Understanding Generalization of Federated Learning via Stability: Heterogeneity Matters

Zhenyu Sun
ECE
Northwestern University

Xiaochun Niu
IEMS
Northwestern University

Ermin Wei
ECE & IEMS
Northwestern University

Abstract

Generalization performance is a key metric in evaluating machine learning models when applied to real-world applications. Good generalization indicates the model can predict unseen data correctly when trained under a limited number of data. Federated learning (FL), which has emerged as a popular distributed learning framework, allows multiple devices or clients to train a shared model without violating privacy requirements. While the existing literature has studied extensively the generalization performances of centralized machine learning algorithms, similar analysis in the federated settings is either absent or with very restrictive assumptions. In this paper, we aim to analyze the generalization performances of federated learning using algorithmic stability, which measures the change of the output model of an algorithm when perturbing one data point. Three widely used algorithms are studied, including FedAvg, SCAFFOLD, and FedProx, under convex and non-convex loss functions. Our analysis shows that the generalization performances of models trained by these three algorithms are closely related to the heterogeneity of clients' datasets as well as the convergence behaviors of the algorithms. Particularly, in the i.i.d. setting, our results recover the classical results of stochastic gradient descent (SGD).

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Federated learning (FL) has recently emerged as an important paradigm for distributed learning in large-scale networks [1]. Unlike the traditional centralized learning, where a model is trained under a large dataset stored at the server [2–4], in federated learning the server hands over computation tasks to the clients, which in turn perform learning algorithms on their local data. After training locally, each client reports its updated model back to the server for model aggregation. The server then aggregates all clients' models to generate a new one that serves as the initialization for the next round of clients' local training. This process is repeated with periodic communication. This local-training framework ensures privacy-preserving and communication-efficient characteristics for federated learning in the sense that no data are transmitted to the server [1].

FedAvg [5], the first proposed algorithm satisfying the federated learning paradigm, implements stochastic gradient descent (SGD) to update local models, which is simple to implement. However, FedAvg suffers from the slow speed of convergence when local data are highly heterogeneous because the local training steps drive the local models away from the global optimal model and towards the local optimal solution [5,6]. This is called client-drift. To mitigate client-drift caused by data heterogeneity, two promising algorithms are proposed. FedProx [7] uses a proximal method such that the trained local model stays relatively close to the global model. Nevertheless, each client has to solve a proximal point optimization problem during each round which could be computationally expensive. Alternatively, SCAFFOLD [8] tries to correct for client-drift based on variance reduction. It is proved that SCAFFOLD outperforms FedAvg when the heterogeneity level of data is large and enjoys a faster convergence speed [8]. Besides, there are several other algorithms proposed later focusing on dealing with client-drift while improving convergence performances [9, 10].

Most existing experimental and theoretical results of FL emphasize convergence to empirical optimal solutions based on training datasets [7, 8, 11, 12] and often ignore their generalization properties. Generalization of FL is important, as it measures the performance of trained models on unseen data by evaluating its testing error. There are only a few existing works studying the generalization properties. Generalization bounds are provided for FL [13–15], which ignores the algorithm choices. These works also require some restrictive assumptions, e.g., binary loss [13, 14], Bernstein condition [15]. In [16, 17], generalization bounds for meta-learning and federated learning are established respectively, when losses are strongly convex and bounded. However, in many practical scenarios, strong convexity does not hold and the loss function may be unbounded. We also note that bounds in [16, 17] are based on uniform stability, which uses a supremum over all single-point perturbations. These tend to be overly conservative compared to a practical alternative notion, *on-average stability*, which takes expectation instead of supremum. Moreover, for the above-mentioned works, the connection between data heterogeneity and generalization performances is not explicitly characterized. Therefore, in this paper, we use on-average stability analysis to obtain generalization bounds that clearly illustrate the dependence of data heterogeneity as well as algorithm convergence speed of three widely-used algorithms: FedAvg, SCAFFOLD, and FedProx. Our bounds are established under general convex and non-convex losses, which can be unbounded.

1.1 Related work

Convergence of federated learning algorithms. Many recent studies are devoted to federated learning problems due to the increasing volume of data among heterogeneous clients and concerns about privacy leakage and communication costs associated with transmitting users' data for central processing [5, 8]. FedAvg applies SGD for local updates of clients and suffers from slow convergence performances when the local datasets across clients are highly heterogeneous [6]. To deal with data heterogeneity and improve convergence speed, FedProx adopts proximal methods for local training [7] and has both convergence guarantees and improved numerical results. SCAFFOLD [8] borrows the idea from variance reduction methods [18] and shows that convergence rates can be highly improved, compared to FedAvg and FedProx. In [19], the effects of heterogeneous objectives on solution bias and convergence slowdown are systematically investigated, and FedNova is proposed to preserve fast convergence. FedPD [20] views federated learning from the primal-dual perspective. In [21], FedLin is aimed to deal with data heterogeneity and system heterogeneity of clients simultaneously. More related works are given therein [22–25].

Generalization of centralized and federated learning. Generalization of centralized learning has gained attraction of researchers since several decades ago. Uniform convergence is commonly considered to bound the generalization error by means of VC dimension or Rademacher complexity [26–29]. However, uniform convergence sometimes renders the bound too loose to be meaningful [30] and from technical perspectives, a finite VC dimension or Rademacher complexity is required to obtain a finite bound, which might be avoided for modern neural-network models. The main limitation of uniform convergence analysis is that it only studies the properties of model classes but ignores training algorithms that generate the models. Taking training algorithms into consideration, the generalization bounds might be tightened, since we can directly ignore a large amount of models which can never be the output of a specific algorithm. Algorithmic stability is a useful notion that specifically helps to investigate generalization errors by considering dependency on particular algorithms [31], which in the meantime makes the analysis for extremely complex model classes (e.g. neural networks) possible. Generalization bounds are built for several stochastic gradient-based methods via algorithmic stability tools [32]- [36]. Several other works on the generalization of centralized learning from theoretical and experimental aspects are listed therein [37]- [39]. In terms of federated learning, [13] provides uniform convergence bound with rate $\mathcal{O}(1/\sqrt{n})$ for agnostic federated learning problems under binary losses, and n is the number of samples collected by all clients. The works [40, 41] capture the heterogeneity effect on generalization with the same rate, while they require a common feature extraction function across clients and the asymptotic convergence cannot be achieved, which means the bounds cannot approach zero as n goes to infinity. This is problematic since one should expect the generalization bound to vanish if there are infinitely many samples. [15] studies the case when some clients do not participate during the training phase and establish bounds with the faster rate $\mathcal{O}(1/n)$ under Bernstein condition and bounded losses. It further requires that clients' distributions are sampled from a meta-distribution, which may be impractical. [16, 17] provide generalization bounds via uniform stability, obtaining rates $\mathcal{O}(1/n)$. Further, [16] requires there is only a one-step local update which does not match the common practice of using multiple local updates in a federated setting. Note these works require bounded and strongly convex loss functions, which are quite restrictive in practice. Some other bounds of federated learning are information-theoretic-

based. However, these bounds are limited by requiring specific forms of loss functions and also fail to capture heterogeneity effect [42, 43]. A comparison of our results to the existing ones is listed in Table 1.

1.2 Our contributions

We summarize our main contributions as follows: (1) We propose a bound on generalization error by using algorithm-dependent on-average stability in federated settings (see Section 3); (2) Based on on-average stability, we provide generalization upper bounds for FedAvg, SCAFFOLD and FedProx respectively with unbounded, convex and non-convex loss functions, which explicitly reveal the effects of data heterogeneity and convergence performances of different algorithms (see Sections 5.1 and 5.2); (3) In i.i.d. setting with convex loss functions, our bounds match existing results of SGD in the sense that FedAvg reduces to SGD method (see Section 5.1); (4) Experimental results are provided, demonstrating the trends in our theoretical bounds (see Section 6).

2 PROBLEM FORMULATION

In this paper, we consider the general federated learning problem, where m clients collaboratively minimize the following global population risk formed by

$$R(\theta) := \sum_{i=1}^{m} p_i \mathbb{E}_{z\sim P_i}[l(\theta; z)], \tag{1}$$

where $\theta \in \mathbb{R}^d$ is the parametrized model and P_i is the underlying distribution of the dataset maintained by client i. We adopt the standard assumption that samples taken from P_i and P_j are independent for any $i, j \in [m]$ such that $i \neq j$, motivated by the observation that local data of clients are commonly unrelated in practical scenarios. We define z as the sample generated by P_i, i.e., $z \sim P_i$, $l(\cdot; z)$ as the loss function evaluated at sample z, and p_i as some constant scalar that measures the contribution of client i's data to the global objective. We also define the local population risk as $R_i(\theta) := \mathbb{E}_{z\sim P_i}[l(\theta; z)]$.

However, in practice, we are unable to minimize the global population risk directly due to the unknown distributions P_i. Thus, one alternative way to get an approximate model is by collecting some empirical sample dataset $\mathcal{S}_i$. More specifically, each local dataset is defined by $\mathcal{S}_i := \{z_{i,j}\}_{j=1}^{n_i}$, where $z_{i,j}$ is the j-th sample of client i and n_i is the number of local samples. Let $\mathcal{S} := \bigcup_{i=1}^{m} \mathcal{S}_i$ be the dataset with all samples and n be the total amount of samples with $n = \sum_{i=1}^{m} n_i$. Moreover, we are interested in the *balanced case*, i.e., $p_i = n_i/n$, meaning the contribution of each client to the global objective is proportional to the local sample size n_i. Thus, we turn to train a model by minimizing the following global empirical risk:

$$\hat{R}_{\mathcal{S}}(\theta) := \sum_{i=1}^{m} p_i \hat{R}_{\mathcal{S}_i}(\theta) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} l(\theta; z_{i,j}), \tag{2}$$

where we use the fact that $p_i = n_i/n$ and $\hat{R}_{\mathcal{S}_i}(\theta)$ is the local empirical risk $\hat{R}_{\mathcal{S}_i}(\theta) := \frac{1}{n_i} \sum_{j=1}^{n_i} l(\theta; z_{i,j})$. Here we use superscript notation $\hat{R}$ to indicate the empirical version of R and will use superscript in a similar fashion for the rest of the paper. Based on the above definitions, we further define the ground-truth model θ^* by minimizing the population risk (1), that is, $\theta^* \in \arg\min_\theta R(\theta)$ and correspondingly the best empirically trained model is defined by $\hat{\theta}_{\mathcal{S}} \in \arg\min_\theta \hat{R}_{\mathcal{S}}(\theta)$.

Our ultimate goal is to obtain the ground-truth model θ^*, which is impossible due to unknown distributions. What we can do practically is to solve for $\hat{\theta}_{\mathcal{S}}$ by implementing appropriate optimization algorithms such that (2) is minimized. Then, a natural question is how we could expect the trained model $\hat{\theta}_{\mathcal{S}}$ to be close to θ^*. Alternatively, we want to test model $\hat{\theta}_{\mathcal{S}}$ on any unseen data such that the testing error is small enough, which means the model $\hat{\theta}_{\mathcal{S}}$ generalizes well on any testing set.

In general, even given good datasets, exactly obtaining $\hat{\theta}_{\mathcal{S}}$ is still a hard optimization problem. A more reasonable approach is to implement some algorithm $\mathcal{A}$ which outputs a model $\mathcal{A}(\mathcal{S})$, noting the model is a function of the training set $\mathcal{S}$.

3 GENERALIZATION & STABILITY

As stated in the previous section, we now focus on the generalization performance of the output of some algorithm $\mathcal{A}(\mathcal{S})$, given a training dataset $\mathcal{S}$. Mathematically, the generalization error of a model $\mathcal{A}(\mathcal{S})$ is defined by

$$\epsilon_{gen} := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}[R(\mathcal{A}(\mathcal{S})) - \hat{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))],$$

where the expectation is taken over $\mathcal{S}$ to model the random sampling of data and over $\mathcal{A}$ to allow the usage of randomized algorithms. For instance, if the stochastic gradient is used in an algorithm then the expectation over $\mathcal{A}$ is average over different samples used to compute the stochastic gradients. A smaller ϵ_{gen} implies the model $\mathcal{A}(\mathcal{S})$ has a better generalization performance on testing datasets.

Generally speaking, it is hard to characterize the generalization error due to the implicit dependency of the

Table 1: Generalization bounds for federated learning. C, SC, and NC denote convex, strongly convex, and non-convex, respectively. The last column represents connections of bounds to data heterogeneity.

