all the d basis vectors in $\mathbb{R}^d$.

6.3 Characterization of federation gains

As mentioned in Section 6.2, when $n_j \ll d$, though client j cannot learn any meaningful model based on its local dataset, by joining FL it can learn a model which is a reasonable

estimation of θ_j^* despite heterogeneity. In this section, we formally characterize, compared with training based on local data only, the gains/loss of a client in joining FL, referred to as *federation gain* henceforth.

Let $\widehat{f}_j \equiv \widehat{f}_j(\mathbf{x}_j, y_j)$ denote any estimator of the true model f_j^* based on the local data $(\mathbf{x}_j, y_j)$ at client j. Let

$$R_j^{\mathsf{Loc}} = \inf_{\widehat{f}_j} \sup_{f_j^* \in \mathcal{H}_B} \mathbb{E}_{\mathbf{x}_j, \xi_j} \left[\left\| \widehat{f}_j - f_j^* \right\|_{\mathcal{H}}^2 \right] \tag{32}$$

denote the minimax risk attainable by the best local estimator $\widehat{f}_j$, where $\mathcal{H}_B = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$[6]. Recall that f_t is the model trained under FL after t rounds. Consequently, f_t can be viewed as a function of the datasets of all the M clients, i.e., $f_t \equiv f_t(\mathbf{x}, y)$. Define the risk of the federated model in estimating f_j^* as

$$R_j^{\mathsf{Fed}} = \inf_{t \geq 0} \sup_{f_j^* \in \mathcal{H}_B} \mathbb{E}_{\mathbf{x}, \xi} \left[\left\| f_t - f_j^* \right\|_{\mathcal{H}}^2 \right], \tag{33}$$

where we take the infimum over time t due to the possible use of the early stopping rule. Notably, to average out the randomness induced by the training data, in (32) and (33) the expectations are taken over local training data $(\mathbf{x}_j, \xi_j)$ and the global training data $(\mathbf{x}, \xi)$, respectively.

Definition 15 (Federation gain) *The* federation gain *of client j in participating FL is defined as the ratio of the local minimax risk and the federated risk:*

$$\mathsf{FG}_j \triangleq \frac{R_j^{\mathsf{Loc}}}{R_j^{\mathsf{Fed}}}.$$

Intuitively, the federation gain is the multiplicative reduction of the error of estimating f_j^* in joining FL compared to the best local estimators. Next we give explicit forms of federation gains for the heterogeneity discussed in details in Sections 6.2.1 and 6.2.2.

Theorem 16 *Consider the same setup as Corollary 13 and assume that $\xi_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Then for $1 \leq j \leq M$, there exists a constant c_1 that only depends on constants α, β such that*

$$\mathsf{FG}_j \geq \frac{c_1}{\kappa} \frac{\min\{\sigma^2 d/n_j, B^2\} + \max\{1 - n_j/d, 0\} B^2}{\sigma^2 d/N + \Gamma^2} \tag{34}$$

Theorem 16 reveals interesting properties of the federation gain. On the extreme case where $\Gamma = 0$ (i.e., there is no model heterogeneity) the federation gain achieves its maximum, which is at least on the order of $\min\{N/n_j, N/d\}$. As the model heterogeneity Γ increases, the federation gain decreases. In particular, for data-scarce clients with local data volume $n_j < d$, the federation gain is at least on the order of $(1 - n_j/d)B^2/\Gamma^2$, which exceeds one when $\Gamma \leq B\sqrt{1 - n_j/d}$. For data-rich clients with local data volume $n_j \geq d$, the

6. Here we impose an upper bound B to the RKHS norm of f_j^* to prevent the minimax risk from blowing up to the infinity when the local data size $n_j < d$ Mourtada (2019).

federation gain is at least on the order of $\min\{\sigma^2 d/n_j, B^2\}/\Gamma^2$, which exceeds one when $\Gamma \le \min\{\sigma\sqrt{d/n_j}, B\}$.

The following theorem further characterizes the federation gain under the subspace model in the presence of covariate heterogeneity.

Theorem 17 *Consider the same setup as Corollary 14. Further, suppose that α, β are fixed positive constants, $f_i^* = f^*$ for all $i \in [M]$, $N \ge C\nu d \log d$, and $\xi_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Then for all $1 \le j \le M$, there exists a constant $c_1 > 0$ depending on α, β such that*

$$\mathsf{FG}_j \ge \frac{c_1}{\kappa} \frac{\min\{\sigma^2 d/n_j, B^2\} + (1 - r_j/d)B^2}{\sigma^2 d/N}. \tag{35}$$

Theorem 17 implies that when $N \gtrsim \nu d \log d$: for data-scarce clients with local data volume $n_j \ll d$, FG_j is dominated by N/d which is unchanged with r_j; for data-rich clients with local data volume $n_j \gg d$, FG_j is dominated by $\frac{1-r_j/d}{d/N}$, which is decreasing in r_j. On the contrary, if $N \ll \nu d \log d$, f_t is not expected to estimate f^* due to the aforementioned coupon collector's problem and hence the federation gain will be small. In conclusion, the federation gain will exhibit a sharp jump at a critical sample complexity $N = \Theta(\nu d \log d)$. This is confirmed by our numerical experiment in Section 7.3.

7. Experimental Results

In this section, we provide experimental results corroborating our theoretical findings.

7.1 Stationary points and estimation errors

We numerically verify that despite the failure of converging to the stationary points of the global emprical risk function, both FedAvg and FedProx can achieve low estimation errors.

We adopt the same simulation setup of Pathak and Wainwright (2020) for fairness in comparison. We let $M = 25$, $d = 100$, and $n_i = 500$. For each client i, suppose $f_i^* = X_i\theta^*$ for some $\theta^* \in \mathbb{R}^d$ and the response vector $y_i \in \mathbb{R}^{n_i}$ is given by $y_i = X_i\theta^* + \xi_i$, where $\xi_i \in \mathbb{R}^{n_i}$ is distributed as $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.5$. The local design matrices X_i are independent random matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Let $\ell(\theta) = \frac{1}{N}\sum_{i=1}^M \|y_i - X_i\theta\|_2^2$ be the global empirical risk function. The difference to Pathak and Wainwright (2020) is that, instead of plotting the sub-optimality in the excess risk $\ell(\theta_t) - \min_\theta \ell(\theta)$, we plot the trajectories of $\|\nabla \ell(\theta_t)\|_2$ to highlight the unreachability to stationary points of $\ell(\theta_t)$.

For both FedAvg and FedProx, we choose the step size $\eta = 0.1$. Fig. 1a confirms the observation in Pathak and Wainwright (2020) that FedAvg with ($s \ge 2$) and FedProx fail to converge to the stationary point of global empirical risk function $\ell(\theta)$. However, Fig. 1b shows that both FedAvg with $s = 5, 10$ and FedProx can achieve almost the same low error $\|\theta_t - \theta^*\|_2$ as FedAvg with $s = 1$, i.e., the standard centralized gradient descent method.

To see the underlying intuition, note that the limiting point obtained by FedAvg is explicitly derived by Pathak and Wainwright (2020):

$$\widehat{\theta}_{\text{Fed}} = \left(\frac{1}{N}\sum_{i=1}^M X_i^\top X_i \sum_{\ell=0}^{s-1}(I - \eta_i X_i^\top X_i)^\ell\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^M \sum_{\ell=0}^{s-1}(I - \eta_i X_i^\top X_i)^\ell X_i^\top y_i\right),$$

where $\eta_i = \eta/n_i$, while the global minimizer, which is the desired stationary point, is given by the ordinary least squares (equivalently, $\widehat{\theta}_{\text{Fed}}$ with $s = 1$)

$$\widehat{\theta}_{\text{OLS}} = \left(\frac{1}{N}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{M} X_i^\top y_i\right).$$

Thus when $s > 1$, $\widehat{\theta}_{\text{Fed}}$ can be far from $\widehat{\theta}_{\text{OLS}}$. Nevertheless, by plugging $y_i = X_i\theta^* + \xi_i$, we have

$$\widehat{\theta}_{\text{Fed}} = \theta^* + \left(\frac{1}{N}\sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1}(I - \eta_i X_i^\top X_i)^\ell\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{M}\sum_{\ell=0}^{s-1}(I - \eta_i X_i^\top X_i)^\ell X_i^\top \xi_i\right),$$
$$\widehat{\theta}_{\text{OLS}} = \theta^* + \left(\frac{1}{N}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{M} X_i^\top \xi_i\right).$$

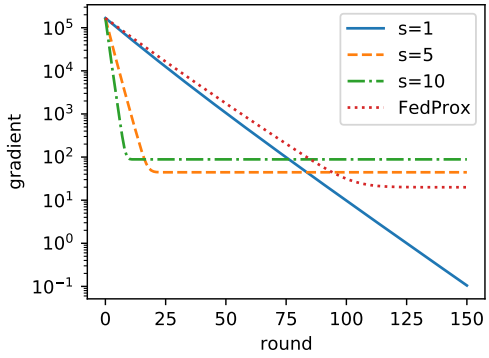
Therefore, both $\widehat{\theta}_{\text{Fed}}$ and $\widehat{\theta}_{\text{OLS}}$ are unbiased estimator of the underlying model parameter θ^* with different variances. Our results in Theorem 12 provide upper bounds for the variance of $\widehat{\theta}_{\text{Fed}}$.

Impact of minibatches Minibatches are often adopted in the real-world implementations of FedAvg and FedProx. Specifically, each client i first partitions its local data into batches of the chosen size B_i. Then for each of the s local steps in FedAvg McMahan et al. (2017), the client i updates $\theta_{i,t}$ via running gradient descent $\frac{n_i}{B_i}$ times, where a different batch in the data partition is used each time. In this way, each of the batches is passed s times in one round. Similarly, in FedProx Li et al. (2020), the client i updates $\theta_{i,t}$ by solving the local proximal optimization (6) $\frac{n_i}{B_i}$ times, where a different batch in the data partition is used each time. In our analysis and previous numerical experiments, we assumed full batch $B_i = n_i$. A natural but interesting question is whether FedAvg and FedProx still enjoys the statistical optimality when using minibatches where $B_i < n_i$.

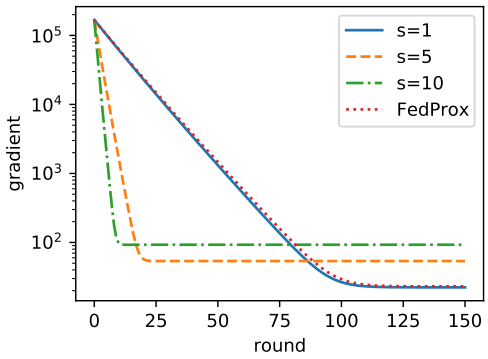
To answer this question, we re-run the experiments with the same setup as above but with three batch sizes B: 20, 50, and 100. We plot the gradient magnitudes and estimation errors in Fig. 2 and Fig. 3. For ease of comparison, we redraw Fig. 1a and Fig. 1b in Fig. 2a and Fig. 3a.

As illustrated in Fig. 2, for $s = 5$ and $s = 10$ the impacts of different batch sizes on the gradient magnitude are negligible. However, strikingly, for FedAvg $s = 1$ with minibatch, its gradient magnitude rises up significantly and hence it can no longer reach the stationary point (This can be rigorously proved by following the arguments in Pathak and Wainwright (2020).). For FedProx with minibatch, its curve mostly coincides with that of FedAvg $s = 1$. In contrast, as shown in Fig. 3, the minibatch has almost no effect on the estimation error. The final estimation errors are almost identical in each of the four figures in Fig. 3. The convergence speed of FedAvg $s = 1$ only decreases a bit with minibatch.

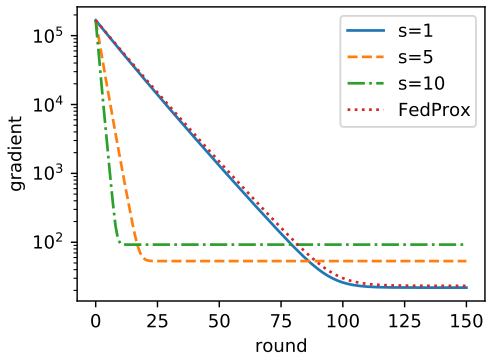
In conclusion, we see that both FedAvg and FedProx with minibatch can achieve low estimation errors despite the unreachability of the stationary points.