Reference	Loss Function	Complexity	Data Heterogeneity
[13, 40, 41]	Binary [13]; Common Feature Extraction [40, 41]	$\mathcal{O}(1/\sqrt{n})$	Yes
[42, 43]	sub-Gaussian	$\mathcal{O}(1/\sqrt{n})$	No
[15]	Bounded, Bernstein Con	$\mathcal{O}(1/n)$	No
[16, 17]	Bounded, SC, Smooth	$\mathcal{O}(1/n)$	No
Ours	Unbounded, C, Smooth	$\mathcal{O}(1/n)$	Yes
Ours	Unbounded, NC, Smooth	$\mathcal{O}(1/n)$	Yes

model and the training dataset. In this paper, we apply the notion of algorithmic stability to provide an upper bound on the generalization error. In particular, we formally define the on-average stability in the context of federated learning. To do this, we first introduce the definition of neighboring datasets.

Definition 1. *Given a global dataset $\mathcal{S} = \bigcup_{l=1}^{m} \mathcal{S}_l$, where $\mathcal{S}_l$ is the local dataset of the l-th client with $\mathcal{S}_l = \{z_{l,1}, \ldots, z_{l,n_l}\}, \forall l \in [m]$, another global dataset is said to be neighboring to $\mathcal{S}$ for client i, denoted by $\mathcal{S}^{(i)}$, if $\mathcal{S}^{(i)} := \bigcup_{l \neq i} \mathcal{S}_l \cup \mathcal{S}_i'$, where $\mathcal{S}_i' = \{z_{i,1}, \ldots, z_{i,j-1}, z_{i,j}', z_{i,j+1}, \ldots, z_{i,n_i}\}$ with $z_{i,j}' \sim P_i$, for some $j \in [n_i]$. And we call $z_{i,j}'$ the perturbed sample in $\mathcal{S}^{(i)}$.*

In other words, $\mathcal{S}$ and $\mathcal{S}^(i)$ are neighboring datasets if they only differ by one data point in $\mathcal{S}_i$ and both are sampled from the same local distribution. Then, we have the following definition of on-average stability for federated learning algorithms, which is established based on on-average stability for centralized learning [26].

Definition 2 (on-average stability for federated learning)**.** *A federated learning algorithm $\mathcal{A}$ is said to have ϵ-on-average stability if given any two neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, then for any $i \in [m]$*

$$\max_{j \in [n_i]} \mathbb{E}_{\mathcal{A}, \mathcal{S}, z_{i,j}'} |l(\mathcal{A}(\mathcal{S}); z_{i,j}') - l(\mathcal{A}(\mathcal{S}^{(i)}); z_{i,j}')| \leq \epsilon,$$

where $z_{i,j}'$ is the perturbed sample in $\mathcal{S}^{(i)}$.

On-average stability basically means any perturbation of samples across all clients cannot lead to a big change in the model trained by the algorithm in expectation. The next theorem shows that on-average stability can be used to bound the generalization error of the model. The proof is given in Appendix B.

Theorem 1. *Suppose that a federated learning algorithm $\mathcal{A}$ is ϵ-on-averagely stable. Then,*

$$\epsilon_{gen} \leq \mathbb{E}_{\mathcal{A}, \mathcal{S}} \left[\left| R(\mathcal{A}(\mathcal{S})) - \hat{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) \right| \right] \leq \epsilon.$$

Therefore, it suffices to characterize the on-average stability of a federated learning algorithm to bound the generalization error of the model. Theorem 1 extends the classical connection of on-average stability and generalization [26], where no heterogeneity characteristic of datasets is considered. Based on Definition 2, we show that when the perturbation of a sample for any local agent has a small influence on algorithm output (i.e., a small ϵ), the generalization error is also small (i.e., ϵ_{gen} is small). This relationship always holds given any clients' local data distributions. Then, in the following, we focus on analyzing the stability of different federated learning algorithms and applying their stability results to the measure of generalization.

4 SUMMARY OF FEDERATED LEARNING ALGORITHMS

In this section, we briefly summarize three widely-used federated learning algorithms: FedAvg, SCAFFOLD, and FedProx, based on which the generalization bounds would be provided. To simplify the analysis, we assume that there is no partial participation among the clients, but our analysis can be extended to partial participation scenarios as well.

Any federated algorithms can be decomposed into two stages: local updating and model aggregation. At the beginning of each communication round (time index t), the server maintains a global model θ_t, which is sent to all clients serving as an initial model of local updating. All clients update their local models θ_{t+1}^i based on their own datasets in parallel. Then, a model aggregation for the start of the next round, i.e., $\theta_{t+1} = \sum_{i=1}^{m} p_i \theta_{t+1}^i$. The three methods only differ in their local updating procedures and the detailed descriptions of algorithms can be found in Appendix A.

For FedAvg and SCAFFOLD, in the t-th communication round, we assume that there are K_i local updates and denote by $\theta_{i,k}$ and $g_i(\cdot)$ the local model at local

iteration k and the sampled gradient of agent i. For FedProx, we let θ_{t+1}^i be the model of client i after local training at round t. Then, for client i, the local updates at iteration k are described as follows.

- FedAvg: Let $\alpha_{i,k}$ be the constant (or diminishing) local stepsize, agent i's local update is for any $k = 0, \dots, K_i - 1$

$$\theta_{i,k+1} = \theta_{i,k} - \alpha_{i,k} g_i(\theta_{i,k}). \tag{3}$$

- SCAFFOLD: Let $\alpha_{i,k}$ be the constant (or diminishing) local stepsize, agent i's local update is for any $k = 0, \dots, K_i - 1$

$$\theta_{i,k+1} = \theta_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t) + g(\theta_t)), \tag{4}$$

  where $g(\theta_t) = \sum_{i=1}^m p_i g_i(\theta_t)$ is the aggregation of all locally sampled gradients.

- FedProx: Let η_i be a constant parameter for the proximal term, agent i's local update is

$$\theta_{t+1}^i = \arg\min_{\theta} \ \hat{R}_{\mathcal{S}_i}(\theta) + \frac{1}{2\eta_i}\|\theta - \theta_t\|^2. \tag{5}$$

5 MAIN RESULTS

In this section, we provide bounds on the generalization errors for FedAvg, SCAFFOLD, and FedProx mentioned in the last section by studying the on-average stability in Definition 2. We allow the loss functions to be unbounded from above, which can be convex or nonconvex. Intuitively, different local distributions affect the global population risk (1) and hence may affect the model generalization as well. To measure the heterogeneity of client i's data, we use D_i to denote the total variation of P_i and P, i.e., $D_i := d_{TV}(P_i, P)$ with $P = \sum_{i=1}^m p_i P_i$. Moreover, we define $D_m := \max_{i\in[m]} D_i$ to measure the furthest distance between the global distribution and any local distribution[1]. A larger value of D_m means greater heterogeneity among the clients. Throughout the analysis we require the following assumptions.

Assumption 1. *The loss function $l(\cdot, z)$ is L-Lipschitz continous for any sample z, that is, $|l(\theta; z) - l(\theta'; z)| \le L\|\theta - \theta'\|$, for any z, θ, θ'.*

Assumption 2. *There exists a constant $\sigma > 0$ such that for any θ, $i \in [m]$, and $z_i \sim P_i$, $\mathbb{E}\left[\|\nabla l(\theta; z_i) - \nabla R_i(\theta)\|^2\right] \le \sigma^2$.*

Assumption 3. *The loss function $l(\cdot, z)$ is β-smooth for any z, that is, $\|\nabla l(\theta; z) - \nabla l(\theta'; z)\| \le \beta\|\theta - \theta'\|$, for any z, θ, θ'.*

[1]Our bounds can be derived under KL divergence as well. However, bounds involving total variation are tighter.

Assumption 1 is standard in literature [32] to establish the connection between model perturbation with stability [2]. Assumptions 2 and 3 serve in our analysis to capture the heterogeneity of different datasets as well as the influence of convergence performances of different algorithms. Detailed proofs of this section are in Appendices C and D.

5.1 Convex loss functions

We first study the case when the loss function is convex with respect to the model parameter.

Assumption 4. *The loss function $l(\cdot, z)$ is convex for any z.*

For each of the three algorithms, FedAvg, SCAFFOLD and FedProx with local updates (3), (4), (5), we apply the method to two neighboring training datasets, i.e., only one data point of one agent is different. We then analyze and bound the difference between the resulting models by data heterogeneity and algorithm performances.

Theorem 2. *Under Assumptions 1-4, denote $\{\theta_t\}_{t=0}^T$ and $\{\theta_t'\}_{t=0}^T$ as the trajectories of the server's models induced by neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Furthermore, suppose the same initialization, i.e., $\theta_0 = \theta_0'$. Then by denoting $\delta_t = \mathbb{E}\|\theta_t - \theta_t'\|$, we have the following bounds on resulting models.*

For FedAvg,

$$\delta_T \le \frac{2}{n}\sum_{t=0}^{T-1} h_t^{\text{avg}}\Big(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma\Big).$$

For SCAFFOLD,

$$\delta_T \le \frac{2}{n}\sum_{t=0}^{T-1} h_t^{\text{scfd}}\Big(2LD_i\gamma_t^1 + \gamma_t^2(\mathbb{E}\|\nabla R(\theta_t)\| + \sigma)\Big).$$

For FedProx,

$$\delta_T \le \frac{2}{n}\sum_{t=0}^{T-1} h_t^{\text{prox}}\Big(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma\Big),$$

where $\gamma_t^1 := 2\tilde{\alpha}_{i,t} + \hat{\alpha}_t$ and $\gamma_t^2 := \gamma_t^1 + \beta\tilde{\alpha}_{i,t}^2$ with $\tilde{\alpha}_{i,t} := \sum_{k=0}^{K_i-1}\alpha_{i,k}$, $\hat{\alpha}_t := \sum_{j=1}^m p_j\tilde{\alpha}_{j,t}$. We also define $\sum_{l=T}^{T-1}\hat{\alpha}_l = 0, \forall\hat{\alpha}_l$. And $h_t^{\text{avg}} = \tilde{\alpha}_{i,t}(1 + \beta\tilde{\alpha}_{i,t})$, $h_t^{\text{scfd}} = \exp\left(2\beta\sum_{l=t+1}^{T-1}\hat{\alpha}_l\right)$, $h_t^{\text{prox}} = \eta_i(1 + \beta\eta_i)$. The expectations are taken with respect to $\mathcal{S}$ and $\mathcal{S}^{(i)}$ jointly as well as the randomness of algorithms.

[2]Note that we only need Assumption 1 holds along the trajectories generated by the algorithms.

Next we discuss the implications of Theorem 2. Firstly, the model differences δ_T of the three algorithms all linearly increase in D_i. Recall that D_i is the total variation of data distribution of client i and the global distribution P, measuring the heterogeneity level of client i's data. This dependency is due to the fact that we only perturb one data point of client i while keeping the others the same and hence only client i's distribution comes into the bound. As D_i increases, perturbing one data point at client i's dataset corresponds to a bigger change in the overall dataset and therefore the distance between the two models increases.

Secondly, the sequence of global gradients evaluated along the trajectories, i.e., $\{\mathbb{E}\|\nabla R(\theta_t)\|\}_{t=0}^{T-1}$, influences the bounds of model differences. Note that this effect is essentially determined by the convergence performances of algorithms, in the sense that $\{\mathbb{E}\|\nabla R(\theta_t)\|\}_{t=0}^{T-1}$ captures how fast $\{\theta_t\}_{t=0}^{T-1}$ approaches to the optimal solution θ^*. Faster converging methods correspond to smaller $\{\mathbb{E}\|\nabla R(\theta_t)\|\}_{t=0}^{T-1}$ terms.

Thirdly, the bounds are also proportional to the sampling variance σ^2 of gradients. A small σ indicates the sampled gradient is accurate and is close to the true gradient $\nabla R(\cdot)$. In particular, when $\sigma = 0$, each client is able to compute $\nabla R_i(\cdot)$ exactly, in which case the bounds are only related to data heterogeneity and algorithm convergence performances.

Finally, all three bounds depend on the stepsizes chosen during the local training process. Different choices of stepsizes result in different convergence rates of algorithms. From the above results, larger stepsizes may make algorithms less "stable", i.e., $\|\theta_T - \theta'_T\|$ becomes bigger, as any difference caused by the perturbed data is magnified by the stepsize.

As we discussed above, the summation of $\mathbb{E}\|\nabla R(\theta_t)\|$ is related to the convergence speed of the algorithm. In the following theorem, we focus on characterizing these terms as a function of the number of iterations. Define $\tilde{D} := \sum_{i=1}^m p_i D_i^2$ and $\Delta_0 = \mathbb{E}[R(\theta_0) - R(\theta^*)]$.