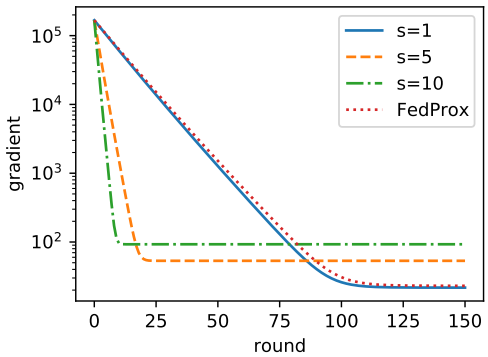
(a) Full $\mathcal{S}_i$ in each local gradient descent update



(b) Batch size 20



(c) Batch size 50



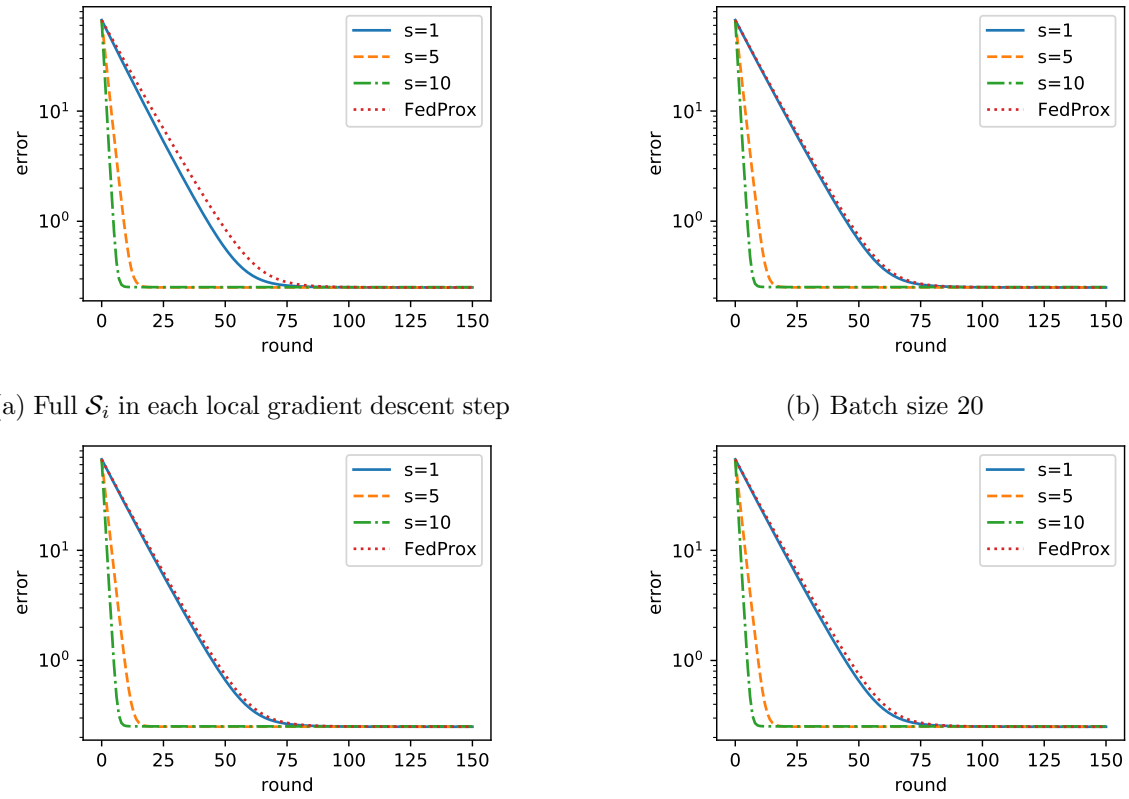
(d) Batch size 100

Figure 2: Impacts of mini batch sizes on the reachability of stationary points

7.2 Federation gains versus model heterogeneity

As mentioned in Section 3, data heterogeneity includes both model heterogeneity (a.k.a. concept shift) and covariate heterogeneity (a.k.a. covariate shift). Complementing our Theorem 16, we provide a numerical study on the impact of model heterogeneity on the federation gain in this section, and the corresponding results of covariate heterogeneity in the next section. We build on our previous experiment setup by allowing for unbalanced local data and the heterogeneity in f_i^*. We choose $M = 20$, $d = 100$, $n_i = 50$ for half of the clients, and $n_i = 500$ for the remaining clients. We refer to the clients with $n_i = 50$ as *data scarce* clients, and to the others as *data rich* clients. We run the experiments with a prescribed set of heterogeneity levels. All the other specifications are the same as before.

We randomly choose a data scarce client and a data rich client, and plot the federation gains against the model heterogeneity $\Gamma = \max_{i,j\in[M]} \|\theta_i^* - \theta_j^*\|_2$ in Fig. 4. Note that in evaluating the federation gains, we use the minimum-norm least squares as the benchmark local estimator, that is $\widehat{\theta}_j = (X_j^\top X_j)^+ X_j^\top y_j$, where the symbol $+$ denotes the Moore-Penrose pseudoinverse. It is known that this estimator can attain the minimax-optimal estimation error rate (Mourtada, 2019). We see that consistent with our theory, despite the



(a) Full $\mathcal{S}_i$ in each local gradient descent step

(b) Batch size 20

(c) Batch size 50

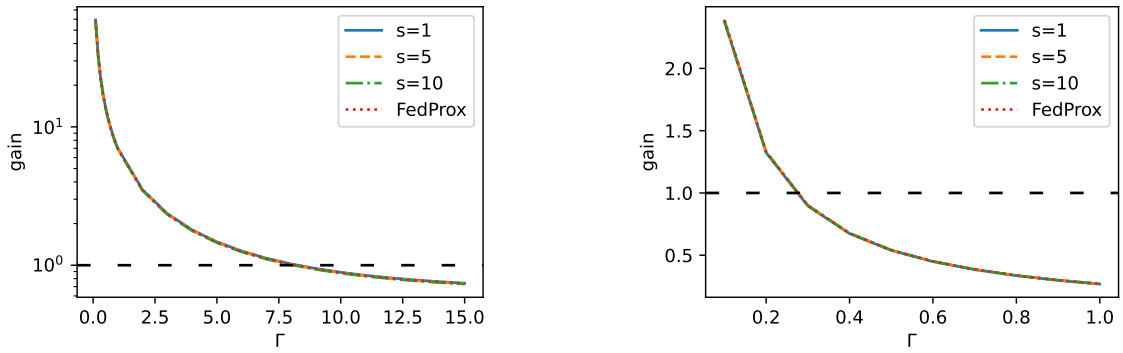
(d) Batch size 100

Figure 3: Impacts of mini batch sizes on the estimation errors

difference in the training behaviors, the models trained under FedAvg with different choices of aggregation periods s and under FedProx have almost indistinguishable federation gains. Moreover, as predicted by our theory, the federation gain drops with increasing model heterogeneity Γ, while the federation gain of the data scarce client is much higher than that of the data rich client. Recall that the federation gain exceeds 1 if and only if the FL model is better than the locally trained model. We observe that the federation gain of a data scarce client drops below 1 at $\Gamma \approx 7.5$, whereas the federation gain of a data rich client drops below 1 at $\Gamma \approx 0.3$. These numbers turn out to be closely match with our theoretically predicted thresholds given after Theorem 16, which are $\Gamma \approx \sqrt{1 - n_j/d}\,\|\theta_j^*\|_2 \approx 7$ and $\Gamma \approx \sigma\sqrt{d/n_j} \approx 0.22$, respectively.

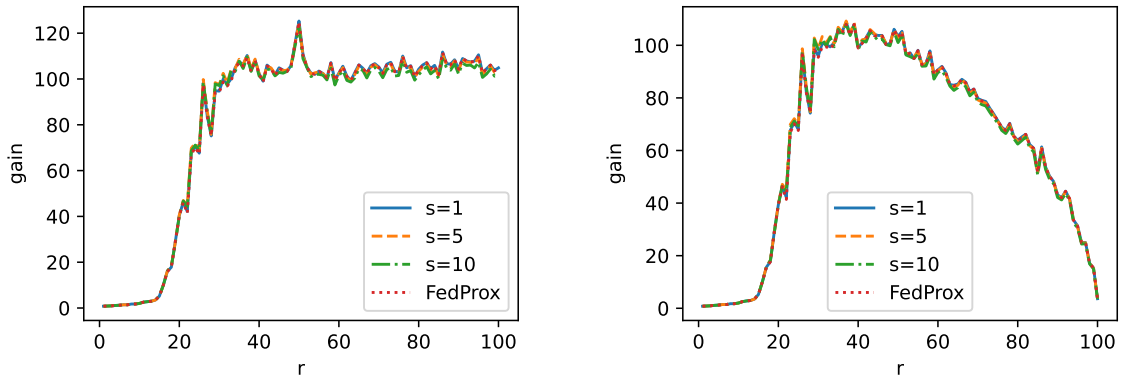
7.3 Federation gain versus covariate heterogeneity

In this section, we study the impact of covariate heterogeneity on the federation gains by focusing on the subspace model. In our experiments, we choose $M = 20$, $d = 100$, $\sigma = 0.5$, $n_i = 50$ for half of the clients, and $n_i = 500$ for the remaining clients. We let the 20 clients share a common underlying truth, i.e., $\theta_j^* = \theta^*$ for all j, which is randomly drawn from $\mathcal{N}(0, I)$. The responses y_i are given as $y_i = X_i\theta^* + \xi_i$. The design matrices $X_i \in \mathbb{R}^{n_i \times d}$ at



(a) A data scarce client (b) A data rich client

Figure 4: Federation gains versus Γ. A data scarce client benefits more from FL participation.



(a) A data scarce client (b) A data rich client

Figure 5: Federation gains versus subspace dimension r.

the clients lie in different subspaces of dimension r; here r ranges from 1 to 100. Specifically, X_i's are generated as follows: we first generate a random index set $E \subseteq [d]$ of cardinality r, and generate a matrix X_i with each row independently distributed as $\mathcal{N}(0, (d/r)I_E)$, where I_E is a diagonal matrix with $(I_E)_{ii} = \mathbf{1}_{\{i \in E\}}$. The scaling $\frac{d}{r}$ ensures that each row of X_i has ℓ^2 norm $\sqrt{d}$ in expectation and hence the signal-to-noise ratio is consistent across different values of r. As $\|X_i\|_2$ increases by a factor of $\sqrt{d/r}$, we rescale the stepsize by choosing $\eta = 0.1/(d/r)$ for the stability of local iterations, according to Corollary 14. Notably, when $r = d$, the stepsize becomes $\eta = 0.1$ which is the same as previous experiments. We randomly choose a data scarce client and a data rich client and record the federation gains. We plot the average federation gains over 20 trials against r – dimension of the subspaces – in Fig. 5. We have the following key observations, matching our theoretical predictions given in Theorem 17:

- First, for any fixed r, a client's federation gains of the model trained by FedAvg with $s = 1, 5, 10$ and FedProx are almost identical. This is consistent with our theory, as we show both FedAvg and FedProx converge to the minimax-optimal mean-squared error rate d/N in Corollary 14.
- Second, up to $r \approx 27$, the curves for the data scarce and the data rich clients are roughly the same. This is because when $r \leq 27$, the main "obstacle" in learning θ^* is the lack of sufficient coverage of each of the 100 dimensions by the data collectively kept by the 20 clients.
- Third, for both curves there are significant jumps starting when $r \approx 16$ to when $r \approx 23$. If we can pool the data together, due to the coupon-collecting effect, as soon as $M \times r \geq d \log d \approx 460$, all the d dimensions can be covered by the design matrices and hence the underlying truth θ^* can be learned with high accuracy. Since $M = 20$, this explains the significant jumps in federations gains in Fig. 5 when r is around 23.
- Finally, the curve trends are different for data scarce and data rich clients. For a data scarce client, as shown in Fig. 5a, as r increases, the federation gain first increases and then stabilizes around 107. In contrast, for a data rich client, as shown in Fig. 5b, as r increases, the federation gain first increases and then quickly decreases when r approaches 100. This distinction is because a data scarce client, on its own, cannot learn θ^* well as $n_i = 50 \ll 100$ no matter how large r is, while a data rich has 500 data tuples and can learn θ^* on its own quite well when r approaches 100.

7.4 Fitting nonlinear functions

In this section, we go beyond linear models. In particular, we focus on fitting U_5 – the degree-5 Chebyshev polynomials of the second kind, which is a special case of the Gegenbauer polynomials and has the explicit expression

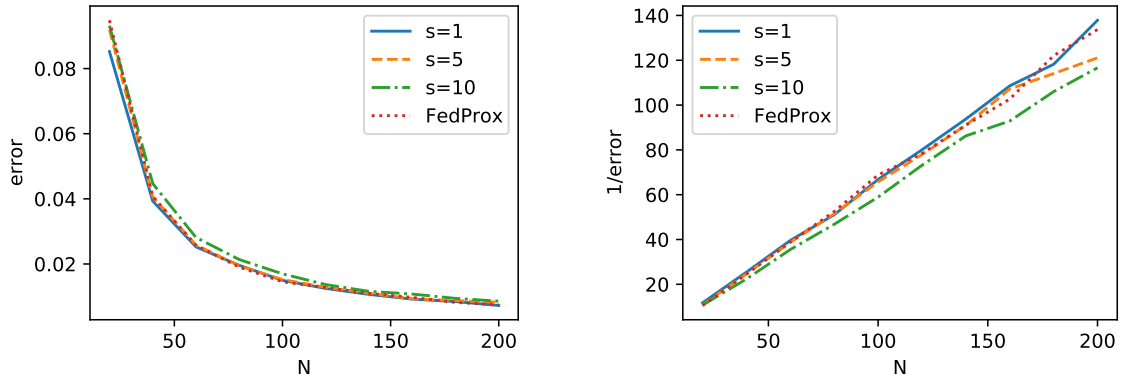
$$U_5(x) = 32x^5 - 32x^3 + 6x.$$

We choose the feature map $\phi(x) = [1, x, \ldots, x^5]^\top$ to be the monomial basis up to degree 5 and run FedAvg and FedProx on the polynomial coefficients. We consider $M = 20$ clients and equal size local dataset $n_i \in \{1, 2, \ldots, 10\}$. Correspondingly, the global dataset size ranges from 20 to 200 as indicated by Fig. 6. The response value is given as $y = U_5(x) + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$. We consider heterogeneous local datasets. Specifically, each client $i \in [M]$ probes the function on disjoint intervals $[-1 + \frac{2(i-1)}{M}, -1 + \frac{2i}{M})$. In the experiments, we generate covariates x_{ij} using the uniform grid. For the fitted function $\widehat{f}$, we evaluate the mean-squared error (MSE) as

$$\left\| \widehat{f} - f^* \right\|_2^2 = \int_{-1}^{1} \left| \widehat{f}(x) - f^*(x) \right|^2 \mathrm{d}x.$$

We run FedAvg and FedProx with the same stepsize $\eta = 0.1$ as before, and evaluate the MSE via Monte Carlo integration. We plot the average MSE over 500 trials.