Theorem 3. *Under Assumptions 1-3, suppose $K_i = K$ and $\alpha_{i,k} \leq 1/(24\beta K)$ for any $i = 1, \ldots, m$. Then, for FedAvg, we have*

$$\sum_{t=0}^{T-1} h_t^{\text{avg}} \mathbb{E}\|\nabla R(\theta_t)\| = \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} T^{\frac{3}{4}} + \big(\Delta_0^2 \tilde{D}\big)^{\frac{1}{6}} T^{\frac{2}{3}} + \sqrt{\Delta_0} T^{\frac{1}{2}} \Big).$$

For SCAFFOLD, if we further set $\alpha_{i,k} \leq 1/[24\beta K(t+1)]$, then

$$\sum_{t=0}^{T-1} h_t^{\text{scfd}} \gamma_t^2 \mathbb{E}\|\nabla R(\theta_t)\| = \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} T^{\frac{5}{6}} + \sqrt{\Delta_0} T^{\frac{7}{12}} \Big).$$

For FedProx, we have

$$\sum_{t=0}^{T-1} h_t^{\text{prox}} \mathbb{E}\|\nabla R(\theta_t)\| = \mathcal{O}\Big(\big(\Delta_0 \tilde{D}\big)^{\frac{1}{2}} T^{\frac{3}{4}} + \sqrt{\Delta_0} T^{\frac{1}{2}} \Big).$$

Theorem 3 bounds the global gradients $\|\nabla R(\theta_t)\|$ along the trajectories of server's outputs. This theorem holds for both convex and non-convex settings. Under suitable selections of stepsizes, Theorem 3 implies that the global gradient $\mathbb{E}\|\nabla R(\theta_t)\|$ converges to zero. This is consistent with convergence results in the optimization perspective [7, 8]. To see this, when dividing the preceding bounds by T, the right-hand side converges to zero in polynomial times, and hence $\mathbb{E}\|\nabla R(\theta_t)\|$ must converge to zero. Moreover, these bounds increase with Δ_0, which measures the distance of the initial model to the optimal one. Thus, starting at a model closer to the optimal solution requires less number of iterations to approximate accurately θ^*.

By combining Theorems 1-3, we establish the generalization bounds for three algorithms, respectively.

Corollary 1. *Suppose Assumptions 1-4 hold and the selection of stepsizes are the same as Theorem 3. Then, we have the following generalization bounds: For FedAvg,*

$$\epsilon_{gen} \leq \mathcal{O}\Big(\frac{T}{n} D_m\Big) + \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} \frac{T^{\frac{3}{4}}}{n} + \big(\Delta_0^2 \tilde{D}\big)^{\frac{1}{6}} \frac{T^{\frac{2}{3}}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{1}{2}}}{n} \Big) + \mathcal{O}\Big(\frac{\sigma T}{n}\Big).$$

For SCAFFOLD,

$$\epsilon_{gen} \leq \tilde{\mathcal{O}}\Big(\frac{T^{\frac{1}{12}}}{n} D_m\Big) + \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} \frac{T^{\frac{5}{6}}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{7}{12}}}{n} \Big) + \mathcal{O}\Big(\frac{T^{\frac{1}{12}}(1+\log T)}{n} \sigma\Big).$$

For FedProx,

$$\epsilon_{gen} \leq \mathcal{O}\Big(\frac{T}{n} D_m\Big) + \mathcal{O}\Big(\big(\Delta_0 \tilde{D}\big)^{\frac{1}{2}} \frac{T^{\frac{3}{4}}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{1}{2}}}{n} \Big) + \mathcal{O}\Big(\frac{T}{n} \sigma\Big).$$

As indicated in Theorem 2, the generalization bound for each algorithm can be separated into three terms corresponding to the three $\mathcal{O}(\cdot)$ terms: heterogeneity level (first), convergence performance (second), sampling variance (third). Note D_m in the first term measures data heterogeneity among all agents. A smaller D_m indicates clients have more similar datasets, which has a positive effect on the generalization of trained models. Moreover, generalization bounds above scale inversely with n, which is the total sample size. This

implies increasing the number of samples gives a better generalization performance. Note that fixing T, the rate is with the order of $\mathcal{O}(1/n)$, which is the same as results in the centralized setting [32, 33, 44].

Furthermore, when assuming all clients maintain i.i.d. data and applying the Lipschitz continuity condition to bound the gradient, i.e., $\|\nabla R(\cdot)\| \leq L$, we have the following bounds under suitable choices of stepsizes.

Corollary 2. *We suppose Assumptions 1-4 hold and all clients have i.i.d. datasets, that is, $P_i = P_j$, for any $i, j \in [m]$. Then for FedAvg and FedProx, if stepsizes are chosen to be constant, we have $\epsilon_{gen} \leq \mathcal{O}(L(L+\sigma)T/n)$. For SCAFFOLD, if stepsizes are chosen with the order of $\mathcal{O}(c/(2\beta t))$, we have $\epsilon_{gen} \leq \mathcal{O}(L(L+\sigma)T^c \log T/n)$.*

From Corollary 2, we observe that the heterogeneity term disappears because $D_m = 0$ in i.i.d. settings. If σ is relatively small compared to L, for FedAvg, our result is aligned with the bound for SGD [32]. The reason is that for i.i.d. case $R_i(\cdot) = R(\cdot)$ and thus the server's model essentially performs SGD (in expectation). For SCAFFOLD, the update reduces to SAGA [18]. Therefore, the generalization bound for SAGA is implied by Corollary 2.

If we set $D_m = 0$ and choose comparable stepsizes, the bounds of Corollary 1 are tighter than those of Corollary 2. The main reason is that using Lipschitz constant L to bound the gradient is usually too loose and the algorithm performances are highly ignored, which, however, should be carefully considered in the analysis. In particular, considering FedAvg with $m = 1$ and $K = 1$, which is then equivalent to the classical SGD method, our result is better than the result provided in [32] when stepsizes are constants, where the order of T reduces from $\mathcal{O}(T)$ to $\mathcal{O}(T^{3/4})$.

5.2 Non-convex losses

In many practical scenarios, the loss functions are non-convex (e.g., neural networks). Therefore, we provide generalization bounds for non-convex losses in this subsection.

Theorem 4. *Under Assumptions 1-3, suppose $K_i = K$ and $\alpha_{i,k} \leq \frac{1}{24\beta K(t+1)}$ for FedAvg and SCAFFOLD. Then for FedAvg, we have*

$$\epsilon_{gen} \leq \tilde{\mathcal{O}}\Big(\frac{T^{\frac{1}{24}}}{n} D_m\Big) + \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} \frac{T^{\frac{5}{6}}}{n} + \big(\Delta_0^2 \tilde{D}\big)^{\frac{1}{6}} \frac{T^{\frac{3}{4}}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{7}{12}}}{n}\Big) + \tilde{\mathcal{O}}\Big(\frac{T^{\frac{1}{24}}}{n}\sigma\Big).$$

For SCAFFOLD,

$$\epsilon_{gen} \leq \tilde{\mathcal{O}}\Big(\frac{T^{\frac{1}{8}}}{n} D_m\Big) + \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}} \frac{T^{\frac{7}{8}}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{5}{8}}}{n}\Big) + \mathcal{O}\Big(\frac{T^{\frac{1}{8}}(\log T + 1)}{n}\sigma\Big).$$

For FedProx, if the eigenvalues of $\nabla^2 R_i(\theta)$ are lower bounded and η_i is chosen small enough and diminishing with order $\mathcal{O}(c/t)$, then

$$\epsilon_{gen} \leq \tilde{\mathcal{O}}\Big(\frac{T^c}{n} D_m\Big) + \mathcal{O}\Big(\big(\Delta_0 \tilde{D}\big)^{\frac{1}{2}} \frac{T^{\frac{3}{4}+c}}{n} + \sqrt{\Delta_0} \frac{T^{\frac{1}{2}+c}}{n}\Big) + \tilde{\mathcal{O}}\Big(\sigma \frac{T^c}{n}\Big). \tag{6}$$

In Theorem 4, the bounds are similar to those of convex cases, i.e., data heterogeneity, algorithm convergence, and sampled variance jointly affect the generalization error of the models. In addition, we remark that in practice T is usually characterized by a function of n and m. Then in this sense, the generalization bounds can be further simplified in terms of the total sample size n and the number of clients m. For example, considering (6) with one full pass local training, i.e., $T = \mathcal{O}(n/m)$, we obtain $\epsilon_{gen} \leq \tilde{\mathcal{O}}(m^{-c} n^{-(1-c)} + m^{-3/4-c} n^{-(1/4-c)})$, meaning the generalization error diminishes as the number of clients participating in the learning process increases.

However, we remark that only upper bounds on generalization errors are provided, and some constants are ignored in our $\mathcal{O}(\cdot)$ notation. In reality, these ignored constants (e.g. stepsizes) could largely affect algorithms' generalization performances. Thus, our bounds might not be tight enough to explain accurately the performances of algorithms. In addition, for different algorithms, the selection of stepsizes is usually a tricky task, meaning optimal stepsizes for different algorithms are chosen during the training process. This implies, in general, it is hard to compare generalization errors among different algorithms by directly analyzing our bounds. Instead, the main insight shown by our results is that explicit dependency of data heterogeneity to generalization is clearly characterized through total variation among local distributions. This is a first step towards this direction, as existing literature [15–17] fails to characterize the connection of data heterogeneity to generalization bounds.

6 EXPERIMENTS

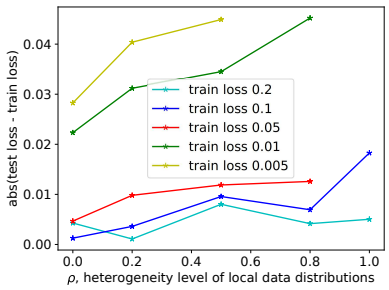
In this section, we numerically evaluate the generalization errors of models trained by FedAvg, SCAFFOLD and FedProx under non-convex loss functions, given different heterogeneity levels of clients' datasets.

Experimental Setups. We investigate classification problems using the MNIST dataset [45]. Each client maintains a three-layer neural network comprising two convolutional layers and a fully connected layer. We focus on a federated learning system involving 10 clients. The training is based on Personalized Federated Platform [46].

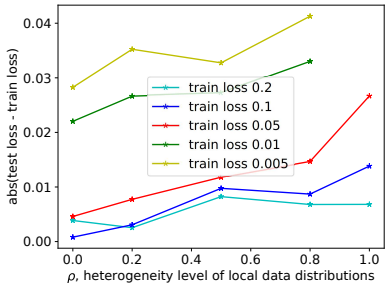
Next, we elaborate on how we construct different clients' datasets with different heterogeneity levels. We introduce various levels of heterogeneity among the clients' local data distributions to examine the impact of heterogeneity on the generalization performance. In the case of extreme heterogeneity, labeled "fully non-i.i.d." scenario, each client has two specific labels out of the ten available MNIST. In the "i.i.d." scenario, the local datasets are uniformly mixed with all ten labels. We also consider the intermediate scenarios labeled "ρ non-i.i.d.", where a fraction ρ of data follows the "fully non-i.i.d." assignment, while the remaining fraction $1-\rho$ adheres to the "i.i.d." assignment. We call ρ the heterogeneity level of local data distributions and run the experiments for $\rho = 0, 0.2, 0.5, 0.8, 1$ cases (5 in total), where $\rho = 0$ and $\rho = 1$ are the "i.i.d." and "fully non-i.i.d." cases, respectively.

In different settings, we start the algorithms from the same initial value with the same training loss. As the training goes on, the training losses decrease. We compare trained models under different levels of training losses. To quantify the generalization errors, we use the absolute difference between the training and testing losses, i.e., $|R(\mathcal{A}(S)) - \hat{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))|$. We terminate the algorithms when either the training loss reaches a desirable level or the number of training steps achieves $T = 1000$.