As shown by Fig. 6a, the four curves of the model prediction errors under FedAvg with different choices of s and FedProx are very similar. Note that for polynomial kernels, the



(a) Estimation errors

(b) Reciprocal of estimation errors

Figure 6: Estimation errors of fitting polynomials

minimax-optimal estimation rate is $O(1/N)$ Raskutti et al. (2012). In comparison, we plot the reciprocal of prediction errors in Fig. 6b. Though the differences in the reciprocal of the prediction errors get amplified as the errors approach zero, each of the four curves in Fig. 6b are mostly straight lines. Moreover, the four curves have similar slopes with the slope of $s = 10$ being slightly smaller than others. This is because a larger s leads to a κ being slightly greater than one and thus an increased error. These observations confirm that both FedAvg and FedProx can achieve nearly-optimal estimation rate in polynomial regression.

Acknowledgments

The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions. J. Xu is supported by the NSF Grants IIS-1838124, CCF-1850743, CCF-1856424, and CCF-2144593. P. Yang is supported by the NSFC Grant 12101353 and Tsinghua University Initiative Scientific Research Program.

References

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2009.

Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 2016.

Zachary Charles and Jakub Konečnỳ. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020.

Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv:1802.07876*, 2018.

Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *The Annals of Probability*, pages 806–816, 1995.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv:2003.13461*, 2020.

Canh Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pages 21394–21405. Curran Associates, Inc., 2020.

Simon Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568. Curran Associates, Inc., 2020.

Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, pages 13–38. Springer, 2000.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv:1909.12488*, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin

Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016a.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv:1610.05492*, 2016b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv:1907.02189*, 2019.

Sen Lin, Guang Yang, and Junshan Zhang. A collaborative learning framework via federated meta-learning. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 289–299. IEEE, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv:1912.10754*, 2019.

Reese Pathak and Martin J Wainwright. Fedsplit: an algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 7057–7066. Curran Associates, Inc., 2020.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(2), 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arxiv:1011.3027*, 2010.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv:1806.00582*, 2018.

Martin Zinkevich, Markus Weimer, Alexander J Smola, and Lihong Li. Parallelized stochastic gradient descent. In *NIPS*, volume 4, page 4. Citeseer, 2010.

Appendix A. Missing Proofs in Section 5

Proof [Proof of Proposition 1] For FedAvg, recall from (7) that $\mathcal{G}_i(f) = \mathcal{L}_i f + \frac{\eta}{n_i} y_i \cdot k_{\mathbf{x}_i}$. Iteratively applying the mapping $\mathcal{G}_i$ s times, we get that

$$f_{i,t} = \mathcal{G}_i^s(f_{t-1}) = \mathcal{L}_i^s f_{t-1} + \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau \frac{\eta}{n_i} y_i \cdot k_{\mathbf{x}_i} = \mathcal{L}_i^s f_{t-1} + y_i \cdot \Psi_i.$$

Combining the last display with (5) yields (9).

For FedProx, it follows from (6) that $f_{i,t}$ solves

$$\frac{\eta}{n_i} \sum_{j=1}^{n_i} (f(x_{ij}) - y_{ij}) k_{x_{ij}} + (f - f_{t-1}) = 0.$$

Then it follows from the definition of the linear operator $\widetilde{\mathcal{L}}_i^{-1}$ that

$$f_{i,t} = \widetilde{\mathcal{L}}_i^{-1}\left(f_{t-1} + \frac{\eta}{n_i}\sum_{j=1}^{n_i} y_{ij}k_{x_{ij}}\right) = \widetilde{\mathcal{L}}_i^{-1}f_{t-1} + y_i \cdot \Psi_i.$$

Combining the last display with (5) yields (9) for FedProx. ■

Proof [Proof of Lemma 2] We first prove the lemma for FedAvg. Note that $f - \mathcal{L}f = \sum_{i=1}^M w_i(f - \mathcal{L}_i^s f)$, and that each term admits the following representation in terms of telescoping sums:

$$f - \mathcal{L}_i^s f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau f - \mathcal{L}_i^{\tau+1} f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau (f - \mathcal{L}_i f) = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau \left(\frac{\eta}{n_i}\sum_{j=1}^{n_i} f(x_{ij})k_{x_{ij}}\right) = f(\mathbf{x}_i)\cdot\Psi_i.$$

Hence applying the definition of Ψ yields that $f(\mathbf{x})\cdot\Psi = \sum_i w_i f(\mathbf{x}_i)\cdot\Psi_i = f - \mathcal{L}f$. Moreover, since

$$\mathcal{L}_i k_{\mathbf{x}_i} = \begin{bmatrix} k_{x_{i1}} - \frac{\eta}{n_i}\sum_{j=1}^{n_i} k(x_{i1}, x_{ij})k_{x_{ij}} \\ \vdots \\ k_{x_{in_i}} - \frac{\eta}{n_i}\sum_{j=1}^{n_i} k(x_{in_i}, x_{ij})k_{x_{ij}} \end{bmatrix} = (I - \eta K_{\mathbf{x}_i})k_{\mathbf{x}_i}, \tag{36}$$

it follows from induction that, for every $\tau \in \mathbb{N}$,

$$\mathcal{L}_i^\tau k_{\mathbf{x}_i} = (I - \eta K_{\mathbf{x}_i})^\tau k_{\mathbf{x}_i}.$$

Therefore, using the definition of P_{ii} yields that $\Psi_i = \frac{\eta}{n_i}P_{ii}k_{\mathbf{x}_i}$ and thus $\Psi = \frac{\eta}{N}Pk_{\mathbf{x}}$. Since $\Psi(\mathbf{x})$ is a N by N matrix that stacks $\Psi(x_{ij})$ in rows, we obtain that $\Psi(\mathbf{x}) = \eta K_{\mathbf{x}}P$. Consequently, since $f(x_{ij}) - \mathcal{L}f(x_{ij}) = \Psi(x_{ij})\cdot f(\mathbf{x})$, we have

$$f(\mathbf{x}) - \mathcal{L}f(\mathbf{x}) = \Psi(\mathbf{x})f(\mathbf{x}) = \eta K_{\mathbf{x}}Pf(\mathbf{x}).$$

Analogously, for FedProx, by the identity

$$f - \widetilde{\mathcal{L}}_i^{-1}f = \widetilde{\mathcal{L}}_i^{-1}(\widetilde{\mathcal{L}}_i f - f) = \widetilde{\mathcal{L}}_i^{-1}\frac{\eta}{n_i}\sum_{j=1}^{n_i} f(x_{ij})k_{x_{ij}} = f(\mathbf{x}_i)\cdot\Psi_i,$$

we get that $f(\mathbf{x})\cdot\Psi = f - \mathcal{L}f$. Similar to (36),

$$\widetilde{\mathcal{L}}_i k_{\mathbf{x}_i} = (I + \eta K_{\mathbf{x}_i})k_{\mathbf{x}_i}.$$

Therefore, $\widetilde{\mathcal{L}}_i^{-1}k_{\mathbf{x}_i} = P_{ii}k_{\mathbf{x}_i}$ and hence $\Psi_i = \frac{\eta}{n_i}P_{ii}k_{\mathbf{x}_i}$. The rest of the proof is identical to that for FedAvg. ■

The following lemma is used in the proof of Lemma 4.

Lemma 18 *Suppose* $\mathbf{x} \in \mathcal{X}^n$ *and* $\Pi \in \mathbb{R}^{m\times n}$. *Then, for any* f *such that* $\|f\|_{\mathcal{H}} \leq 1$, *it holds that*

$$\Pi K_{\mathbf{x}} \Pi^\top \succeq \frac{1}{n} \Pi f(\mathbf{x}) f(\mathbf{x})^\top \Pi^\top. \tag{37}$$

Furthermore, the following is true:

$$\left\| \Pi K_{\mathbf{x}} \Pi^\top \right\|_2 = \max_{\|f\|_{\mathcal{H}} \leq 1} \frac{1}{n} \|\Pi f(\mathbf{x})\|_2^2 . \tag{38}$$

Proof For any $u \in \mathbb{R}^n$, we have

$$\begin{aligned} u^\top \Pi K_{\mathbf{x}} \Pi^\top u &= \frac{1}{n} \left\langle (u^\top \Pi) \cdot k_{\mathbf{x}}, (u^\top \Pi) \cdot k_{\mathbf{x}} \right\rangle_{\mathcal{H}} = \frac{1}{n} \max_{\|f\|_{\mathcal{H}} \leq 1} \left\langle f, (u^\top \Pi) \cdot k_{\mathbf{x}} \right\rangle_{\mathcal{H}}^2 \\ &= \frac{1}{n} \max_{\|f\|_{\mathcal{H}} \leq 1} (u^\top \Pi f(\mathbf{x}))^2 = \frac{1}{n} \max_{\|f\|_{\mathcal{H}} \leq 1} u^\top (\Pi f(\mathbf{x}))(\Pi f(\mathbf{x}))^\top u, \end{aligned} \tag{39}$$

where the second equality used the fact that $\|g\|_{\mathcal{H}} = \max_{\|f\|_{\mathcal{H}} \leq 1} \langle f, g \rangle_{\mathcal{H}}$. Then (37) follows from (39). Note that $\Pi K_{\mathbf{x}} \Pi^\top$ is a symmetric matrix. Then,

$$\left\| \Pi K_{\mathbf{x}} \Pi^\top \right\|_2 = \max_{\|u\|_2 \leq 1} \frac{1}{n} u^\top \Pi K_{\mathbf{x}} \Pi^\top u = \max_{\|f\|_{\mathcal{H}} \leq 1, \|u\|_2 \leq 1} \frac{1}{n} (u^\top \Pi f(\mathbf{x}))^2 = \max_{\|f\|_{\mathcal{H}} \leq 1} \frac{1}{n} \|\Pi f(\mathbf{x})\|_2^2 ,$$

where the second equality follows from (39). ■

Proof [Proof of Lemma 4] FedAvg: For any $f \in \mathcal{H}$ we have

$$\langle f, \mathcal{L}_i f \rangle_{\mathcal{H}} \overset{(a)}{=} \|f\|_{\mathcal{H}}^2 - \frac{\eta}{n_i} \|f(\mathbf{x}_i)\|_2^2 \leq \|f\|_{\mathcal{H}}^2 ,$$

where equality (a) follows from the definition of $\mathcal{L}_i$, proving $\|\mathcal{L}_i\|_{\text{op}} \leq 1$. We show $\mathcal{L}_i$ is positive by lower bounding $\langle f, \mathcal{L}_i f \rangle_{\mathcal{H}}$ as follow:

$$\langle f, \mathcal{L}_i f \rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2 - \frac{\eta}{n_i} \|f(\mathbf{x}_i)\|_2^2 \overset{(a)}{\geq} \|f\|_{\mathcal{H}}^2 (1 - \eta \|K_{\mathbf{x}_i}\|_2) \overset{(b)}{\geq} \|f\|_{\mathcal{H}}^2 (1 - \gamma) \overset{(c)}{>} 0,$$

where inequality (a) follows from Lemma 18, (b) holds by definition of γ, and (c) is true by the assumption that $\gamma < 1$. Hence, we obtain that $\mathcal{L}_i$ is positive with $\|\mathcal{L}_i\|_{\text{op}} \leq 1$, which immediately implies that both $\mathcal{L}_i^s$ and $\mathcal{L} = \sum_i w_i \mathcal{L}_i^s$ are positive and their operator norm is upper bounded by 1.

FedProx: For any $f \in \mathcal{H}$, let $g \triangleq \widetilde{\mathcal{L}}_i f$. Next we show that $0 \leq \langle \widetilde{\mathcal{L}}_i^{-1} f, f \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}^2$. By definition, it holds that $\left\langle f, \widetilde{\mathcal{L}}_i^{-1} f \right\rangle_{\mathcal{H}} = \left\langle \widetilde{\mathcal{L}}_i g, g \right\rangle_{\mathcal{H}}$. We have

$$\left\langle \widetilde{\mathcal{L}}_i g, g \right\rangle_{\mathcal{H}} = \left\langle g + \frac{\eta}{n_i} \sum_{j=1}^{n_i} g(\mathbf{x}_{ij}) k_{\mathbf{x}_{ij}}, g \right\rangle_{\mathcal{H}} = \|g\|_{\mathcal{H}}^2 + \frac{\eta}{n_i} \|g(\mathbf{x}_i)\|_2^2 \geq 0.$$

In addition, we have

$$\|f\|_{\mathcal{H}}^2 = \left\langle \widetilde{\mathcal{L}}_i g, \widetilde{\mathcal{L}}_i g \right\rangle_{\mathcal{H}} = \|g\|_{\mathcal{H}}^2 + 2\frac{\eta}{n_i}\|g(\mathbf{x}_i)\|_2^2 + \left(\frac{\eta}{n_i}\right)^2 \sum_{j=1}^{n_i}\sum_{j'=1}^{n_i} g(x_{ij})g(x_{ij'})k(x_{ij}, x_{ij'})$$
$$\overset{(a)}{\geq} \|g\|_{\mathcal{H}}^2 + 2\frac{\eta}{n_i}\|g(\mathbf{x}_i)\|_2^2 \geq \|g\|_{\mathcal{H}}^2 + \frac{\eta}{n_i}\|g(\mathbf{x}_i)\|_2^2 = \left\langle \widetilde{\mathcal{L}}_i g, g \right\rangle_{\mathcal{H}},$$

where inequality (a) is true because the kernel function k is positive semi-definite. Hence, we obtain that $\widetilde{\mathcal{L}}_i^{-1}$ is positive with $\left\|\widetilde{\mathcal{L}}_i^{-1}\right\|_{\text{op}} \leq 1$ regardless of γ, and so is $\mathcal{L} = \sum_i w_i \widetilde{\mathcal{L}}_i^{-1}$.

Note that P is positive definite when $\gamma < 1$ for FedAvg and is positive definite regardless of γ for FedProx. Positive definiteness ensure that the matrix $K_{\mathbf{x}}P$ is similar [7] to $P^{1/2}K_{\mathbf{x}}P^{1/2}$ which has non-negative eigenvalues only. So it suffices to prove $\|\eta P^{1/2}K_{\mathbf{x}}P^{1/2}\|_2 \leq 1$. By Lemma 18,

$$\left\|\eta P^{1/2}K_{\mathbf{x}}P^{1/2}\right\|_2 = \max_{\|f\|_{\mathcal{H}}\leq 1} \frac{\eta}{N} f(\mathbf{x})^\top P f(\mathbf{x}) = \max_{\|f\|_{\mathcal{H}}\leq 1} \frac{\eta}{N} \left\langle f, (f(\mathbf{x})^\top P)\cdot(k_{\mathbf{x}}) \right\rangle_{\mathcal{H}}.$$

Applying Lemma 2 yields that

$$\frac{\eta}{N}\langle f, f(\mathbf{x})\cdot(Pk_{\mathbf{x}})\rangle_{\mathcal{H}} = \langle f, f(\mathbf{x})\cdot\Psi\rangle_{\mathcal{H}} = \langle f, f - \mathcal{L}f\rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2 - \langle f, \mathcal{L}f\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}^2,$$

where the last inequality holds because that $\mathcal{L}$ is positive. ■

Proof [Proof of Lemma 5] By the block structure, it holds that $\|P^{-1}\|_2 = \max_{i\in[M]}\|P_{ii}^{-1}\|_2 = \frac{1}{\lambda_{\min}(P_{ii})}$ and $\|P\|_2 = \max_{i\in[M]}\|P_{ii}\|_2$. For FedAvg, we have $\|P_{ii}\|_2 \leq s$, and

$$\lambda_{\min}(P_{ii}) \geq \sum_{\tau=0}^{s-1}(1-\gamma)^\tau = \frac{1-(1-\gamma)^s}{\gamma}.$$

For FedProx, we have $\|P_{ii}\|_2 \leq 1$ and $\lambda_{\min}(P_{ii}) \geq (1+\gamma)^{-1}$. Finally, since the eigenvalues of $K_{\mathbf{x}}P$ coincide with those of $P^{1/2}K_{\mathbf{x}}P^{1/2}$, (15) follows from Ostrowski's inequality (see e.g. (Horn and Johnson, 2012, Theorem 4.5.9)) and the fact that $s/\kappa \leq \lambda_{\min}(P_{ii}) \leq \|P_{ii}\|_2 \leq s$ for FedAvg and $1/\kappa \leq \lambda_{\min}(P_{ii}) \leq \|P_{ii}\|_2 \leq 1$ for FedProx. ■

Appendix B. Proofs in Section 6.1

In this section, we present the missing proofs of results in Section 6.1. We focus on proving the results for FedAvg. The proof for FedProx follows verbatim using the facts that $\|P\|_2 \leq 1$ and that $\left\|P^{-1}\right\|_2 \leq \kappa$.

One key idea is to apply the eigenvalue decomposition of $K_{\mathbf{x}}P$ and to project $f(\mathbf{x})$ to the eigenspace of $K_{\mathbf{x}}P$ for $f \in \mathcal{H}$. We first describe the eigen-decomposition and present

7. Recall that two matrices $A, B \in \mathbb{R}^{n\times n}$ are similar if there exists an invertible matrix Q such that $B = Q^{-1}AQ$.

bounds on relevant matrix norms. Recall that P is positive-definite, and thus $K_{\mathbf{x}}P$ is similar to $P^{1/2}K_{\mathbf{x}}P^{1/2}$, whose eigenvalue decomposition is denoted as

$$P^{1/2}K_{\mathbf{x}}P^{1/2} = U\Lambda U^{\top}, \tag{40}$$

where U is unitary, i.e., $U^{\top} = U^{-1}$, and $\Lambda = \mathsf{diag}\{\Lambda_1, \ldots, \Lambda_N\}$ is a $N \times N$ diagonal matrix with non-negative entries. Let $V \triangleq P^{-1/2}U$ and $L \triangleq I - \eta K_{\mathbf{x}}P$. Then,

$$K_{\mathbf{x}} = V\Lambda V^{\top}, \qquad K_{\mathbf{x}}P = V\Lambda V^{-1}, \qquad L^t = V(I-\eta\Lambda)^t V^{-1}. \tag{41}$$

By the definition of V,

$$\|V\|_2^2 = \left\|P^{-1}\right\|_2 \leq \kappa/s, \qquad \left\|V^{-1}\right\|_2^2 = \|P\|_2 \leq s, \tag{42}$$

where the upper bounds of $\left\|P^{-1}\right\|_2$ and $\|P\|_2$ are derived in the proof of Lemma 5. Since $\gamma < 1$, Lemma 4 shows that $\|\eta\Lambda\|_2 \leq 1$ and thus $\|I - (I-\eta\Lambda)^t\|_2 \leq 1$. Therefore, using the eigenvalue decomposition in (41) and the upper bounds in (42), we have

$$\left\|I - L^t\right\|_2 = \left\|V(I-(I-\eta\Lambda)^t)V^{-1}\right\|_2 \leq \sqrt{\kappa}. \tag{43}$$

Lemma 19

$$\left\|I - L^t\right\|_{\mathrm{F}} \leq \sqrt{\kappa}\left\|I - (I-\eta\Lambda)^t\right\|_{\mathrm{F}} \leq \sqrt{\kappa N \eta t s}\mathcal{R}\left(\frac{1}{\sqrt{\eta t s}}\right),$$

where $\mathcal{R}$ is the empirical Rademacher complexity defined in (22)*.*

Proof Applying the the eigenvalue decomposition (41) and the inequality $\|AB\|_{\mathrm{F}} \leq \|A\|_2 \|B\|_{\mathrm{F}}$ (see, e.g., (Horn and Johnson, 2012, 5.6.P20)), we obtain

$$\left\|I - L^t\right\|_{\mathrm{F}} = \left\|V(I-(I-\eta\Lambda)^t)V^{-1}\right\|_{\mathrm{F}} \leq \|V\|_2 \left\|V^{-1}\right\|_2 \left\|I-(I-\eta\Lambda)^t\right\|_{\mathrm{F}} \leq \sqrt{\kappa}\left\|I-(I-\eta\Lambda)^t\right\|_{\mathrm{F}},$$

where the last inequality used (42).

Next we prove the second inequality. Since $0 \leq \eta\Lambda_i \leq 1$, it holds that

$$\begin{aligned}
\left\|I-(I-\eta\Lambda)^t\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{N}\left(1-(1-\eta\Lambda_i)^t\right)^2 \\
&\leq \sum_{i=1}^{N}\min\left\{1, \eta^2 t^2 \Lambda_i^2\right\} \\
&\leq \sum_{i=1}^{N}\min\left\{1, \eta t \Lambda_i\right\} \\
&\leq \sum_{i=1}^{N}\min\left\{1, \eta t \lambda_i s\right\},
\end{aligned}$$

where the last inequality holds because $\Lambda_i \leq \lambda_i \|P\| \leq \lambda_i s$ in view of (15). The conclusion follows from the definition of $\mathcal{R}$ in (22). ■

B.1 Proof of Proposition 7

We show the convergence of the prediction error (16). Plugging $y = f(\mathbf{x}) + \Delta_f + \xi$ into (13), we get that

$$f_t(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P\left(f(\mathbf{x}) + \Delta_f + \xi\right).$$

Subtracting $f(\mathbf{x})$ from both hand sides yields that

$$f_t(\mathbf{x}) - f(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\left(f_{t-1}(\mathbf{x}) - f(\mathbf{x})\right) + \eta K_{\mathbf{x}} P\left(\Delta_f + \xi\right).$$

Unrolling the above recursion and recalling $L \triangleq I - \eta K_{\mathbf{x}} P$, we deduce that

$$\begin{aligned} f_t(\mathbf{x}) - f(\mathbf{x}) &= L^t(f_0(\mathbf{x}) - f(\mathbf{x})) + \sum_{\tau=0}^{t-1} L^\tau \eta K_{\mathbf{x}} P\left(\xi + \Delta_f\right) \\ &= L^t(f_0(\mathbf{x}) - f(\mathbf{x})) + \left(I - L^t\right)\left(\xi + \Delta_f\right), \end{aligned} \tag{44}$$

where the last equality follows from the identity that $\sum_{\tau=0}^{t-1}(I-A)^\tau A = I - (I-A)^t$. It follows that

$$\begin{aligned} \|f_t(\mathbf{x}) - f(\mathbf{x})\|_2^2 &\le 3\left\|L^t(f_0(\mathbf{x}) - f(\mathbf{x}))\right\|_2^2 + 3\left\|(I - L^t)\xi\right\|_2^2 + 3\left\|(I - L^t)\Delta_f\right\|_2^2 \\ &\le 3\left\|L^t(f_0(\mathbf{x}) - f(\mathbf{x}))\right\|_2^2 + 3\left\|(I - L^t)\xi\right\|_2^2 + 3\kappa\left\|\Delta_f\right\|_2^2, \end{aligned} \tag{45}$$

where the last inequality holds due to (43). To finish the proof of Proposition 7, it suffices to apply the following two lemmas, which bound the first (bias) and the second (variance) terms in (45), respectively.

Lemma 20 (Bias) *For all iterations $t = 1, 2, \ldots$, it holds that*

$$\frac{1}{N}\left\|L^t(f_0(\mathbf{x}) - f(\mathbf{x}))\right\|_2^2 \le \kappa\delta_1(t)\left\|f_0 - f\right\|_{\mathcal{H}}^2,$$

where

$$\delta_1(t) \triangleq \frac{1}{s}\max_{1\le i\le N}\left(1 - \eta\Lambda_i\right)^{2t}\Lambda_i \le \frac{1}{2e\eta t s}.$$

Proof For any $f \in \mathcal{H}$, it follows from Lemma 18 that $\frac{1}{N} f(\mathbf{x}) f(\mathbf{x})^\top \preceq \|f\|_{\mathcal{H}}^2 K_{\mathbf{x}}$. Then,

$$\frac{1}{N}\left\|L^t f(\mathbf{x})\right\|_2^2 = \frac{1}{N}\left\|L^t f(\mathbf{x}) f(\mathbf{x})^\top (L^t)^\top\right\|_2 \le \|f\|_{\mathcal{H}}^2\left\|L^t K_{\mathbf{x}} (L^t)^\top\right\|_2.$$

Applying (41) yields that

$$\left\|L^t K_{\mathbf{x}} (L^t)^\top\right\|_2 = \left\|V[(I - \eta\Lambda)^{2t}\Lambda]V^\top\right\|_2 \le \left\|(1 - \eta\Lambda)^{2t}\Lambda\right\|_2 \|V\|_2^2 \le \left\|(1 - \eta\Lambda)^{2t}\Lambda\right\|_2 \kappa/s,$$

where the last inequality used (42). The conclusion follows by noting that $f - f_0 \in \mathcal{H}$, $\|\eta\Lambda\|_2 \le 1$ by Lemma 4, and $(1-x)^t x \le \frac{1}{et}$ for all $x \le 1$. ■

Lemma 21 (Variance) *For all iterations $t = 1, 2, \ldots,$ it holds that*

$$\frac{1}{N}\mathbb{E}\left[\left\|(I - L^t)\xi\right\|_2^2\right] \leq \kappa\sigma^2\delta_2(t), \tag{46}$$

where

$$\delta_2(t) \triangleq \frac{1}{N}\sum_{i=1}^{N}\left(1 - (1 - \eta\Lambda_i)^t\right)^2 \leq \frac{1}{N}\sum_{i=1}^{N}\min\left\{1, \eta t\Lambda_i\right\} \leq \eta ts\mathcal{R}^2\left(\frac{1}{\sqrt{\eta ts}}\right).$$

Proof Note that

$$\left\|(I - L^t)\xi\right\|_2^2 = \xi^\top Q\xi = \mathsf{Tr}\left(\xi\xi^\top Q\right), \tag{47}$$

where $Q = (I - L^t)^\top(I - L^t) \succeq 0$. By the assumption $\mathbb{E}\left[\xi\xi^\top\right] \preceq \sigma^2 I$ and the fact

$$\mathsf{Tr}(YZ) \geq 0, \quad \text{if } Y \succeq 0 \text{ and } Z \succeq 0, \tag{48}$$

we have

$$\mathbb{E}\|(I - L^t)\xi\|_2^2 = \mathsf{Tr}\left(\mathbb{E}[\xi\xi^\top]Q\right) \leq \sigma^2\mathsf{Tr}(Q) = \sigma^2\left\|I - L^t\right\|_{\mathrm{F}}^2. \tag{49}$$

Then (46) follows from Lemma 19. ■

B.2 Proofs of Theorems 8 – 11

We first deduce the convergence result with early stopping from Proposition 7.