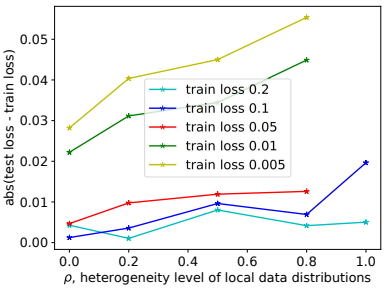
Numerical Results. The generalization errors of FedAvg, SCAFFOLD, and FedProx are shown in Fig. 1. The x-axis shows the heterogeneity level of local data distributions (ρ) and the y-axis shows the generalization errors of the algorithms. We note that the algorithms in some heterogeneous cases ($\rho = 0.8$ or $\rho = 1$) did not achieve some levels of training losses (e.g. 0.005) before they terminated. So there are less than 5 points in the corresponding training loss curves. The figure shows that the generalization error increases as data heterogeneity increases, which is aligned with our theoretical results. Moreover, vertically, the generalization error also increases as the training loss level decreases. Noting that a smaller training loss generally needs more iterations in the training process. Hence these numerical results are also consistent with our bounds, which implies the generalization errors increase as T gets bigger. We provide additional experiments under different approaches of generating non-i.i.d. data in Appendix E, which also demonstrate our theoretical results.



(a) FedAvg



(b) SCAFFOLD



(c) FedProx

Figure 1: Generalization errors of FedAvg, SCAFFOLD, and FedProx

7 CONCLUSION

In this paper, we provide generalization upper bounds for FedAvg, SCAFFOLD, and FedProx by means of on-average stability under both convex and non-convex loss functions. Our bounds explicitly capture the effect of data heterogeneity and algorithm convergence properties on the generalization performances of different algorithms, which indicates that the heterogeneity level of datasets is highly related to the generalization of FL. In particular, under the i.i.d. case, FedAvg reduces to the SGD method and our results are shown to be consistent with those of SGD methods. Our

numerical simulations demonstrate the theoretical results. For future studies, how to evaluate model generalization to slowly time-vary data distributions or some new distribution is a potential topic.

8 ACKNOWLEDGEMENTS

This work was supported by NSF ECCS-2030251, CMMI-2024774 and ECCS-2216970.

References

[1] Peter Kairouz et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14: 1-210, 2019.

[2] Jeffrey Dean et al. Large scale distributed deep networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[3] R.H. Byrd, S.L. Hansen, J. Nocedal, Y.Singer. A Stochastic Quasi-Newton Method for Large-Scale Optimization. *SIAM on Optimization*, 26(2): 1008-1031, 2016.

[4] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221): 1-51, 2018.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

[6] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, Vikas Chandra. Federated learning with non-i.i.d. data. *arXiv preprint arXiv:1806.00582*, 2018.

[7] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2, (2020): 429-450.

[8] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. *International Conference on Machine Learning*, PMLR 119:5132–5143, 2020.

[9] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*. 33, 2020.

[10] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. *Proceedings of the 38th International Conference on Machine Learning*. PMLR 139:12253-12266, 2021.

[11] Hong-You Chen and Wei-Lun Chao. FedBE: Making Bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2021.

[12] Xiaotong Yuan and Ping Li. On convergence of FedProx: Local dissimilarity invariant Bounds, non-smoothness and beyond. *Advances in Neural Information Processing Systems*. 35, 2022.

[13] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. *International Conference on Machine Learning*, PMLR pp. 4615-4625, 2019.

[14] Zhenyu Sun and Ermin Wei. A Communication-efficient Algorithm with Linear Convergence for Federated Minimax Learning. *Advances in Neural Information Processing Systems*. 35, 2022.

[15] Xiaolin Hu, Shaojie Li,and Yong Liu. Generalization bounds for federated learning: fast rates, unparticipating clients and unbounded losses. *International Conference on Learning Representation*, 2023.

[16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*. 34, 2021.

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

[18] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

[19] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. FedPD: a federated learning framework with optimal rates and adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69: 6055-6070, 2021.

[21] 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, 2021.

[22] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. FedDANE: A federated newton-type method. *53rd Asilomar Conference on Signals, Systems, and Computers*. pp. 1227-1231, 2019.

[23] Reese Pathak and Martin J. Wainwright. FedSplit: an algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. *arXiv preprint arXiv:2010.05273*, 2021.

[25] Xiaochun Niu and Ermin Wei. FedHybrid: A hybrid federated optimization method for heterogeneous clients. *IEEE Transactions on Signal Processing*, 71:150-163, 2023.

[26] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11(90): 2635-2670, 2010.

[27] Mehryar Mohri,Afshin Rostamizadeh,and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

[28] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *International Conference on Machine Learning*, PMLR 97:7085-7094, 2019.

[29] Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. *International Conference on Algorithmic Learning Theory*, 98:162-183. PMLR, 2019.

[30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107-115, 2021.

[31] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research* 2: 499-526, 2002.

[32] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *International Conference on Machine Learning*, PMLR 48:1225-1234, 2016.

[33] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. *International Conference on Machine Learning*, PMLR 80:2815-2824, 2018.

[34] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. *International Conference on Machine Learning*, PMLR 119:5809-5819, 2020.

[35] Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research* 21(34): 1-44, 2020.

[36] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. *Conference On Learning Theory*, PMLR 75:605-638, 2018.

[37] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[38] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. *International Conference on Machine Learning*, PMLR 139:3964-3975, 2021.

[39] Hao Wang, Rui Gao, and Flavio P. Calmon. Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *Journal of Machine Learning Research*, 24: 1-43, 2023.

[40] Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351-2363, 2020.

[41] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. *International Conference on Machine Learning*, PMLR 139:12878-12889, 2021.

[42] L. P. Barnes, Alex Dytso, and H. V. Poor. Improved information theoretic generalization bounds for distributed and federated learning. *IEEE International Symposium on Information Theory (ISIT)*, 2022.

[43] Milad Sefidgaran, Romain Chor, and Abdellatif Zaidi. Rate-distortion theoretic bounds on generalization error for distributed learning. *Advances in Neural Information Processing Systems*, 35, 19687-19702.

[44] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[45] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 141–142, 2012.

[46] Tsing, Yuwen Yang, and NewAlexandria. TsingZ0/PFL-Non-IID: First Release (v0.1.0). Zenodo. https://doi.org/10.5281/zenodo.7780680, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix of "Understanding Generalization of Federated Learning via Stability: Heterogeneity Matters"

A Federated learning algorithms

In this section, we summarize FedAvg, SCAFFOLD and FedProx in detail in Algorithms 1,2,3, respectively.

Algorithm 1 FedAvg

Input: θ^0 as initialization of the server

for $t = 0, 1, \ldots, T-1$ **do**
  $\theta_{i,0}^{t+1} = \theta^t$, $\forall i = 1, \ldots, m$
  for $k = 0, 1, \ldots, K_i - 1$ **do** (in parallel for all agents)
    $\theta_{i,k+1}^{t+1} = \theta_{i,k}^{t+1} - \alpha_{i,k} g_i(\theta_{i,k}^{t+1})$
  end for
  $\theta^{t+1} = \sum_{i=1}^m p_i \theta_{i,K_i}^{t+1}$
end for

Output: θ^T given by the server

Algorithm 2 SCAFFOLD

Input: θ^0 as initialization of the server

for $t = 0, 1, \ldots, T-1$ **do**
  Server broadcasts θ^t
  Agent computes $g_i(\theta^t)$ and send it to the server
  Server computes $g(\theta^t) = \sum_{i=1}^m p_i g_i(\theta^t)$ and broadcasts it
  Each agent i for $i = 1, \ldots, m$ sets $\theta_{i,0}^{t+1} = \theta^t$
  for $k = 0, 1, \ldots, K_i - 1$ **do** (in parallel for all agents)
    $\theta_{i,k+1}^{t+1} = \theta_{i,k}^{t+1} - \alpha_{i,k}\big(g_i(\theta_{i,k}^{t+1}) - g_i(\theta^t) + g(\theta^t)\big)$
  end for
  $\theta^{t+1} = \sum_{i=1}^m p_i \theta_{i,K_i}^{t+1}$
end for

Output: θ^T given by the server

Algorithm 3 FedProx

Input: θ^0 as initialization of the server

for $t = 0, 1, \ldots, T-1$ **do**
  $\theta_i^{t+1} = \arg\min_\theta \hat{R}_{\mathcal{S}_i}(\theta) + \frac{1}{2\eta_i}\|\theta - \theta^t\|^2$ (in parallel for all agents)
  $\theta^{t+1} = \sum_{i=1}^m p_i \theta_i^{t+1}$
end for

Output: θ^T given by the server

B Proof of Theorem 1

In this section, we provide the proof of Theorem 1.

Given $\mathcal{S}$ and $\mathcal{S}^{(i)}$ which are neighboring datasets defined in Definition 1,

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\hat{R}_{\mathcal{S}_i}(\mathcal{A}(\mathcal{S}))\right] &= \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n_i}\sum_{j=1}^{n_i} l(\mathcal{A}(\mathcal{S}); z_{i,j})\right] \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}_{\mathcal{S}}\left[l(\mathcal{A}(\mathcal{S}); z_{i,j})\right] \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}_{\mathcal{S},z'_{i,j}}\left[l(\mathcal{A}(\mathcal{S}^{(i)}); z'_{i,j})\right].\end{aligned}$$

Moreover, we have

$$\mathbb{E}_{\mathcal{S}}\left[R_i(\mathcal{A}(\mathcal{S}))\right] = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}_{\mathcal{S},z'_{i,j}}\left[l(\mathcal{A}(\mathcal{S}); z'_{i,j})\right],$$

since $z'_{i,j}$ and $\mathcal{S}$ are independent for any j. Thus,

$$\begin{aligned}\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[R(\mathcal{A}(\mathcal{S})) - \hat{R}(\mathcal{A}(\mathcal{S}))\right] &\leq \mathbb{E}_{\mathcal{A},\mathcal{S}}\left[\sum_{i=1}^{m}\frac{n_i}{n}\left(R_i(\mathcal{A}(\mathcal{S})) - \hat{R}_{\mathcal{S}_i}(\mathcal{A}(\mathcal{S}))\right)\right] \\ &= \sum_{i=1}^{m}\frac{n_i}{n}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}_{\mathcal{S},z'_{i,j}}\left(l(\mathcal{A}(\mathcal{S}); z'_{i,j}) - l(\mathcal{A}(\mathcal{S}^{(i)}); z'_{i,j})\right)\right] \\ &\leq \epsilon,\end{aligned}$$

where the last inequality follows Definition 2. This completes the proof.

C Generalization bounds for convex losses

In this section, we drop index t when context is clear for simplicity. We first provide the bound involving data heterogeneity by means of total variation between local distribution and global one.

Lemma 1. *Under Assumption 1 and given $i \in [m]$, for any θ we have*

$$\|\nabla R_i(\theta) - \nabla R(\theta)\| \leq 2LD_i,$$

where $D_i = d_{TV}(P_i, P)$ with $P = \sum_{i=1}^{m} p_i P_i$.

Proof. Let $\mathcal{Z}_i$ and $\mathcal{Z}$ be the supports of P_i and P, respectively.

$$\begin{aligned}\|\nabla R_i(\theta) - \nabla R(\theta)\| &= \|\nabla_\theta \int_{\mathcal{Z}_i} l(\theta; z) dP_i(z) - \nabla_\theta \int_{\mathcal{Z}} l(\theta; z) dP(z)\| \\ &= \|\int_{\mathcal{Z}_i \cup \mathcal{Z}} \left(\nabla_\theta l(\theta; z) dP_i(z) - \nabla_\theta l(\theta; z) dP(z)\right)\| \\ &\leq \int_{\mathcal{Z}_i \cup \mathcal{Z}} \|\nabla_\theta l(\theta; z) dP_i(z) - \nabla_\theta l(\theta; z) dP(z)\| \\ &= \int_{\mathcal{Z}_i \cup \mathcal{Z}} \|\nabla l(\theta; z)\| \|dP_i(z) - dP(z)\| \\ &\leq \int_{\mathcal{Z}_i \cup \mathcal{Z}} L|dP_i(z) - dP(z)| \\ &= 2L d_{TV}(P_i, P)\end{aligned}$$

by noting the definition of total variation of two distributions P and Q is

$$d_{TV}(P,Q) = \frac{1}{2}\int |dP - dQ|.$$

□

When the loss function is convex, the gradient descent operator has the non-expansiveness property stated by the following lemma.

Lemma 2. *Suppose $f(x)$ is a β-Lipschitz smooth, convex function with respect to x. Consider gradient descent operator $G_\alpha(x) := x - \alpha\nabla f(x)$. Then, for $\alpha \leq 1/\beta$,*

$$\|G_\alpha(x) - G_\alpha(y)\| \leq \|x - y\|.$$

Proof. Since f is β-smooth and convex, we know that

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2.$$

Using this fact,

$$\begin{aligned}
\|G_\alpha(x) - G_\alpha(y)\|^2 &= \|x - y - \alpha(\nabla f(x) - \nabla f(y))\|^2 \\
&= \|x-y\|^2 + \alpha^2\|\nabla f(x) - \nabla f(y)\|^2 - \alpha\langle \nabla f(x) - \nabla f(y), x - y\rangle \\
&\leq \|x-y\|^2 + \alpha(\alpha - \beta^{-1})\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \|x-y\|^2
\end{aligned}$$

when $\alpha \leq 1/\beta$. □

The proximal operator is also non-expansive, which is shown by the following lemma.