Proof [Proof of Theorem 8] Plugging the upper bounds of $\delta_1(t)$ and $\delta_2(t)$ in Lemma 20 and Lemma 21 into (18) in Proposition 7, we get that

$$\begin{aligned}\frac{1}{N}\mathbb{E}_\xi\left[\|f_t(\mathbf{x}) - f(\mathbf{x})\|_2^2\right] &\leq \frac{3\kappa}{2e\eta ts}\|f_0 - f\|_{\mathcal{H}}^2 + 3\kappa\sigma^2\eta ts\mathcal{R}^2\left(\frac{1}{\sqrt{\eta ts}}\right) + \frac{3\kappa}{N}\|\Delta_f\|_2^2 \\ &\leq \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\|\Delta_f\|_2^2, \quad \forall 1 \leq t \leq T,\end{aligned}$$

where the last inequality holds because by the definition of early stopping time T given in (23), we have $\eta ts\mathcal{R}(1/\sqrt{\eta ts}) \leq 1/(\sqrt{2e}\sigma)$ for all $t \leq T$ given that $\mathcal{R}(\epsilon)/\epsilon^2$ is non-increasing in ϵ. ■

Then we deduce the convergence results that hold with high probability.

Proof [Proof of Theorem 10] For any $t \geq 0$, let

$$q_t \triangleq \mathbb{P}\left\{\frac{1}{N}\left\|\left[I - L^t\right]\xi\right\|_2^2 \geq \frac{1}{N}\mathbb{E}\left[\left\|\left[I - L^t\right]\xi\right\|_2^2\right] + \frac{\delta}{N}\right\},$$

where $\delta = N\kappa/(e\eta ts)$. By definition of ε_t in (24) and Theorem 8, we have $\varepsilon_t \leq q_t$. Recall from (47) that $\left\|(I - L^t)\xi\right\|_2^2 = \mathsf{Tr}\left(\xi\xi^\top Q\right)$. It remains to show the concentration of the quadratic expression $\mathsf{Tr}\left(\xi\xi^\top Q\right)$.

Sub-Gaussian noise. Using Hanson-Wright's inequality Rudelson and Vershynin (2013), we get

$$\mathbb{P}\left\{\mathsf{Tr}\left(\xi\xi^\top Q\right) - \mathbb{E}\left[\mathsf{Tr}\left(\xi\xi^\top Q\right)\right] \geq \delta\right\} \leq \exp\left(-c_1 \min\left\{\frac{\delta}{\sigma^2\left\|Q\right\|_2}, \frac{\delta^2}{\sigma^4\|Q\|_{\mathrm{F}}^2}\right\}\right), \tag{50}$$

where $c_1 > 0$ is a universal constant. Note that

$$\|Q\|_2 \leq \left\|V^{-1}\right\|_2^2 \|V\|_2^2 = \|P\|_2 \left\|P^{-1}\right\|_2 \leq \kappa, \tag{51}$$

$$\|Q\|_{\mathrm{F}}^2 = \mathsf{Tr}(QQ^\top) \leq \|Q\|_2 \,\mathsf{Tr}(Q), \tag{52}$$

where the last inequality follows from (48). Applying Lemma 19 yields that

$$\frac{\delta^2}{\sigma^4\|Q\|_{\mathrm{F}}^2} \geq \frac{\delta}{\sigma^2\left\|Q\right\|_2}\frac{\delta}{\sigma^2\mathsf{Tr}(Q)} \geq \frac{\delta}{\sigma^2\left\|Q\right\|_2}\frac{\delta}{\sigma^2\kappa N\eta t s\mathcal{R}^2\left(\frac{1}{\sqrt{\eta t s}}\right)},$$

$$\frac{\delta}{\sigma^2\left\|Q\right\|_2} \geq \frac{\delta}{\sigma^2\kappa}.$$

Thus we obtain that

$$\min\left\{\frac{\delta}{\sigma^2\left\|Q\right\|_2}, \frac{\delta^2}{\sigma^4\|Q\|_{\mathrm{F}}^2}\right\} \geq \frac{\delta}{\sigma^2\kappa}\min\left\{1, \frac{\delta}{\sigma^2\kappa N\eta t s\mathcal{R}^2\left(\frac{1}{\sqrt{\eta t s}}\right)}\right\}. \tag{53}$$

Combining (49), (50), and (53) yields that

$$q \leq \exp\left(-\frac{c_1\delta}{\sigma^2\kappa}\min\left\{1, \frac{\delta}{\sigma^2\kappa N\eta t s\mathcal{R}^2\left(1/\sqrt{\eta t s}\right)}\right\}\right). \tag{54}$$

Recalling $\delta = N\kappa/(e\eta t s)$, we deduce that for $t \leq T$,

$$\begin{aligned} q &\leq \exp\left(-\frac{c_1 N}{\sigma^2 e\eta t s}\min\left\{1, \frac{1}{\sigma^2 e(\eta t s)^2\mathcal{R}^2\left(\frac{1}{\sqrt{\eta t s}}\right)}\right\}\right) \\ &\leq \exp\left(-\frac{c_1 N}{\sigma^2 e\eta t s}\right), \end{aligned}$$

where the last inequality holds due to $\eta t s\mathcal{R}(1/\sqrt{\eta t s}) \leq 1/(\sqrt{2e}\sigma)$ for $t \leq T$.

Heavy-tailed noise. We first prove a concentration inequality analogous to the Hanson-Wright inequality. Note that $\mathsf{Tr}\left(\xi\xi^\top Q\right) = \xi^\top Q\xi$. We decompose the deviation into two parts and bound their tail probabilities separately:

$$\begin{aligned} \mathbb{P}\left[\left|\xi^\top Q\xi - \mathbb{E}[\xi^\top Q\xi]\right| > \delta\right] &\leq \mathbb{P}\left\{\left|\sum_i Q_{ii}(\xi_i^2 - \mathbb{E}\xi_i^2)\right| > \frac{\delta}{2}\right\} + \mathbb{P}\left\{\left|\sum_{i\neq j} Q_{ij}\xi_i\xi_j\right| > \frac{\delta}{2}\right\} \\ &\leq \frac{\mathbb{E}\left|\sum_i Q_{ii}(\xi_i^2 - \mathbb{E}\xi_i^2)\right|^{p/2}}{(\delta/2)^{p/2}} + \frac{\mathbb{E}|\sum_{i\neq j} Q_{ij}\xi_i\xi_j|^p}{(\delta/2)^p}. \end{aligned} \tag{55}$$

The first term involves a sum of independent random variables. Since $\mathbb{E}(\xi_i^2 - \mathbb{E}\xi_i^2)^2 \leq \mathbb{E}|\xi_i|^4 \triangleq M_4$ and $\mathbb{E}|\xi_i^2 - \mathbb{E}\xi_i^2|^{p/2} \leq 2^{p/2}\mathbb{E}|\xi_i|^p$, by the Rosenthal-type inequality (Pinelis, 1994, Theorem 5.2),

$$\mathbb{E}\left|\sum_i Q_{ii}(\xi_i^2 - \mathbb{E}\xi_i^2)\right|^{p/2} \leq C_p\left(M_p\sum_i |Q_{ii}|^{p/2} + \left(M_4\sum_i |Q_{ii}|^2\right)^{p/4}\right) \leq 2C_p\,\|Q\|_{\mathrm{F}}^{p/2}\,M_p, \tag{56}$$

where C_p only depends on p, and we used $M_4^{1/4} \leq M_p^{1/p}$ and $\|x\|_p \leq \|x\|_q$ for $p \geq q$. For the second term, the decoupling inequality gives de la Peña and Montgomery-Smith (1995):

$$\mathbb{E}\left|\sum_{i\neq j} Q_{ij}\xi_i\xi_j\right|^p \leq 4^p\mathbb{E}\left|\xi^\top Q\xi'\right|^p,$$

where ξ' is an independent copy of ξ. By the moment inequalities for decoupled U-statistics (Giné et al., 2000, Proposition 2.4), we have

$$\mathbb{E}\left|\xi^\top Q\xi'\right|^p \leq C_p\left(\sigma^{2p}\,\|Q\|_{\mathrm{F}}^p + \sigma^p M_p\|Q\|_{2,p}^p + M_p^2\|Q\|_{p,p}^p\right) \leq 3C_p\,\|Q\|_{\mathrm{F}}^p\,M_p^2, \tag{57}$$

where $\|\cdot\|_{p,q}$ denotes the $L_{p,q}$-norm given by $\|A\|_{p,q}^q = \sum_j(\sum_i |A_{ij}|^p)^{q/p}$. Finally, we apply (56) – (57) in the upper bound (55) and obtain

$$\mathbb{P}\left[\left|\xi^\top Q\xi - \mathbb{E}[\xi^\top Q\xi]\right| > \delta\right] \leq C_p'\left(M_p\left(\frac{\|Q\|_{\mathrm{F}}}{\delta}\right)^{p/2} + M_p^2\left(\frac{\|Q\|_{\mathrm{F}}}{\delta}\right)^p\right) \tag{58}$$

for some constant $C_p' \geq 1$ only depends on p. We claim that

$$\mathbb{P}\left[\left|\xi^\top Q\xi - \mathbb{E}[\xi^\top Q\xi]\right| > \delta\right] \leq 2C_p'M_p\left(\frac{\|Q\|_{\mathrm{F}}}{\delta}\right)^{p/2}.$$

This is because when $\|Q\|_{\mathrm{F}} \geq \delta$, the last display equation automatically holds; otherwise, it follows from (58).

Recalling $\delta = N\kappa/(e\eta ts)$ and $\|Q\|_{\mathrm{F}} \leq \sqrt{\|Q\|_2}\sqrt{\mathsf{Tr}(Q)} \leq \kappa\sqrt{N\eta ts}\mathcal{R}(1/\sqrt{\eta ts})$, we deduce that for $t \leq T$,

$$\begin{aligned} q &\leq 2C_p'M_p\left(\frac{\kappa\sqrt{N\eta ts}\mathcal{R}(1/\sqrt{\eta ts})e\eta ts}{N\kappa}\right)^{p/2} \\ &\leq 2C_p'M_p\left(\frac{e\eta ts}{2N\sigma^2}\right)^{p/4}. \end{aligned}$$

■

Finally we deduce the exponential convergence result from Proposition 7 in the special case of finite-rank kernels.

Proof [Proof of Theorem 11] Since $\Lambda_i = 0$ for $i > d$, in view of Lemma 20, we have

$$\delta_1(t) \leq \frac{1}{s} \max_{1 \leq i \leq d} \left(1 - \eta \Lambda_i\right)^{2t} \Lambda_i \leq \frac{1}{\eta s} \left(1 - \eta \lambda_d s/\kappa\right)^{2t} \leq \frac{1}{\eta s} \exp\left(-2\lambda_d \eta t s/\kappa\right),$$

where the second inequality holds due to $\Lambda_d \geq \lambda_d s/\kappa$ in view of (15). Moreover, in view of Lemma 21, we have $\delta_2(t) \leq d/N$. It follows from Proposition 7 that

$$\frac{1}{N}\mathbb{E}_\xi\left[\|f_t(\mathbf{x}) - f(\mathbf{x})\|_2^2\right] \leq 3\frac{\kappa}{\eta s}\|f_0 - f\|_{\mathcal{H}}^2 \exp\left(-2\frac{\eta s}{\kappa}\lambda_d t\right) + 3\kappa\sigma^2\frac{d}{N} + \frac{3\kappa}{N}\|\Delta_f\|_2^2, \quad \forall t.$$

■

B.3 Convergence of prediction error in $L^2(\mathbb{P})$ norm

We establish the convergence of prediction error in $L^2(\mathbb{P})$ norm when the the covariates $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ as stated in Remark 9, that is,

$$\|f_T - f^*\|_2^2 \triangleq \int_{\mathcal{X}} (f_T - f)^2 d\mathbb{P}.$$

The convergence is characterized in terms of the eigenvalues of the kernel function $k(x, y)$. Let $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq 0$ denote the eigenvalues of $k(x, y)$ in $L^2(\mathbb{P})$. Define the population Rademacher complexity

$$\bar{R}(\epsilon) = \sqrt{\frac{1}{N}\sum_{i=1}^{\infty} \min\{\bar{\lambda}_i, \epsilon^2\}}.$$

and the critical radius

$$\bar{\epsilon}_N = \inf\left\{\epsilon > 0 : \bar{R}(\epsilon) \leq \frac{\epsilon^2}{\sqrt{2e}\sigma}\right\}.$$

Theorem 22 *Suppose that x_{ij}'s are sampled i.i.d. according to distribution $\mathbb{P}$ and $\gamma < 1$. The coordinates of the noise vector ξ are N independent zero-mean and sub-Gaussian variables (with a constant sub-Gaussian norm σ). There exist universal constants c_1, c_2, c_3 such that with probability at least $1 - c_1 \exp\left(-c_2 N \bar{\epsilon}_N^2\right)$,*

$$\|f_T - f\|_2^2 \leq c_3 \kappa \left(\bar{\epsilon}_N^2\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{\|\Delta_f\|_2^2}{N}\right).$$

As immediate corollaries of Theorem 22, we show that both FedAvg (with $s \geq 1$) and FedProx achieve the centralized minimax-optimal estimation error rate for the following specific kernels Yang and Barron (1999); Raskutti et al. (2012).

Finite rank kernels Consider the class of RKHS with finite-rank kernels, that is, there exists a finite r such that $\bar{\lambda}_j = 0$ for $j > r+1$. For example, the kernel $k(x, y) = (1+\langle x, y\rangle)^a$ for $x, y \in \mathbb{R}^d$ generates the RKHS of all multivariate polynomials of $2d$ variables and degree at most $2a$.

Corollary 23 *Suppose in addition to the conditions of Theorem 22, the kernel k has finite rank r and $\Delta_{f^*} = 0$. Then with probability $1 - c_1 \exp(-c_2 r)$,*

$$\|f_T - f^*\|_2^2 \leq c_3 \kappa \sigma^2 \frac{r}{N}$$

for some universal constant $c_3 > 0$.

Kernels with polynomial eigenvalue decay Consider the kernels whose eigenvalues exhibit a polynomial decay with exponent β, that is,

$$\bar{\lambda}_i \leq C i^{-2\beta} \quad \text{for } \beta > 1/2 \text{ and a constant } C > 0. \tag{59}$$

For example, the first-order Sobolev kernel $k(x, y) = \min\{x, y\}$ on the unit square $[0, 1]^2$ satisfies the eigenvalue decay (59) with $\beta = 1$.

Corollary 24 *Suppose in addition to the conditions of Theorem 22, the kernel k satisfies (59) and $\Delta_{f^*} = 0$. Then with probability at least $1 - c_1 \exp\left(-c_2 N^{1/(2\beta+1)}\right)$,*

$$\|f_T - f^*\|_2^2 \leq c_3 \kappa \left(\frac{\sigma^2}{N}\right)^{\frac{2\beta}{2\beta+1}} \tag{60}$$

for some universal constant $c_3 > 0$.

Next we prove Theorem 22. We focus on proving the result for FedAvg. The proof for FedProx follows verbatim. We need two standard results (see e.g. (Raskutti et al., 2014, Lemma 10, 11) and (Wainwright, 2019, Theorem 14.1, Proposition 14.25)), which connect the prediction error in $L^2(\mathbb{P}_N)$ norm and that in $L^2(\mathbb{P})$, and the critical radius ϵ_N and $\bar{\epsilon}_N$, respectively.

Lemma 25 *Let $\mathcal{G}$ denote the set of functions $g \in \mathcal{H}$ such that $\|g\|_\infty \leq B$ and $\|g\|_{\mathcal{H}} \leq B$ for some $B > 0$. Suppose that x_{ij}'s are sampled i.i.d. according to distribution $\mathbb{P}$ and σ is a constant. Then there exists universal constants (c_1, c_2, c_3) such that for any $\delta \geq \bar{\epsilon}_N$, with probability at least $1 - c_1 \exp(-c_2 N \delta^2)$.*

$$\left| \|g\|_N^2 - \|g\|_2^2 \right| \leq c_3 B^2 \delta^2, \quad \forall g \in \mathcal{G}.$$

Lemma 26 *There exist constants c_1, c_2, c_3, c_4 such that with probability at least $1 - c_1 \exp(-c_2 N \bar{\epsilon}_N^2)$,*

$$\frac{c_3}{\sqrt{\eta T s}} \leq \bar{\epsilon}_N \leq \frac{c_4}{\sqrt{\eta T s}}.$$

Now we are ready to prove Theorem 22.

Proof [Proof of Theorem 22] It follows from Lemmas 26 and 27 that with probability at least $1-\exp\left(-c_1 N\epsilon_N^2/\sigma^2\right)$,

$$\|f_T - f\|_{\mathcal{H}} \le \|f_0 - f\|_{\mathcal{H}} + 1 + \sqrt{\frac{\eta s T}{N}}\|\Delta_f\|_2 \triangleq B.$$

Applying Lemma 25 with $\delta = \bar{\epsilon}_N$, $g = f_T - f$, and Lemma 26, we get that with probability at least $1 - c_1\exp(-c_2 N\bar{\epsilon}_N^2)$,

$$\|f_T - f\|_2^2 \le \|f_T - f^*\|_N^2 + c_3 B^2 \bar{\epsilon}_N^2.$$

Recall that from Theorem 10 in the main paper, we get that with probability at least $1-\exp(-c_1 N\bar{\epsilon}_N^2)$,

$$\|f_T - f\|_N^2 \le \frac{3\kappa}{2e\eta T s}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 3\right) + 3\frac{\kappa}{N}\|\Delta_f\|_2^2.$$

Combining the last two displayed equation yields that with probability at least $1-\exp(-c_1 N\bar{\epsilon}_N^2)$,

$$\|f_T - f\|_2^2 \le c_4\kappa\left(\bar{\epsilon}_N^2\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{\|\Delta_f\|_2^2}{N}\right).$$

■

Proof [Proof of Corollary 23] Since the kernel k has finite rank r, we have $\bar{\lambda}_i = 0$ for all $i > r$ and hence

$$\bar{R}(\epsilon) = \sqrt{\frac{1}{N}\sum_{i=1}^{\infty}\min\{\bar{\lambda}_i, \epsilon^2\}} = \sqrt{\frac{1}{N}\sum_{i=1}^{r}\min\{\bar{\lambda}_i, \epsilon^2\}} \le \sqrt{\frac{r}{N}}\epsilon.$$

Therefore, by the definition of $\bar{\epsilon}_N$, we get that

$$\frac{\bar{\epsilon}_N^2}{\sqrt{2e}\sigma} = \bar{R}_k(\epsilon_N) \le \sqrt{\frac{r}{N}}\bar{\epsilon}_N,$$

and hence

$$\bar{\epsilon}_N \le \sigma\sqrt{\frac{2er}{N}},$$

which completes the proof in view of Theorem 22. ■

Proof [Proof of Corollary 24] Since the kernel k satisfies the eigenvalue decay (59), we have

$$\bar{R}(\epsilon) \le \sqrt{\frac{1}{N}\sum_{i=1}^{\infty}\min\{Ci^{-2\beta}, \epsilon^2\}} \le \frac{C'}{\sqrt{N}}\epsilon^{1-1/(2\beta)},$$

where the last inequality follows from (Raskutti et al., 2014, Corollary 3). Therefore, by the definition of $\bar{\epsilon}_N$, we get that

$$\frac{\bar{\epsilon}_N^2}{\sqrt{2e}\sigma} = \bar{R}(\epsilon_N) \leq \frac{C'}{\sqrt{N}} \bar{\epsilon}_N^{1-1/(2\beta)},$$

and hence

$$\bar{\epsilon}_N \leq \left(\frac{\sqrt{2e} C' \sigma}{\sqrt{N}} \right)^{\frac{2\beta}{2\beta+1}},$$

which completes the proof in view of Theorem 22. ■

B.4 Upper bound to the RKHS norm

Lemma 27 *Suppose that the coordinates of the noise vector ξ are N independent zero-mean and sub-Gaussian variables (with sub-Gaussian norm bounded by σ)/ There exists a universal constant c such that, for any $t \leq T$, with probability at least $1 - \exp\left(-cN/(\sigma^2 \eta t s)\right)$,*

$$\|f_t - f\|_{\mathcal{H}} \leq \|f_0 - f\|_{\mathcal{H}} + 1 + \sqrt{\frac{\eta s t}{N}} \|\Delta_f\|_2, \quad \forall\ f \in \mathcal{H}.$$

Proof Similar to (66), for any $f \in \mathcal{H}$, we use Δ_f in (21) and obtain

$$f_t - f = \mathcal{L}^t (f_0 - f) + \sum_{\tau=0}^{t-1} \mathcal{L}^\tau ((\xi + \Delta_f) \cdot \Psi). \tag{61}$$

It follows from Lemma 4 that $\|\mathcal{L}\|_{\text{op}} \leq 1$ and

$$\|\mathcal{L}^t (f_0 - f)\|_{\mathcal{H}} \leq \|f_0 - f\|_{\mathcal{H}}. \tag{62}$$

For the second term of (61), using the matrix Σ defined in (67), we have $\|\sum_{\tau=0}^{t-1} \mathcal{L}^\tau (a \cdot \Psi)\|_{\mathcal{H}}^2 = a^\top \Sigma a$ for any $a \in \mathbb{R}^N$. Applying Lemma 28 with $\mathcal{T} = \sum_{\tau=0}^{t-1} \mathcal{L}^\tau$ yields that

$$\|\Sigma\|_2 \leq \frac{s\eta}{N} \left\| \sum_{\tau=0}^{t-1} \mathcal{L}^\tau (\mathcal{I} - \mathcal{L}^t) \right\|_{\text{op}} \leq \frac{s\eta t}{N}.$$

Therefore,

$$\left\| \sum_{\tau=0}^{t-1} \mathcal{L}^\tau (\Delta_f \cdot \Psi) \right\|_{\mathcal{H}} = \sqrt{\Delta_f \Sigma \Delta_f} \leq \sqrt{\|\Sigma\|_2} \|\Delta_f\|_2 \leq \sqrt{\frac{\eta s t}{N}} \|\Delta_f\|_2. \tag{63}$$

Finally we consider $\|\sum_{\tau=0}^{t-1} \mathcal{L}^\tau (\xi \cdot \Psi)\|_{\mathcal{H}}^2 = \xi^\top \Sigma \xi$. Recall the early stopping rule (23), which implies that $\eta t s \mathcal{R}(1/\sqrt{\eta t s}) \leq 1/(\sqrt{2e}\sigma)$ for $t \leq T$. Then, by Lemma 30,

$$\mathbb{E}[\xi^\top \Sigma \xi] = \mathbb{E}[\mathsf{Tr}(\xi \xi^\top \Sigma)] \leq \sigma^2 \mathsf{Tr}(\Sigma) \leq \left(\sigma \eta t s \mathcal{R}\left(\frac{1}{\sqrt{\eta t s}} \right) \right)^2 \leq \frac{1}{2e}. \tag{64}$$

Using the Hanson-Wright inequality Rudelson and Vershynin (2013), for a universal constant c_1,

$$\mathbb{P}\left\{\xi^\top\Sigma\xi - \mathbb{E}[\xi^\top\Sigma\xi] \geq \delta\right\} \leq \exp\left(-c_1 \min\left\{\frac{\delta}{\sigma^2\left\|\Sigma\right\|_2}, \frac{\delta^2}{\sigma^4\|\Sigma\|_{\mathrm{F}}^2}\right\}\right).$$

Since $\|\Sigma\|_{\mathrm{F}}^2 \leq \|\Sigma\|_2 \mathsf{Tr}(\Sigma)$. Choosing $\delta = \frac{1}{2e}$ and invoking $\sigma^2\mathsf{Tr}(\Sigma) \leq \delta$ from (64) and $\|\Sigma\|_2 \leq \eta st/N$, we get that

$$\mathbb{P}\left\{\xi^\top\Sigma\xi - \mathbb{E}[\xi^\top\Sigma\xi] \geq \delta\right\} \leq \exp\left(-c_1 N/(2\sigma^2 e\eta st)\right). \tag{65}$$

Hence, combining (62) – (65), we conclude the proof from (61). ■

Lemma 28 *Suppose $\mathcal{T} : \mathcal{H} \mapsto \mathcal{H}$ is a self-adjoint linear operator. Let A be a $N \times N$ matrix with $A_{ij} = \langle\mathcal{T}(\Psi_i), \mathcal{T}(\Psi_j)\rangle_{\mathcal{H}}$. Then,*

$$\|A\|_2 \leq \frac{s\eta}{N}\left\|\mathcal{T}(\mathcal{I}-\mathcal{L})\mathcal{T}\right\|_{\mathrm{op}}, \qquad \mathsf{Tr}(A) \leq \frac{s\eta}{N}\mathsf{Tr}(\mathcal{T}(\mathcal{I}-\mathcal{L})\mathcal{T}).$$