Lemma 3. *Suppose f is convex. Define the proximal operator by*

$$\text{prox}_f(x) := \arg\min_y f(y) + \frac{1}{2}\|y - x\|^2.$$

Then, for any x_1, x_2, we have

$$\|\text{prox}_f(x_1) - \text{prox}_f(x_2)\| \leq \|x_1 - x_2\|.$$

Proof. Let $u_1 = \text{prox}_f(x_1)$ and $u_2 = \text{prox}_f(x_2)$. According to the first-order optimality condition, we have

$$\begin{aligned}
\nabla f(u_1) + u_1 - x_1 &= 0 \\
\nabla f(u_2) + u_2 - x_2 &= 0
\end{aligned}$$

Since f is convex, we further have

$$\begin{aligned}
0 &\leq \langle \nabla f(u_1) - \nabla f(u_2), u_1 - u_2\rangle \\
&= \langle x_1 - u_2 - (x_2 - u_2), u_1 - u_2\rangle \\
&= \langle x_1 - x_2, u_1 - u_2\rangle - \|u_1 - u_2\|^2
\end{aligned}$$

and hence

$$\|u_1 - u_2\|^2 \leq \langle x_1 - x_2, u_1 - u_2\rangle \leq \|x_1 - x_2\|\|u_1 - u_2\|$$

which completes the proof. □

C.1 Analysis for FedAvg under convex losses

Lemma 4. *Suppose Assumptions 1-4 hold. Then for FedAvg with $\alpha_{i,k} \leq 1/\beta$,*

$$\mathbb{E}\|\theta_{i,k} - \theta_t\| \leq \tilde{\alpha}_{i,t}\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),\ \forall k = 1, \ldots, K_i,$$

where $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Considering local update (3) of FedAvg

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k+1} - \theta_t\| &= \mathbb{E}\|\theta_{i,k} - \alpha_{i,k} g_i(\theta_{i,k}) - \theta_t\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t))\| + \alpha_{i,k}\mathbb{E}\|g_i(\theta_t)\| \\
&\overset{(a)}{\leq} \mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}\mathbb{E}\|g_i(\theta_t)\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|g_i(\theta_t) - \nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t)\|) \\
&\overset{(b)}{\leq} \mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma),
\end{aligned}$$

where (a) follows Lemma 2; (b) follows Assumption 2. Unrolling the above and noting $\theta_{i,0} = \theta_t$ yields

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k} - \theta_t\| &\leq \mathbb{E}\|\theta_{i,0} - \theta_t\| + \sum_{l=0}^{k-1} \alpha_{i,l}\big(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma\big) \\
&\leq \sum_{l=0}^{K_i-1} \alpha_{i,l}\big(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma\big) \\
&= \tilde{\alpha}_i\big(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma\big) \\
&\leq \tilde{\alpha}_i\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),
\end{aligned}$$

where the last inequality follows Lemma 1. □

Lemma 5. *Given Assumptions 1-4 and considering* (3) *of FedAvg, for $\alpha_{i,k} \leq 1/\beta$ we have*

$$\mathbb{E}\|g_i(\theta_{i,k})\| \leq (1 + \beta\tilde{\alpha}_{i,t})\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),$$

where $g_i(\cdot)$ is the sampled gradient of client i, $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Using Lemmas 1 and 4, we obtain

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq \mathbb{E}\|g_i(\theta_{i,k}) - \nabla R_i(\theta_{i,k})\| + \mathbb{E}\|\nabla R_i(\theta_{i,k})\| \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{i,k})\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_{i,k}) - \nabla R_i(\theta_t)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \beta\mathbb{E}\|\theta_{i,k} - \theta_t\| + \sigma \\
&\leq (1 + \beta\tilde{\alpha}_i)\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big).
\end{aligned}$$

□

Theorem 5 (FedAvg part of Theorem 2)**.** *Suppose Assumptions 1-4 hold and consider FedAvg (Algorithm 1). Let $\{\theta_t\}_{t=0}^T$ and $\{\theta'_t\}_{t=0}^T$ be two trajectories of the server induced by neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Suppose $\theta_0 = \theta'_0$. Then,*

$$\mathbb{E}\|\theta_T - \theta'_T\| \leq \frac{2}{n}\sum_{t=0}^{T-1} \tilde{\alpha}_{i,t}(1 + \beta\tilde{\alpha}_{i,t})\big(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma\big),$$

where $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$ and $D_i = d_{TV}(P_i, P)$.

Proof. Note that in the phase of local update, each client runs stochastic gradient descent (SGD) using its own local gradient $g_i(\cdot)$ sampled uniformly from its dataset. Given time index t, for client j with $j \neq i$, the local datasets are identical since the perturbed data point only occurs at client i. Thus, when $j \neq i$, we have for any $k = 0, \ldots, K_j - 1$,

$$\begin{aligned}\mathbb{E}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &= \mathbb{E}\|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| \\ &\leq \mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\|\end{aligned}$$

where we use Lemma 2 in the last inequality. Here we drop t for simplicity. Unrolling it gives

$$\mathbb{E}\|\theta_{j,K_j} - \theta'_{j,K_j}\| \leq \mathbb{E}\|\theta_t - \theta'_t\|, \quad \forall j \neq i. \tag{7}$$

For client i, there are two cases to consider. In the first case, SGD selects the index of an sample at local step k on which is identical in $\mathcal{S}$ and $\mathcal{S}^{(i)}$. In this sense, we have

$$\|\theta_{i,k+1} - \theta'_{i,k+1}\| \leq \|\theta_{i,k} - \theta'_{i,k}\|$$

due to the non-expansiveness of gradient descent operator by Lemma 2. And this case happens with probability $1 - 1/n_i$ (since only one sample is perturbed for client i).

In the second case, SGD encounters the perturbed sample at local time step k, which happens with probability $1/n_i$. We denote the gradient of this perturbed sample as $g'_i(\cdot)$. Then,

$$\begin{aligned}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &= \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g'_i(\theta'_{i,k}))\| \\ &\leq \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta'_{i,k}))\| + \alpha_{i,k}\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| \\ &\leq \|\theta_{i,k} - \theta'_{i,k}\| + \alpha_{i,k}\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\|.\end{aligned}$$

Combining these two cases we have for client i

$$\begin{aligned}\mathbb{E}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq \mathbb{E}\|\theta_{i,k} - \theta'_{i,k}\| + \frac{\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| \\ &\leq \mathbb{E}\|\theta_{i,k} - \theta'_{i,k}\| + \frac{2\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta_{i,k})\|,\end{aligned}$$

where the last inequality follows that $g_i(\cdot)$ and $g'_i(\cdot)$ are sampled from the same distribution. Then unrolling it we have

$$\mathbb{E}\|\theta_{i,K_i} - \theta'_{i,K_i}\| \leq \mathbb{E}\|\theta_t - \theta'_t\| + \frac{2}{n_i}\sum_{k=0}^{K_i-1}\alpha_{i,k}\mathbb{E}\|g_i(\theta_{i,k})\|. \tag{8}$$

Combining (7) and (8) gives

$$\begin{aligned}\mathbb{E}\|\theta_{t+1} - \theta'_{t+1}\| &= \mathbb{E}\|\sum_{j=1}^{m} p_j(\theta_{j,K_j} - \theta'_{j,K_j})\| \\ &\leq \sum_{j=1}^{m} p_j\mathbb{E}\|\theta_{j,K_j} - \theta'_{j,K_j}\| \\ &\leq \mathbb{E}\|\theta_t - \theta'_t\| + \frac{2p_i}{n_i}\sum_{k=0}^{K_i-1}\alpha_{i,k}\mathbb{E}\|g_i(\theta_{i,k})\| \\ &\leq \mathbb{E}\|\theta_t - \theta'_t\| + \frac{2}{n}\tilde{\alpha}_{i,t}(1 + \beta\tilde{\alpha}_{i,t})\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),\end{aligned}$$

where we use Lemma 5 in the last step. Iterating the above over t and noting $\theta_0 = \theta'_0$, we conclude the proof.

□

C.2 Analysis for SCAFFOLD under convex losses

Lemma 6. *Suppose Assumptions 1-4 hold. Running SCAFFOLD with $\alpha_{i,k} \leq 1/\beta$, then for any $i \in [m]$*

$$\mathbb{E}\|\theta_{i,k} - \theta_t\| \leq \tilde{\alpha}_{i,t}(\mathbb{E}\|R(\theta_t)\| + \sigma), \quad \forall k = 1, \ldots, K_i$$

where $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Considering local update (4) of SCAFFOLD

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k+1} - \theta_t\| &= \mathbb{E}\|\theta_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t) + g(\theta_t)) - \theta_t\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t))\| + \alpha_{i,k}\mathbb{E}\|g(\theta_t)\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}\mathbb{E}\|g(\theta_t)\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|R(\theta_t)\| + \sigma)
\end{aligned}$$

where we use the non-expansiveness property of gradient descent operator and Assumption 2. Therefore, for any $k = 1, \ldots, K_i - 1$,

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k} - \theta_t\| &\leq \sum_{l=0}^{k-1} \alpha_{i,k}(\mathbb{E}\|R(\theta_t)\| + \sigma) \\
&\leq \tilde{\alpha}_{i,t}(\mathbb{E}\|R(\theta_t)\| + \sigma),
\end{aligned}$$

which completes the proof. □

Lemma 7. *Given Assumptions 1-4 and considering SCAFFOLD (Algorithm 2), with $\alpha_{i,k} \leq 1/\beta$ we have the following inequalities*

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq (1 + \beta\tilde{\alpha}_{i,t})(\mathbb{E}\|\nabla R(\theta_t)\| + \sigma) + 2LD_i, \\
\mathbb{E}\|g_i(\theta_t)\| &\leq 2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma
\end{aligned}$$

for any $i \in [m]$, $k = 0, \ldots, K_i - 1$ and $t = 0, 1, \ldots$.

Proof. Note that based on Assumption 2,

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq \mathbb{E}\|\nabla R_i(\theta_{i,k})\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{i,k}) - \nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma \\
&\leq \beta\mathbb{E}\|\theta_{i,k} - \theta_t\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq (1 + \beta\tilde{\alpha}_{i,t})(\mathbb{E}\|\nabla R(\theta_t)\| + \sigma) + 2LD_i,
\end{aligned}$$

where we use Lemmas 1 and 6.

Similarly, using same techniques we have

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_t)\| &\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq 2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma.
\end{aligned}$$

□

Theorem 6 (SCAFFOLD part of Theorem 2)**.** *Suppose Assumptions 1-4 hold and consider SCAFFOLD (Algorithm 2). Let $\{\theta_t\}_{t=0}^T$ and $\{\theta'_t\}_{t=0}^T$ be two trajectories of the server induced by neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Suppose $\theta_0 = \theta'_0$. Then*

$$\mathbb{E}\|\theta_T - \theta'_T\| \leq \frac{2}{n}\sum_{t=0}^{T-1} \exp\Big(2\beta \sum_{l=t+1}^{T-1} \hat{\alpha}_l\Big)\Big(2LD_i\gamma_t^1 + \gamma_t^2\mathbb{E}\|\nabla R(\theta_t)\| + \sigma\gamma_t^2\Big)$$

where

$$\gamma_t^1 := 2\tilde{\alpha}_{i,t} + \hat{\alpha}_t, \quad \gamma_t^2 := \gamma_t^1 + \beta\tilde{\alpha}_{i,t}^2$$

with $\tilde{\alpha}_{i,t} := \sum_{k=0}^{K_i-1} \alpha_{i,k}$, $\hat{\alpha}_t := \sum_{j=1}^m p_j\tilde{\alpha}_{j,t}$, and $\sum_{l=T}^{T-1} \hat{\alpha}_l = 0, \forall \hat{\alpha}_l$.