Proof By definition, $a^\top Aa = \|\mathcal{T}(a\cdot\Psi)\|_{\mathcal{H}}^2$ for any $a \in \mathbb{R}^N$. Therefore,

$$\|A\|_2 = \max_{\|a\|_2\leq 1}\|\mathcal{T}(a\cdot\Psi)\|_{\mathcal{H}}^2 = \max_{\|a\|_2\leq 1}\max_{\|f\|_{\mathcal{H}}\leq 1}\langle f, \mathcal{T}(a\cdot\Psi)\rangle_{\mathcal{H}}^2.$$

For any $f \in \mathcal{H}$ and $a \in \mathbb{R}^N$ with $\|f\|_{\mathcal{H}} \leq 1$ and $\|a\|_2 \leq 1$, we have

$$\langle f, \mathcal{T}(a\cdot\Psi)\rangle_{\mathcal{H}} = \langle\mathcal{T}f, a\cdot\Psi\rangle_{\mathcal{H}} = \frac{\eta}{N}a^\top Pg(\mathbf{x}) \leq \frac{\eta}{N}\left\|Pg(\mathbf{x})\right\|_2,$$

where $g = \mathcal{T}f$ and the second equality used Lemma 2. Using $\|P\|_2 \leq s$ and Lemma 2, we get

$$\langle f, \mathcal{T}(a\cdot\Psi)\rangle_{\mathcal{H}}^2 \leq \frac{s\eta^2}{N^2}g(\mathbf{x})^\top Pg(\mathbf{x}) = \frac{s\eta}{N}\langle g, g-\mathcal{L}g\rangle_{\mathcal{H}} = \frac{s\eta}{N}\langle f, \mathcal{T}(\mathcal{I}-\mathcal{L})\mathcal{T}f\rangle_{\mathcal{H}}.$$

Next we prove the second inequality. Let $\{\phi_1, \phi_2, \dots\}$ be an orthonormal basis of $\mathcal{H}$, and let $f_i \triangleq \mathcal{T}\phi_i$. By the definition of $\mathcal{T}$ and Lemma 2,

$$\begin{aligned}\mathsf{Tr}(A) &= \sum_j \|\mathcal{T}(\Psi_j)\|_{\mathcal{H}}^2 = \sum_{ij}\langle\phi_i, \mathcal{T}(\Psi_j)\rangle_{\mathcal{H}}^2 = \sum_{ij}\langle f_i, e_j\cdot\Psi\rangle_{\mathcal{H}}^2 \\ &= \sum_{ij}\left(\frac{\eta}{N}e_j^\top Pf_i(\mathbf{x})\right)^2 = \sum_i\left\|\frac{\eta}{N}Pf_i(\mathbf{x})\right\|_2^2.\end{aligned}$$

Since $\|P\|_2 \leq s$, we further have

$$\mathsf{Tr}(A) \leq \frac{s\eta^2}{N^2}\sum_i f_i(\mathbf{x})^\top Pf_i(\mathbf{x}) = \frac{s\eta}{N}\sum_i\langle f_i, f_i - \mathcal{L}f_i\rangle_{\mathcal{H}} = \frac{s\eta}{N}\sum_i\langle\phi_i, \mathcal{T}(I-\mathcal{L})\mathcal{T}\phi_i\rangle_{\mathcal{H}}.$$

■

Appendix C. Proofs in Section 6.2

Again we focus on proving the results for FedAvg. The proof for FedProx follows verbatim using the facts that $\|P\|_2 \le 1$ and that $\left\|P^{-1}\right\|_2 \le \kappa$.

C.1 Proof of Theorem 12

Since the desired conclusion (25) trivially holds when $\rho_N = 0$, we assume $\rho_N > 0$ in the proof.

It follows from Proposition 1 and (27) that

$$\begin{aligned} f_t - \bar{f} &= \mathcal{L}(f_{t-1} - \bar{f}) - (\bar{f} - \mathcal{L}\bar{f}) + y \cdot \Psi = \mathcal{L}(f_{t-1} - \bar{f}) + \xi \cdot \Psi \\ &= \mathcal{L}^t(f_0 - \bar{f}) + \sum_{\tau=0}^{t-1} \mathcal{L}^\tau(\xi \cdot \Psi). \end{aligned} \tag{66}$$

To analyze (66), we show properties of $\mathcal{L}$ and the matrix Σ of size $N \times N$ with

$$\Sigma_{ij} = \left\langle \sum_{\tau=0}^{t-1} \mathcal{L}^\tau(e_i \cdot \Psi), \sum_{\tau=0}^{t-1} \mathcal{L}^\tau(e_j \cdot \Psi) \right\rangle_{\mathcal{H}}. \tag{67}$$

Lemma 29 *For $f \in \mathcal{H}$, define $\mathcal{P}f = \frac{1}{N} f(\mathbf{x}) \cdot k_{\mathbf{x}}$. Then,*

$$\mathcal{I} - s\eta\mathcal{P} \preceq \mathcal{L} \preceq \mathcal{I} - s\eta\mathcal{P}/\kappa, \tag{68}$$

where $\mathcal{T}_1 \preceq \mathcal{T}_2$ means $\mathcal{T}_2 - \mathcal{T}_1$ is positive.

Moreover, assume $\phi(x)$ is d-dimensional. Then there is a one-to-one correspondence between the eigenvalues of $\mathcal{P}$ and those of the d by d matrix $\frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x})$.

Proof We first show that

$$\mathcal{I} - s(\mathcal{I} - \mathcal{L}_i) \preceq \mathcal{L}_i^s = (\mathcal{I} - (\mathcal{I} - \mathcal{L}_i))^s \preceq \mathcal{I} - s(\mathcal{I} - \mathcal{L}_i)/\kappa. \tag{69}$$

Since all terms in (69) are polynomial in $\mathcal{L}_i$, it suffices to show the ordering of corresponding eigenvalues. Suppose λ_j is the j-th eigenvalue of $\mathcal{I} - \mathcal{L}_i$. It is shown in the proof of Lemma 4 that $0 \le \lambda_j \le \gamma$. Then,

$$1 - s\lambda_j \le (1 - \lambda_j)^s \le 1 - s\lambda_j/\kappa.$$

To see the second inequality, we note the function $x \mapsto \frac{x}{1-(1-x)^s}$ is monotone increasing in $[0, 1]$ and thus

$$\kappa = \frac{\gamma s}{1 - (1 - \gamma)^s} \ge \frac{\lambda_j s}{1 - (1 - \lambda_j)^s}.$$

Then (68) follows from (69) and (8) as $\mathcal{L} = \sum_i w_i \mathcal{L}_i^s$ and $(I - \mathcal{L}_i)f = \frac{\eta}{n_i} f(\mathbf{x}_i) \cdot k_{\mathbf{x}_i}$.

It remains to establish the correspondence between the eigenvalues of $\mathcal{P}$ and those of $\frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x})$. Recall that $\{\phi_\ell\}_{\ell=1}^d$ forms an orthonormal basis of $\mathcal{H}$. Thus, it suffices to show that the matrix representation of $\mathcal{P}$ is $\frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x})$, i.e., for any $f = a \cdot \phi$ with $a \in \mathbb{R}^d$, we have $\mathcal{P}f = \left(\frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x}) a\right) \cdot \phi$. This follows from the fact that $\mathcal{P}\phi = \frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x})\phi$ for $\phi = (\phi_1, \ldots, \phi_d)$.

■

Lemma 30 *Let $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \dots\}$ be the eigenvalues of $\mathcal{L}$. Then,*

$$\mathsf{Tr}(\Sigma) \le \frac{\eta s}{N} \sum_i \frac{(1-\widetilde{\lambda}_i^t)^2}{1-\widetilde{\lambda}_i} \le (\eta t s)^2 \mathcal{R}^2 \left(\frac{1}{\sqrt{\eta t s}}\right),$$

where $\mathcal{R}$ is the empirical Rademacher complexity defined in (22).

Proof Applying Lemma 28 with $\mathcal{T} = \sum_{\tau=0}^{t-1} \mathcal{L}^\tau$ yields that

$$\mathsf{Tr}(\Sigma) \le \frac{s\eta}{N} \mathsf{Tr}\left(\sum_{\tau=0}^{t-1} \mathcal{L}^\tau (I - \mathcal{L}^t)\right).$$

Let $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \dots\}$ denote the eigenvalues of $\mathcal{L}$, where $\widetilde{\lambda}_i \in [0,1]$ by Lemma 5. Applying the facts $1 - x^t \le \min\{1, t(1-x)\}$ and $\min\{\frac{1}{x}, t^2 x\} \le \min\{t, t^2 x\}$ for $t \ge 0$ and $0 \le x \le 1$, we obtain

$$\begin{aligned}
\mathsf{Tr}\left(\sum_{\tau=0}^{t-1} \mathcal{L}^\tau (I - \mathcal{L}^t)\right) &= \sum_i \frac{(1-\widetilde{\lambda}_i^t)^2}{1-\widetilde{\lambda}_i} \\
&\le \sum_i \min\left\{\frac{1}{1-\widetilde{\lambda}_i}, t^2(1-\widetilde{\lambda}_i)\right\} \\
&\le \sum_i \min\left\{t, t^2(1-\widetilde{\lambda}_i)\right\}.
\end{aligned}$$

By Lemma 29, we have $1 - \widetilde{\lambda}_i \le s\eta\lambda_i$ for $1 \le i \le N$ and $1 - \widetilde{\lambda}_i = 0$ for $i > N$. It follows that

$$\mathsf{Tr}(\Sigma) \le \frac{s\eta}{N} \sum_{i=1}^{N} \min\left\{t, t^2 s\eta\lambda_i\right\} = t^2 s^2 \eta^2 \mathcal{R}^2 \left(\frac{1}{\sqrt{\eta t s}}\right),$$

where the last equality used the definition of $\mathcal{R}$ in (22). ■

For the first term of (66), by Lemma 29, we get

$$\|\mathcal{L}\|_{\mathrm{op}} \le 1 - \frac{s\eta\lambda_{\min}(\mathcal{P})}{\kappa} = 1 - \frac{s\eta\rho_N}{\kappa}. \tag{70}$$

By linearity, the norm of the second term in (66) can be represented as

$$\left\| \sum_{\tau=0}^{t-1} \mathcal{L}^\tau (\xi \cdot \Psi) \right\|_{\mathcal{H}}^2 = \xi^\top \Sigma \xi = \mathsf{Tr}(\xi\xi^\top \Sigma).$$

By the assumption $\mathbb{E}\left[\xi\xi^\top\right] \preceq \sigma^2 I$ and (48), we have

$$\mathbb{E}\left[\mathsf{Tr}(\xi\xi^\top\Sigma)\right] \le \sigma^2 \mathsf{Tr}(\Sigma) \le \sigma^2 \frac{\eta s}{N} \sum_i \frac{(1-\widetilde{\lambda}_i^t)^2}{1-\widetilde{\lambda}_i}.$$

Recall that $s\eta\rho_N/\kappa \le 1-\widetilde{\lambda}_i \le 1$ by Lemma 4 and (70), and that ϕ is d-dimensional. We obtain

$$\mathbb{E}\mathsf{Tr}(\xi\xi^\top\Sigma) \le \sigma^2\frac{\eta s}{N}\sum_{i=1}^{d}\frac{1}{s\eta\rho_N/\kappa} = \frac{\sigma^2\kappa d}{N\rho_N}. \tag{71}$$

Applying (70) and (71) to (66) yields the desired (25).

It remains to establish (26). By the definition of $\bar{f}$,

$$\bar{f} - f_j^* = (\mathcal{I}-\mathcal{L})^{-1}\left(\Delta_{f_j^*}\cdot\Psi\right).$$

where $\Delta_{f_j^*} = (f_1^*(\mathbf{x}_1) - f_j^*(\mathbf{x}_1), \ldots, f_M^*(\mathbf{x}_M) - f_j^*(\mathbf{x}_M))$ defined in (21). Then by linearity,

$$\|\bar{f} - f_j^*\|_{\mathcal{H}}^2 = \Delta_{f_j^*}^\top S\Delta_{f_j^*} \le \left\|\Delta_{f_j^*}\right\|_2^2 \|S\|_2,$$

where S is a matrix of size $N\times N$ with $S_{ij} = \left\langle(\mathcal{I}-\mathcal{L})^{-1}(e_i\cdot\Psi), (\mathcal{I}-\mathcal{L})^{-1}(e_j\cdot\Psi)\right\rangle_{\mathcal{H}}$. Applying Lemma 28 with $\mathcal{T} = (\mathcal{I}-\mathcal{L})^{-1}$ yields that

$$\|S\|_2 \le \frac{s\eta}{N}\left\|(\mathcal{I}-\mathcal{L})^{-1}\right\|_{\mathrm{op}}.$$

It follows from (70) that $\|S\|_2 \le \frac{\kappa}{N\rho_N}$, which implies (26).