Proof. Similar to the idea used in the proof of Theorem 5, given time index t and client j with $j \neq i$, note that the local gradients $g_j(\cdot)$ are identical for client j in the sense that local datasets for client j are the same. However, since SCAFFOLD uses the global sampled gradient $g(\cdot)$ during the local update, it is still possible to encounter the perturbed sample. Thus, for $j \neq i$, we distinguish two cases. In the first case, SCAFFOLD does not sample the perturbed gradient of client i, i.e., $g(\cdot) = g'(\cdot)$ at local step k. Then, with probability equal to $1 - 1/n_i$

$$\begin{aligned}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\leq \|\theta_{j,k} - \theta'_{j,k}\| + 2\alpha_{j,k}\beta\|\theta_t - \theta'_t\|\end{aligned}$$

where the second inequality follows Lemma 2 and Assumption 3.

In the second case, the perturbed data point of client i is sampled to calculate the global gradient $g'(\cdot)$, meaning $g(\cdot) - g'(\cdot) = p_i(g_i(\cdot) - g'_i(\cdot))$, where we denote the gradient evaluated at the perturbed sample as $g'_i(\cdot)$. This happens with probability $1/n_i$ and hence we have

$$\begin{aligned}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g'(\theta'_t)\| \\ &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g(\theta'_t)\| + \alpha_{j,k}\|g(\theta'_t) - g'(\theta'_t)\| \\ &\leq \|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\|\theta_t - \theta'_t\| + \alpha_{j,k}p_i\|g_i(\theta'_t) - g'_i(\theta'_t)\|.\end{aligned}$$

Combining these two cases, we conclude that for client j with $j \neq i$

$$\begin{aligned}\mathbb{E}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &\leq \mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{\alpha_{j,k}p_i}{n_i}\mathbb{E}\|g_i(\theta'_t) - g'_i(\theta'_t)\| \\ &\leq \mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\alpha_{j,k}}{n}\mathbb{E}\|g_i(\theta_t)\|, \\ &\leq \mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\alpha_{j,k}}{n}(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma)\end{aligned}$$

where we use $p_i = n_i/n$ and g_i, g'_i are drawn from the same distribution; we also use Lemma 7 in the last step. Unrolling the above over k we obtain

$$\mathbb{E}\|\theta_{j,K_j} - \theta'_{j,K_j}\| \leq (1 + \beta\tilde{\alpha}_{j,t})\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\tilde{\alpha}_{j,t}}{n}(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma), \quad \forall j \neq i. \tag{9}$$

Next, we specifically consider client i. Similar to the above analysis, there are two cases as well. In the first case, at local step k client i does not select the perturbed sample to compute the gradient. This happens with probability $1 - 1/n_i$. Then,

$$\begin{aligned}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta'_{i,k}))\| + \alpha_{i,k}\|g_i(\theta_t) - g_i(\theta'_t)\| \\ &\quad + \alpha_{i,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\leq \|\theta_{i,k} - \theta'_{i,k}\| + 2\alpha_{i,k}\beta\|\theta_t - \theta'_t\|.\end{aligned}$$

In the second case, the perturbed sample is selected to calculate local gradient for client i, which has the probability equal to $1/n_i$. Then,

$$\begin{aligned}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g'_i(\theta'_{i,k}))\| + \alpha_{i,k}\|g_i(\theta_t) - g'_i(\theta'_t)\| \\ &\quad + \alpha_{i,k}\|g(\theta_t) - g'(\theta'_t)\| \\ &\leq \|\theta_{i,k} - \theta'_{i,k}\| + \alpha_{i,k}\|g_i(\theta_t) - g_i(\theta'_t)\| + \alpha_{i,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\quad + \alpha_{i,k}\big(\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| + (1 + p_i)\|g_i(\theta'_t) - g'_i(\theta'_t)\|\big) \\ &\leq \|\theta_{i,k} - \theta'_{i,k}\| + 2\beta\alpha_{i,k}\|\theta_t - \theta'_t\| + \alpha_{i,k}\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| \\ &\quad + \alpha_{i,k}(1 + p_i)\|g_i(\theta'_t) - g'_i(\theta'_t)\|\end{aligned}$$

where the non-expansiveness of gradient descent operator and Lipschitz smoothness are utilized.

Combining these two cases for client i and further leveraging Lemma 7, we obtain

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k+1}-\theta'_{i,k+1}\| &\leq \mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+2\beta\alpha_{i,k}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta'_{i,k})-g'_i(\theta'_{i,k})\| \\
&\quad+\frac{\alpha_{i,k}(1+p_i)}{n_i}\mathbb{E}\|g_i(\theta'_t)-g'_i(\theta'_t)\| \\
&\leq \mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+2\beta\alpha_{i,k}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta_{i,k})\| \\
&\quad+\frac{2\alpha_{i,k}(1+p_i)}{n_i}\mathbb{E}\|g_i(\theta_t)\| \\
&\leq \mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+2\beta\alpha_{i,k}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2\alpha_{i,k}}{n_i}(1+\beta\tilde{\alpha}_{i,t})(\mathbb{E}\|\nabla R(\theta_t)\|+\sigma) \\
&\quad+\frac{2\alpha_{i,k}(1+p_i)}{n_i}(2LD_i+\mathbb{E}\|\nabla R(\theta_t)\|+\sigma)+\frac{2\alpha_{i,k}}{n_i}2LD_i.
\end{aligned}$$

Unrolling it gives

$$\begin{aligned}
\mathbb{E}\|\theta_{i,K_i}-\theta'_{i,K_i}\| &\leq (1+\beta\tilde{\alpha}_{i,t})\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2\tilde{\alpha}_{i,t}}{n_i}\big(2LD_i+(1+\beta\tilde{\alpha}_{i,t})(\mathbb{E}\|\nabla R(\theta_t)\|+\sigma)\big) \\
&\quad+\frac{2\tilde{\alpha}_{i,t}(1+p_i)}{n_i}(2LD_i+\mathbb{E}\|\nabla R(\theta_t)\|+\sigma). \qquad (10)
\end{aligned}$$

By (9) and (10), we obtain

$$\begin{aligned}
\mathbb{E}\|\theta_{t+1}-\theta'_{t+1}\| &\leq \sum_{j=1}^m p_j\mathbb{E}\|\theta_{j,K_j}-\theta'_{j,K_j}\| \\
&\leq (1+\beta\hat{\alpha}_t)\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2\gamma_t^1}{n}2LD_i+\frac{2\gamma_t^2}{n}(\mathbb{E}\|\nabla R(\theta_t)\|+\sigma),
\end{aligned}$$

and we further keep iterate it over t to obtain

$$\mathbb{E}\|\theta_T-\theta'_T\|\leq\frac{2}{n}\sum_{t=0}^{T-1}\exp\Big(2\beta\sum_{l=t+1}^{T-1}\hat{\alpha}_l\Big)\Big(2LD_i\gamma_t^1+\gamma_t^2\mathbb{E}\|\nabla R(\theta_t)\|+\sigma\gamma_t^2\Big)$$

where we use the fact $1+x\leq e^x,\forall x$. □

C.3 Analysis for FedProx under convex losses

Lemma 8. *Suppose Assumptions 1, 2 and 4 hold. Considering FedProx with local update* (5)*, then for any $\eta_i>0$, we have for any $i\in[m]$*

$$\mathbb{E}\|\theta_{t+1}^i-\theta_t\|\leq\eta_i(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma),\quad\forall t=0,1,\ldots.$$

Proof. Recalling the local update (5) of FedProx and according to the first-order optimality condition, we have

$$\eta_i\nabla\hat{R}_{\mathcal{S}_i}(\theta_{t+1}^i)+\theta_{t+1}^i-\theta_t=0.$$

Moreover, since the function $\eta_i\hat{R}_{\mathcal{S}_i}(\theta)+\frac{1}{2}\|\theta-\theta_t\|$ is 1-strongly-convex when Assumption 4 holds, we have

$$\|\theta_{t+1}^i-\theta_t\|\leq\|\eta_i\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)+\theta_t-\theta_t\|=\eta_i\|\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)\|$$

by combining the first-order optimality condition. Moreover, note that

$$\begin{aligned}
\mathbb{E}\|\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)\| &\leq \mathbb{E}\|\nabla R_i(\theta_t)\|+\sigma \\
&\leq \mathbb{E}\|\nabla R(\theta_t)\|+\mathbb{E}\|\nabla R_i(\theta_t)-\nabla R(\theta_t)\|+\sigma \\
&\leq \mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma,
\end{aligned}$$

where we use Lemma 1 and note

$$\mathbb{E}\|\nabla \hat{R}_{\mathcal{S}_i}(\theta_t) - \nabla R_i(\theta_t)\| \leq \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}\|\nabla l(\theta_t; z_{i,j}) - \nabla R_i(\theta_t)\| \leq \sigma.$$

Thus, we have

$$\mathbb{E}\|\theta_{t+1}^i - \theta_t\| \leq \eta_i(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma),$$

which completes the proof. □

Lemma 9. *Suppose Assumptions 1-4 hold and consider FedProx with local update* (5)*. Then, for any $i \in [m]$ and $j \in [n_i]$, we have*

$$\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j})\| \leq (1+\beta\eta_i)(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma), \quad \forall t = 0, 1, \ldots.$$

Proof. For any $i \in [m]$ and time t,

$$\begin{aligned}
\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j})\| &\leq \mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla R_i(\theta_{t+1}^i)\| + \mathbb{E}\|\nabla R_i(\theta_{t+1}^i)\| \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{t+1}^i)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_{t+1}^i) - \nabla R_i(\theta_t)\| + \sigma \\
&\leq \beta\mathbb{E}\|\theta_{t+1}^i - \theta_t\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq (1+\beta\eta_i)(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma),
\end{aligned}$$

where we use Lemma 1 and Lemma 8 in the last step. □

Theorem 7 (FedProx part of Theorem 2)**.** *Suppose Assumptions 1-4 hold and consider FedProx (Algorithm 3). Let $\{\theta_t\}_{t=0}^T$ and $\{\theta_t'\}_{t=0}^T$ be two trajectories of the server induced by neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Suppose $\theta_0 = \theta_0'$. Then,*

$$\mathbb{E}\|\theta_T - \theta_T'\| \leq \frac{2}{n}\sum_{t=0}^{T-1}\eta_i(1+\beta\eta_i)\Big(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma\Big).$$

Proof. Denoting $\mathrm{prox}_f(x) := \arg\min_y f(y) + \frac{1}{2}\|y - x\|^2$, we can rewrite the local update (5) as

$$\theta_{t+1}^i = \mathrm{prox}_{\eta_i \hat{R}_{\mathcal{S}_i}}(\theta_t).$$

There are two different cases for local updates. For client j with $j \neq i$, we note $\hat{R}_{\mathcal{S}_j}(\cdot) = \hat{R}_{\mathcal{S}_j'}(\cdot)$ in the sense that there is no perturbation for client j. In this case, using Lemma 3 we obtain

$$\begin{aligned}
\|\theta_{t+1}^i - (\theta_{t+1}^i)'\| &= \|\mathrm{prox}_{\eta_i \hat{R}_{\mathcal{S}_i}}(\theta_t) - \mathrm{prox}_{\eta_i \hat{R}_{\mathcal{S}_i}}(\theta_t')\| \\
&\leq \|\theta_t - \theta_t'\|.
\end{aligned}$$

For client i, we note that $\hat{R}_i(\cdot) - \hat{R}_i'(\cdot) = \frac{1}{n_i}(l(\cdot; z_{i,j}) - l(\cdot; z_{i,j}'))$, where $z_{i,j}'$ is the perturbed data point. And we also use $\hat{R}_i$ and $\hat{R}_i'$ to represent $\hat{R}_{\mathcal{S}_i}$ and $\hat{R}_{\mathcal{S}_i'}$ for simplicity. Then, we have

$$\begin{aligned}
\theta_{t+1}^i &= \arg\min_\theta \eta_i \hat{R}_i(\theta) + \frac{1}{2}\|\theta - \theta_t\|^2 \\
(\theta_{t+1}^i)' &= \arg\min_\theta \eta_i \hat{R}_i'(\theta) + \frac{1}{2}\|\theta - \theta_t'\|^2.
\end{aligned}$$

According to the first-order optimality condition, it yields

$$\begin{aligned}
\theta_{t+1}^i - \theta_t &= -\eta_i \nabla \hat{R}_i(\theta_{t+1}^i) \\
(\theta_{t+1}^i)' - \theta_t' &= -\eta_i \nabla \hat{R}_i'((\theta_{t+1}^i)') \\
&= -\eta_i \nabla \hat{R}_i((\theta_{t+1}^i)') + \frac{\eta_i}{n_i}\left(\nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l((\theta_{t+1}^i)'; z_{i,j}')\right).
\end{aligned}$$