C.2 Proofs of Corollaries 13 – 14

Proof [Proof of Corollary 13] In view of (Vershynin, 2010, Theorem 5.39) and the union bound, with probability at least $1-e^{-d}$,

$$\left\|\frac{1}{N}\phi(\mathbf{x})^\top\phi(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\Sigma_{ij}\right\|_2 \le c_1\max\left\{\sqrt{\frac{d}{N}}, \frac{d}{N}\right\},$$

where c_1 is a universal constant. By the assumption, $\alpha I \preceq \Sigma_{ij} \preceq \beta I$ for some fixed constant $\alpha, \beta > 0$. Therefore,

$$\rho_N = \lambda_{\min}\left(\frac{1}{N}\phi(\mathbf{x})^\top\phi(\mathbf{x})\right) \ge \alpha - c_1\max\left\{\sqrt{\frac{d}{N}}, \frac{d}{N}\right\} \ge \alpha/2,$$

where the first inequality follows from Weyl's inequality and the second inequality holds by choosing $N \ge d\max\{4c_1^2/\alpha^2, 2c_1/\alpha\}$. The desired (28) readily follows from (25).

It remains to prove (29). By the definition of $\|\Delta_{f_j^*}\|_2$, we have

$$\left\|\Delta_{f_j^*}\right\|_2^2 = \sum_{i=1}^{M}\left\|f_i^*(\mathbf{x}_i) - f_j^*(\mathbf{x}_i)\right\|_2^2 = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\langle\phi(x_{ij}), \theta_i^* - \theta_j^*\right\rangle^2. \tag{72}$$

Recall that $\Gamma = \max_{i,j}\left\|\theta_i^* - \theta_j^*\right\|_2$. Thus $\left\langle\phi(x_{ij}), \theta_i^* - \theta_j^*\right\rangle$ are independent and sub-Gaussian random variables with the sub-Gaussian norm bounded by $c_2\Gamma$ for a constant c_2. It follows from the Hanson-Wright inequality that

$$\mathbb{P}\left\{\left\|\Delta_{f_j^*}\right\|_2^2 \ge \mathbb{E}\left[\left\|\Delta_{f_j^*}\right\|_2^2\right] + t\right\} \le \exp\left(-c_3\min\left(\frac{t^2}{\Gamma^4 N}, \frac{t}{\Gamma^2}\right)\right).$$

Setting $t = c_4\Gamma^2 N$ for some large constant c_4, we get that with probability at least $1 - \exp(-N)$,

$$\left\|\Delta_{f_j^*}\right\|_2^2 \le \mathbb{E}\left[\left\|\Delta_{f_j^*}\right\|_2^2\right] + c_4\Gamma^2 N \le \Gamma^2(\beta + c_4)N,$$

where the last inequality holds due to $\Sigma_{ij} \preceq \beta I$. The conclusion (29) follows from (26). ■

Proof [Proof of Corollary 14] We first lower bound ρ_N. By assumption $F_i^\top F_i/n_i \succeq \alpha I_d$, and then

$$\frac{1}{N}\phi(\mathbf{x})^\top\phi(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^M \phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_i) = \frac{1}{N}\sum_{i=1}^M \frac{d}{r_i}U_i F_i^\top F_i U_i^\top \succeq \alpha\sum_{i=1}^M w_i\frac{d}{r_i}U_iU_i^\top. \tag{73}$$

Note that $\mathbb{E}\left[U_iU_i^\top\right] = \frac{r_i}{d}I_d$. Thus,

$$\sum_{i=1}^M w_i\frac{d}{r_i}\mathbb{E}\left[U_iU_i^\top\right] = I_d, \tag{74}$$

Let

$$Y_i = w_i\frac{d}{r_i}\left[U_iU_i^\top - \mathbb{E}\left[U_iU_i^\top\right]\right].$$

Next we use the matrix Bernstein inequality to bound the deviation $\sum_{i=1}^M Y_i$. Note that $\|Y_i\|_{\mathrm{op}} \le 2w_i d/r_i$ and

$$\begin{aligned}
\left\|\sum_{i=1}^M \mathbb{E}\left[Y_i^2\right]\right\|_{\mathrm{op}} &= \left\|\sum_{i=1}^M w_i^2\frac{d^2}{r_i^2}\left(\mathbb{E}\left[U_iU_i^\top\right] - \left(\mathbb{E}\left[U_iU_i^\top\right]\right)^2\right)\right\|_{\mathrm{op}} \\
&= \left\|\sum_{i=1}^M w_i^2\frac{d^2}{r_i^2}\left(\frac{r_i}{d}I_d - \left(\frac{r_i}{d}\right)^2 I_d\right)\right\|_{\mathrm{op}} \\
&\le \sum_{i=1}^M w_i^2\frac{d}{r_i}.
\end{aligned}$$

Therefore, by the matrix Bernstein inequality, with probability at least $1 - d^{-1}$, for a universal constant $c_3 > 0$,

$$\begin{aligned}
\left\|\sum_{i=1}^M Y_i\right\|_{\mathrm{op}} &\le c_3\sqrt{\sum_{i=1}^M w_i^2\frac{d}{r_i}\log d} + c_3\max_{1\le i\le M} w_i\frac{d}{r_i}\log d \\
&\overset{(a)}{\le} c_3\sqrt{\frac{\nu d\log d}{N}} + c_3\frac{\nu d\log d}{N} \\
&\overset{(b)}{\le} \frac{1}{2},
\end{aligned} \tag{75}$$

where (a) holds by definition $\nu = \max_{i\in[M]} n_i/r_i$ and $w_i = n_i/N$; (b) holds by the assumption that $N \geq C\nu d \log d$ for a sufficiently large constant C. Therefore, combining (74) and (75),

$$\frac{1}{N}\phi(\mathbf{x})^\top \phi(\mathbf{x}) \succeq \alpha \sum_{i=1}^{M} w_i \frac{d}{r_i} U_i U_i^\top \succeq \frac{\alpha}{2} I_d.$$

Thus the desired conclusion (30) readily follows from (25) in Theorem 12. The proof of (31) follows similarly from (72) and

$$\sum_{i=1}^{M} \left\| f_i^*(\mathbf{x}_i) - f_j^*(\mathbf{x}_i) \right\|_2^2 \leq \sum_{i=1}^{M} \|\phi(\mathbf{x}_i)\|_2^2 \left\| \theta_i^* - \theta_j^* \right\|_2^2,$$

$$\sum_{i=1}^{M} \|\phi(\mathbf{x}_i)\|_2^2 \leq \sum_{i=1}^{M} \frac{d}{r_i} \|F_i\|_2^2 \leq \beta d \sum_{i=1}^{M} \frac{n_i}{r_i} \leq \beta \nu M d.$$

■

Appendix D. Proofs in Section 6.3

Proof [Proof of Theorem 16] It follows from Corollary 13 that

$$R_j^{\mathsf{Fed}} \lesssim \kappa \left(\sigma^2 d/N + \Gamma^2 \right).$$

Then the desired conclusion readily follows from the following claim:

$$R_j^{\mathsf{Loc}} \gtrsim \min\{\sigma^2 d/n_j, B^2\} + \max\{1 - n_j/d, 0\} B^2.$$

It remains to check the claim. Note that to estimate the model $f_j^* \in \mathcal{H}_B$, it is equivalent to estimating the model coefficient θ_j^* in the ℓ^2 ball of radius B centered at the origin.

For any n_j and d, we bound the minimax risk from below using the celebrated Fano's inequality. Let $\mathcal{V}$ denote a 1/2-packing set of the unit ℓ^2 ball in ℓ^2 norm. By simple volume ratio argument (see e.g. (Wainwright, 2019, Lemma 5.5 and Lemma 5.6), such a set of cardinarlity $|\mathcal{V}| \geq 2^d$ exists. For each $v \in \mathcal{V}$, define $\theta_v = 4\delta v$, where δ will be optimized later. For every pair of $v \neq v'$, $\|\theta_v - \theta_{v'}\|_2 = 4\delta \|v - v'\|_2 \geq 2\delta$. Also, let $\mathcal{P}_v$ denote the distribution of y_j conditional on $\theta_j^* = \theta_v$. Let D_{KL} denote the Kullback–Leibler divergence. Then by the convexity of D_{KL}, we have

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathcal{P}_v \| \mathcal{P}_{v'}\right) &\leq \mathbb{E}_{\mathbf{x}_j}\left[D_{\mathrm{KL}}\left(\mathcal{N}\left(\phi(\mathbf{x}_j)\theta_v, \sigma^2 \mathbf{I}\right) \| \mathcal{N}\left(\phi(\mathbf{x}_j)\theta_v', \sigma^2 \mathbf{I}\right)\right)\right] \\
&= \mathbb{E}_{\mathbf{x}_j}\left[\frac{\|\phi(\mathbf{x}_j)(\theta_v - \theta_v')\|_2^2}{2\sigma^2} \right] \\
&\leq n_j \beta \left\| \theta_v - \theta_v' \right\|_2^2 / (2\sigma^2) \\
&\leq n_j \beta 32 \delta^2 / \sigma^2.
\end{aligned}$$

Therefore,

$$\max_{v,v'} D_{\mathrm{KL}}\left(\mathcal{P}_v \| \mathcal{P}_{v'}\right) \leq 32 n_j \beta \delta^2/\sigma^2.$$

Finally, applying Fano's inequality (see e.g. (Wainwright, 2019, Proposition 15.12)), we get

$$R_j^{\mathsf{Loc}} \geq \delta^2 \left(1 - \frac{32 n_j \beta \delta^2/\sigma^2 + \log 2}{d \log 2}\right).$$

Picking $\delta^2 = \frac{1}{64}\min\{\sigma^2 d/(\beta n_j), B^2\}$. Then by construction, $\|\theta_v\|_2 \leq B$. Further, it follows from the last displayed equation that

$$R_j^{\mathsf{Loc}} \geq c(\beta)\min\{\sigma^2 d/(\beta n_j), B^2\},$$

where $c(\beta)$ is a constant that only depends on β.

When $n_j < d$, we bound the minimax risk from below by assuming θ_j^* is uniformly distributed over the ℓ^2 sphere $\mathcal{S}$ of radius B. Moreover, we use the standard genie-aided argument by assuming that the estimator also has access to $\bar{y}_j \triangleq \phi(\mathbf{x}_j)\theta_j^*$. In this case, the posterior distribution of θ_j^* (conditional on $\{\mathbf{x}_j, y_j, \bar{y}_j\}$) is the uniform distribution over $\mathcal{S}' \triangleq \mathcal{S} \cap \{\theta : \phi(\mathbf{x}_j)\theta = \bar{y}_j\}$. Construct matrix V (resp. $V_\perp$) by choosing its columns as a set of basis vectors in the row (resp. null) space of $\phi(\mathbf{x}_j)$. Then $\mathcal{S}' = \{\theta : \theta = V_\perp \alpha + V\beta\}$, where β is the unique solution such that $\phi(\mathbf{x}_j)V\beta = \bar{y}_j$ and α satisfies $\|\alpha\|_2^2 = B^2 - \|\beta\|_2^2$. Therefore, for any estimator $\widehat{\theta}_j\left(\mathbf{x}_j, y_j, \bar{y}_j\right)$,

$$\mathbb{E}\left[\left\|\widehat{\theta}_j - \theta_j^*\right\|_2^2 \mid \mathbf{x}_j, y_j, \bar{y}_j\right] \geq \inf_{\theta} \mathbb{E}\left[\left\|\theta - \theta_j^*\right\|_2^2 \mid \mathbf{x}_j, y_j, \bar{y}_j\right] = B^2 - \|\beta\|_2^2$$

Taking the average over both hand sides and using $V^\top \theta_j^* = \beta$ so that

$$\mathbb{E}_{\theta_j^*, \mathbf{x}_j, \xi_j}\left[\left\|\widehat{\theta}_j - \theta_j^*\right\|_2^2\right] \geq B^2 - \mathbb{E}\left[\left\|V^\top \theta_j^*\right\|_2^2\right] \geq \left(1 - \frac{n_j}{d}\right) B^2, \quad \forall \widehat{\theta}_j,$$

where the last inequality holds because the rank of V is at most n_j and the prior distribution of θ_j^* is uniform over the sphere $\mathcal{S}$, so that

$$\mathbb{E}\left[\left\|V^\top \theta_j^*\right\|_2^2 \mid \mathbf{x}_j\right] = \left\langle VV^\top, \mathbb{E}\left[\theta_j^* (\theta_j^*)^\top\right]\right\rangle = \frac{B^2}{d}\left\langle VV^\top, \mathbf{I}\right\rangle = \frac{B^2}{d}\|V\|_{\mathrm{F}}^2 \leq \frac{B^2 n_j}{d}.$$

Therefore,

$$R_j^{\mathsf{Loc}} \geq \inf_{\widehat{\theta}_j} \mathbb{E}_{\theta_j^*, \mathbf{x}_j, \xi_j}\left[\left\|\widehat{\theta}_j - \theta_j^*\right\|_2^2\right]] \geq \left(1 - \frac{n_j}{d}\right) B^2.$$

■

Proof [Proof of Theorem 17] It follows from (30) in Corollary 14 that

$$R_j^{\mathsf{Fed}} \lesssim \sigma^2 \kappa d/N.$$

Then the desired conclusion readily follows from the following claim:

$$R_j^{\mathsf{Loc}} \gtrsim \min\{\sigma^2 d/n_j, B^2\} + (1 - r_j/d)B^2.$$

The proof of the claim follows verbatim as that in Theorem 16 with the rank of V being at most r_j, and is omitted for simplicity. ■