Moreover, by the monotone property of $\nabla \hat{R}_i(\cdot)$ for convex losses i.e., Lemma 3,

$$\begin{aligned}\|(\theta_{t+1}^i)' - \theta_{t+1}^i\|^2 &\leq \langle \theta_t' - \theta_t, (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle - \eta_i\langle \nabla \hat{R}_i((\theta_{t+1}^i)') - \nabla \hat{R}_i(\theta_{t+1}^i), (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle \\ &\quad + \frac{\eta_i}{n_i}\langle (\theta_{t+1}^i)' - \theta_{t+1}^i, \nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l(\theta_{t+1}^i; z_{i,j}')\rangle \\ &\leq \langle \theta_t' - \theta_t, (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle + \frac{\eta_i}{n_i}\langle (\theta_{t+1}^i)' - \theta_{t+1}^i, \nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l(\theta_{t+1}^i; z_{i,j}')\rangle\end{aligned}$$

which further implies by symmetry of $z_{i,j}$ and $z_{i,j}'$,

$$\|(\theta_{t+1}^i)' - \theta_{t+1}^i\| \leq \|\theta_t' - \theta_t\| + \frac{\eta_i}{n_i}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla l(\theta_{t+1}^i; z_{i,j}')\|.$$

Combining two cases gives

$$\begin{aligned}\mathbb{E}\|\theta_{t+1} - \theta_{t+1}'\| &\leq \sum_{j=1}^m p_j \mathbb{E}\|\theta_{t+1}^j - (\theta_{t+1}^j)'\| \\ &\leq \mathbb{E}\|\theta_t - \theta_t'\| + \frac{\eta_i}{n}\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla l(\theta_{t+1}^i; z_{i,j}')\|, \\ &\leq \mathbb{E}\|\theta_t - \theta_t'\| + \frac{2\eta_i}{n}\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j})\| \\ &\leq \mathbb{E}\|\theta_t - \theta_t'\| + \frac{2\eta_i}{n}(1 + \beta\eta_i)(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma).\end{aligned}$$

Unrolling it over t completes the proof. □

C.4 Proof of Theorem 3

Our results are established based on the following convergence results of three algorithms, which are formaly shown in Theorem 8. These results are based on the following assumptions.

Assumption 5. *There exist constants $G \geq 0$ and $B \geq 1$ such that*

$$\sum_{i=1}^m p_i \|\nabla R_i(\theta)\|^2 \leq 2G^2 + B^2\|\nabla R(\theta)\|^2, \quad \forall \theta.$$

Assumption 6. *There exist constants $G_i \geq 0$ such that for any $i \in [m]$,*

$$\|\nabla R_i(\theta) - \nabla R(\theta)\| \leq G_i, \quad \forall \theta.$$

In fact, Assumption 6 is a stronger assumption compared to Assumption 5, which is shown by the following proposition.

Proposition 1. *Assumption 6 implies Assumption 5.*

Proof. Note that given Assumption 6

$$\|\nabla R_i(\theta)\| \leq \|\nabla R(\theta)\| + \|\nabla R_i(\theta) - \nabla R(\theta)\| \leq G_i + \|\nabla R(\theta)\|,$$

which implies

$$\|\nabla R_i(\theta)\|^2 \leq 2G_i^2 + 2\|\nabla R(\theta)\|^2.$$

Taking the weighted sum of p_i and we conclude $G^2 = \sum_{i=1}^m p_i G_i^2$, $B^2 = 2$. □

In the next proposition, we characterize G_i defined in Assumption 6 by directly usting Lemma 1.

Proposition 2. *Suppose Assumption 1 holds. Then, $G_i = 2Ld_{TV}(P_i, P)$ defined in Assumption 6.*

Then, we state the existing convergence results for FedAvg, SCAFFOLD and FedProx in the following theorem.

Theorem 8. *[7, 8] Suppose Assumption 2 holds and $K_i = K, \forall i \in [m]$.*

For FedAvg (Algorithm 1) with Assumptions 3,5 satisfied and $\alpha_{i,k} \leq \frac{1}{(1+B^2)8\beta K}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla R(\theta_t)\|^2 \leq \mathcal{O}\left(\frac{\sqrt{\Delta_0}}{\sqrt{TKm}} + \frac{(\Delta_0 G)^{2/3}}{T^{2/3}} + \frac{B^2\Delta_0}{T}\right). \tag{11}$$

For SCAFFOLD (Algorithm 2) with Assumption 3 and $\alpha_{i,k} \leq \frac{1}{24\beta K}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla R(\theta_t)\|^2 \leq \mathcal{O}\left(\frac{\sqrt{\Delta_0}}{\sqrt{TKm}} + \frac{\Delta_0}{T}\right). \tag{12}$$

Suppose Assumption 6 hold. For FedProx (Algorithm 3) with eigenvalues of $\nabla^2 R(\theta)$ lower bounded and η_i chosen small enough, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla R(\theta_t)\|^2 \leq \mathcal{O}\left(\frac{\Delta_0\sum_{i=1}^m p_i G_i^2}{\sqrt{T}} + \frac{\Delta_0}{T}\right), \tag{13}$$

where $\Delta_0 := \mathbb{E}[R(\theta_0) - R(\theta^*)]$.

Proof of FedAvg and FedProx parts of Theorem 3. It follows the fact that stepsizes $\alpha_{i,k}$ and η_i are upper bounded by some constant c and

$$\left(\sum_{t=0}^{T-1} c\mathbb{E}\|\nabla R(\theta_t)\|\right)^2 \leq T\sum_{t=0}^{T-1} c^2\big(\mathbb{E}\|\nabla R(\theta_t)\|\big)^2 \leq T\sum_{t=0}^{T-1} c^2\mathbb{E}\|\nabla R(\theta_t)\|^2, \tag{14}$$

where the second inequality follows Jensen's inequality. Combining Propositions 1 and 2 with (11),(13) completes the proof.

Proof of SCAFFOLD part of Theorem 3. To get the result for SCAFFOLD in Theorem 3, we further note that γ_t^2 is upper bounded by some constant $\bar{\gamma}$ and when $\alpha_{i,k} \leq 1/[24\beta K(t+1)]$

$$\begin{aligned}\sum_{t=0}^{T-1}\exp\left(2\beta\sum_{l=t+1}^{T-1}\hat{\alpha}_l\right)\gamma_t^2\mathbb{E}\|\nabla R(\theta_t)\| &\leq \sum_{t=0}^{T-1}\exp\left(\frac{1}{12}\log(T)\right)\gamma_t^2\mathbb{E}\|\nabla R(\theta_t)\| \\ &\leq T^{1/12}\sum_{t=0}^{T-1}\bar{\gamma}\mathbb{E}\|\nabla R(\theta_t)\|. \end{aligned}\tag{15}$$

Combining (15) with (12) and (14) completes the proof.

C.5 Proofs of Corollaries 1 and 2

To obtain Corollary 1, we note that under Assumption 1,

$$\mathbb{E}_{\mathcal{A},\mathcal{S},z'_{i,j}}|l(\theta_T; z'_{i,j}) - l(\theta'_T; z'_{i,j})| \leq L\mathbb{E}\|\theta_T - \theta'_T\|,\ \forall j \in [n_i]$$

and then combining Theorems 1,2,3 provides the results.

To obtain Corollary 2, we start from Theorem 2. Note that given Assumption 1, we can bound $\mathbb{E}\|\nabla R(\theta_t)\|$ by Lipschitz constant L, i.e., $\mathbb{E}\|\nabla R(\theta_t)\| \leq L$. Moreover, under the i.i.d. case, meaning $D_i = 0, \forall i \in [m]$, we conclude the proof by using the same techniques as those in (14) and (15).

Remark 1. *Note that the bounds in Corollary 2 are also looser, compared to those in Corollary 1 even when $D_{max} = 0$ (which corresponds to the i.i.d. case). To see this, note that bounds in Corollary 2 are linear in T, while bounds in Corollary 1 are with $\mathcal{O}(T^q)$ for some $q < 1$. Moreover, more information is captured in Corollary 1, e.g., number of clients m, distance of the initial point to the optimal one Δ_0, etc.*

D Generalization bounds for non-convex losses

D.1 Analysis for FedAvg under non-convex losses

Lemma 10. *Suppose Assumptions 1-3 hold. Then for FedAvg with $\alpha_{i,k} \leq c/\beta$ for some $c > 0$,*

$$\mathbb{E}\|\theta_{i,k} - \theta_t\| \leq (1+c)^{K_i-1}\tilde{\alpha}_{i,t}\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),\ \forall k = 1, \ldots, K_i,$$

where $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Considering local update (3) of FedAvg

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k+1} - \theta_t\| &= \mathbb{E}\|\theta_{i,k} - \alpha_{i,k}g_i(\theta_{i,k}) - \theta_t\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t))\| + \alpha_{i,k}\mathbb{E}\|g_i(\theta_t)\| \\
&\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}\mathbb{E}\|g_i(\theta_t)\| \\
&\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|g_i(\theta_t) - \nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t)\|) \\
&\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma),
\end{aligned}$$

where we use Assumptions 2 and 3. Unrolling the above and noting $\theta_{i,0} = \theta_t$ yields

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k} - \theta_t\| &\leq \sum_{l=0}^{k-1} \alpha_{i,l}\big(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma\big)(1+c)^{k-1-l} \\
&\leq \sum_{l=0}^{K_i-1} \alpha_{i,l}\big(\mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma\big)(1+c)^{K_i-1} \\
&\leq (1+c)^{K_i-1}\tilde{\alpha}_{i,t}\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),
\end{aligned}$$

where the last inequality follows Lemma 1. □

Lemma 11. *Given Assumptions 1-3 and considering (3) of FedAvg, for $\alpha_{i,k} \leq c/\beta$ with some $c > 0$, we have*

$$\mathbb{E}\|g_i(\theta_{i,k})\| \leq (1 + (1+c)^{K_i-1}\beta\tilde{\alpha}_{i,t})\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big),$$

where $g_i(\cdot)$ is the sampled gradient of client i, $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Using Lemmas 1 and 10, we obtain

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq \mathbb{E}\|g_i(\theta_{i,k}) - \nabla R_i(\theta_{i,k})\| + \mathbb{E}\|\nabla R_i(\theta_{i,k})\| \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{i,k})\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_{i,k}) - \nabla R_i(\theta_t)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \beta\mathbb{E}\|\theta_{i,k} - \theta_t\| + \sigma \\
&\leq (1 + (1+c)^{K_i-1}\beta\tilde{\alpha}_i)\big(\mathbb{E}\|\nabla R(\theta_t)\| + 2LD_i + \sigma\big).
\end{aligned}$$

□

Theorem 9 (FedAvg part of Theorem 4)**.** *Suppose Assumptions 1-3 hold and consider FedAvg (Algorithm 1). Let $K_i = K, \forall i \in [m]$ and $\alpha_{i,k} \leq \frac{1}{24\beta K(t+1)}$. Then,*

$$\epsilon_{gen} \leq \mathcal{O}\Big(\frac{T^{\frac{1}{24}}\log T}{n}(D_{max} + \sigma)\Big) + \mathcal{O}\Big(\big(\frac{\Delta_0}{Km}\big)^{\frac{1}{4}}\frac{T^{\frac{5}{6}}}{n} + \big(\Delta_0^2\tilde{D}\big)^{\frac{1}{6}}\frac{T^{\frac{3}{4}}}{n} + \sqrt{\Delta_0}\frac{T^{\frac{7}{12}}}{n}\Big).$$

Proof. The proof is similar to that of Theorem 5. Given time index t and for client j with $j \neq i$, we have

$$\begin{aligned}
\mathbb{E}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &= \mathbb{E}\|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| \\
&\leq (1+\beta\alpha_{j,k})\mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\|.
\end{aligned}$$

And unrolling it gives

$$\begin{aligned}\mathbb{E}\|\theta_{j,K_j}-\theta'_{j,K_j}\| &\leq \prod_{k=0}^{K_j-1}(1+\beta\alpha_{j,k})\mathbb{E}\|\theta_t-\theta'_t\| \\ &\leq e^{\beta\tilde{\alpha}_{j,t}}\mathbb{E}\|\theta_t-\theta'_t\|, \quad \forall j\neq i, \end{aligned} \tag{16}$$

where we use $1+x\leq e^x, \forall x$.

For client i, there are two cases to consider. In the first case, SGD selects non-perturbed samples in $\mathcal{S}$ and $\mathcal{S}^{(i)}$, which happens with probability $1-1/n_i$. Then, we have

$$\|\theta_{i,k+1}-\theta'_{i,k+1}\|\leq(1+\beta\alpha_{i,k})\|\theta_{i,k}-\theta'_{i,k}\|.$$

In the second case, SGD encounters the perturbed sample at time step k, which happens with probability $1/n_i$. Then, we have

$$\begin{aligned}\|\theta_{i,k+1}-\theta'_{i,k+1}\| &= \|\theta_{i,k}-\theta'_{i,k}-\alpha_{i,k}(g_i(\theta_{i,k})-g'_i(\theta'_{i,k}))\| \\ &\leq \|\theta_{i,k}-\theta'_{i,k}-\alpha_{i,k}(g_i(\theta_{i,k})-g_i(\theta'_{i,k}))\|+\alpha_{i,k}\|g_i(\theta'_{i,k})-g'_i(\theta'_{i,k})\| \\ &\leq (1+\beta\alpha_{i,k})\|\theta_{i,k}-\theta'_{i,k}\|+\alpha_{i,k}\|g_i(\theta'_{i,k})-g'_i(\theta'_{i,k})\|.\end{aligned}$$

Combining these two cases for client i we have

$$\begin{aligned}\mathbb{E}\|\theta_{i,k+1}-\theta'_{i,k+1}\| &\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+\frac{\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta'_{i,k})-g'_i(\theta'_{i,k})\| \\ &\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+\frac{\alpha_{i,k}}{n_i}\mathbb{E}\|g_i(\theta_{i,k})\|, \\ &\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+\frac{2\alpha_{i,k}}{n_i}(1+(1+c)^{K_i-1}\beta\tilde{\alpha}_{i,t})\big(\sigma \\ &\quad +\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i\big) \\ &\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k}-\theta'_{i,k}\|+\frac{2\alpha_{i,k}\tilde{c}}{n_i}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma)\end{aligned}$$

where we use Lemma 11 and we let $\tilde{c}$ be an upper bound of $1+(1+c)^{K_i-1}\beta\tilde{\alpha}_{i,t}$ since $\tilde{\alpha}_{i,t}$ is bounded above. Then unrolling it gives

$$\begin{aligned}\mathbb{E}\|\theta_{i,K_i}-\theta'_{i,K_i}\| &\leq \prod_{k=0}^{K_i-1}(1+\beta\alpha_{i,k})\mathbb{E}\|\theta_t-\theta'_t\|+\bigg(\frac{2}{n_i}\sum_{k=0}^{K_i-1}\alpha_{i,k}\tilde{c}\prod_{l=k+1}^{K_i-1}(1+\beta\alpha_{i,l}) \\ &\quad \cdot(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma)\bigg) \\ &\leq e^{\beta\tilde{\alpha}_{i,t}}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2}{n_i}\tilde{c}\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma).\end{aligned} \tag{17}$$

By (16) and (17) we have

$$\begin{aligned}\mathbb{E}\|\theta_{t+1}-\theta'_{t+1}\| &\leq \sum_{i=1}^{m}p_i\mathbb{E}\|\theta_{i,K_i}-\theta'_{i,K_i}\| \\ &\leq e^{\beta\tilde{\alpha}_{i,t}}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2}{n}\tilde{c}\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma)\end{aligned}$$

where we also use $p_i=n_i/n$ in the last step. Further, unrolling the above over t and noting $\theta_0=\theta'_0$, we obtain

$$\mathbb{E}\|\theta_T-\theta'_T\|\leq\frac{2\tilde{c}}{n}\sum_{t=0}^{T-1}\exp\left(\beta\sum_{l=t+1}^{T-1}\tilde{\alpha}_{i,t}\right)\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma)$$

When the diminishing stepsizes are chosen in the statement of the theorem, we further combine Theorem 1 and the same techniques used in Theorem 3, we conclude the proof. □

D.2 Analysis for SCAFFOLD under non-convex losses

Lemma 12. *Suppose Assumptions 1-3 hold. Running SCAFFOLD with $\alpha_{i,k} \leq c/\beta$ for some $c > 0$, then for any $i \in [m]$*

$$\mathbb{E}\|\theta_{i,k} - \theta_t\| \leq (1+c)^{K_i-1}\tilde{\alpha}_{i,t}(\mathbb{E}\|R(\theta_t)\| + \sigma), \quad \forall k = 1, \ldots, K_i$$

where $\tilde{\alpha}_{i,t} = \sum_{k=0}^{K_i-1} \alpha_{i,k}$.

Proof. Considering local update (4) of SCAFFOLD

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k+1} - \theta_t\| &= \mathbb{E}\|\theta_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t) + g(\theta_t)) - \theta_t\| \\
&\leq \mathbb{E}\|\theta_{i,k} - \theta_t - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta_t))\| + \alpha_{i,k}\mathbb{E}\|g(\theta_t)\| \\
&\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}\mathbb{E}\|g(\theta_t)\| \\
&\leq (1+\beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta_t\| + \alpha_{i,k}(\mathbb{E}\|R(\theta_t)\| + \sigma)
\end{aligned}$$

where we use Assumptions 2 and 3. Therefore, for any $k = 1, \ldots, K_i - 1$,

$$\begin{aligned}
\mathbb{E}\|\theta_{i,k} - \theta_t\| &\leq \sum_{k=0}^{K_i-1} \alpha_{i,k}(\mathbb{E}\|R(\theta_t)\| + \sigma)(1+c)^{K_i-1} \\
&= \tilde{\alpha}_{i,t}(1+c)^{K_i-1}(\mathbb{E}\|R(\theta_t)\| + \sigma)
\end{aligned}$$

which completes the proof. □

Lemma 13. *Given Assumptions 1-3 and considering SCAFFOLD (Algorithm 2), with $\alpha_{i,k} \leq c/\beta$ for some $c > 0$ we have the following inequalities*

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq (1 + \beta\tilde{\alpha}_{i,t}(1+c)^{K_i-1})(\mathbb{E}\|\nabla R(\theta_t)\| + \sigma) + 2LD_i, \\
\mathbb{E}\|g_i(\theta_t)\| &\leq 2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma
\end{aligned}$$

for any $i \in [m]$, $k = 0, \ldots, K_i - 1$ and $t = 0, 1, \ldots$.

Proof. Note that based on Assumption 2,

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_{i,k})\| &\leq \mathbb{E}\|\nabla R_i(\theta_{i,k})\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{i,k}) - \nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma \\
&\leq \beta\mathbb{E}\|\theta_{i,k} - \theta_t\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq (1 + \beta\tilde{\alpha}_{i,t}(1+c)^{K_i-1})(\mathbb{E}\|\nabla R(\theta_t)\| + \sigma) + 2LD_i,
\end{aligned}$$

where we use Lemmas 1 and 12.

Similarly, using same techniques we have

$$\begin{aligned}
\mathbb{E}\|g_i(\theta_t)\| &\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq 2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma.
\end{aligned}$$

□

Theorem 10 (SCAFFOLD part of Theorem 4)**.** *Suppose Assumptions 1-3 hold and consider SCAFFOLD (Algorithm 2). Let $K_i = K$ and $\alpha_{i,k} \leq \frac{1}{24\beta K(t+1)}$, $\forall i \in [m]$*

$$\epsilon_{gen} \leq \mathcal{O}\Big(\frac{T^{\frac{1}{8}}\log T}{n} D_{max}\Big) + \mathcal{O}\Big((\frac{\Delta_0}{Km})^{\frac{1}{4}}\frac{T^{\frac{7}{8}}}{n} + \sqrt{\Delta_0}\frac{T^{\frac{5}{8}}}{n}\Big) + \mathcal{O}\Big(\frac{T^{\frac{1}{8}}(\log T + 1)}{n}\sigma\Big),$$

where $\Delta_0 = \mathbb{E}[R(\theta_0) - R(\theta^*)]$.

Proof. Similar to the proof of Theorem 6, considering client j with $j \neq i$, there are two cases. In the first case, SCAFFOLD does not select the perturbed sample from client i's dataset at local step k. Then, with probability equal to $1 - 1/n_i$,

$$\begin{aligned}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\leq (1 + \beta\alpha_{j,k})\|\theta_{j,k} - \theta'_{j,k}\| + 2\alpha_{j,k}\beta\|\theta_t - \theta'_t\|\end{aligned}$$

where the second inequality follows Assumption 3.

In the second case, there is with probability $1/n_i$ that the perturbed sample is selected during the local update of step k. Then,

$$\begin{aligned}\|\theta_{j,k+1} - \theta'_{j,k+1}\| &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g'(\theta'_t)\| \\ &\leq \|\theta_{j,k} - \theta'_{j,k} - \alpha_{j,k}(g_j(\theta_{j,k}) - g_j(\theta'_{j,k}))\| + \alpha_{j,k}\|g_j(\theta_t) - g_j(\theta'_t)\| \\ &\quad + \alpha_{j,k}\|g(\theta_t) - g(\theta'_t)\| + \alpha_{j,k}\|g(\theta'_t) - g'(\theta'_t)\| \\ &\leq (1 + \beta\alpha_{i,k})\|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\|\theta_t - \theta'_t\| + \alpha_{j,k}p_i\|g_i(\theta'_t) - g'_i(\theta'_t)\|.\end{aligned}$$

We again use Assumption 3 in the last step. Combining two cases, we have for client j with $j \neq i$

$$\mathbb{E}\|\theta_{j,k+1} - \theta'_{j,k+1}\| \leq (1 + \beta\alpha_{j,k})\mathbb{E}\|\theta_{j,k} - \theta'_{j,k}\| + 2\beta\alpha_{j,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\alpha_{j,k}}{n}\mathbb{E}\|g_i(\theta_t)\|.$$

Unrolling it over k we obtain

$$\begin{aligned}\mathbb{E}\|\theta_{j,K_j} - \theta'_{j,K_j}\| &\leq \prod_{k=0}^{K_j-1}(1 + \beta\alpha_{j,k})\mathbb{E}\|\theta_t - \theta'_t\| + \Bigg(\sum_{k=0}^{K_j-1}\Big(\prod_{l=k+1}^{K_j-1}(1 + \beta\alpha_{j,l})\Big) \\ &\quad \cdot(2\beta\alpha_{j,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\alpha_{j,k}}{n}\mathbb{E}\|g_i(\theta_t)\|)\Bigg) \\ &\leq (1 + \beta\tilde{\alpha}_{j,t})e^{\beta\tilde{\alpha}_{j,t}}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2\tilde{\alpha}_{j,t}}{n}e^{\beta\tilde{\alpha}_{j,t}}\mathbb{E}\|g_i(\theta_t)\| \qquad (18)\end{aligned}$$

where we use the fact $1 + x \leq e^x$ and Lemma 13 in the last step.

For client i, there are two cases as well. In the first case, the perturbed sample is not selected at step k, which happens with probability $1 - 1/n_i$. Then,

$$\begin{aligned}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g_i(\theta'_{i,k}))\| + \alpha_{i,k}\|g_i(\theta_t) - g_i(\theta'_t)\| \\ &\quad + \alpha_{i,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\leq (1 + \beta\alpha_{i,k})\|\theta_{i,k} - \theta'_{i,k}\| + 2\alpha_{i,k}\beta\|\theta_t - \theta'_t\|.\end{aligned}$$

In the second case, the perturbed sample is selected at local step k with probability $1/n_i$. Then,

$$\begin{aligned}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq \|\theta_{i,k} - \theta'_{i,k} - \alpha_{i,k}(g_i(\theta_{i,k}) - g'_i(\theta'_{i,k}))\| + \alpha_{i,k}\|g_i(\theta_t) - g'_i(\theta'_t)\| \\ &\quad + \alpha_{i,k}\|g(\theta_t) - g'(\theta'_t)\| \\ &\leq (1 + \beta\alpha_{i,k})\|\theta_{i,k} - \theta'_{i,k}\| + \alpha_{i,k}\|g_i(\theta_t) - g_i(\theta'_t)\| + \alpha_{i,k}\|g(\theta_t) - g(\theta'_t)\| \\ &\quad + \alpha_{i,k}\big(\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| + (1 + p_i)\|g_i(\theta'_t) - g'_i(\theta'_t)\|\big) \\ &\leq (1 + \beta\alpha_{i,k})\|\theta_{i,k} - \theta'_{i,k}\| + 2\beta\alpha_{i,k}\|\theta_t - \theta'_t\| + \alpha_{i,k}\|g_i(\theta'_{i,k}) - g'_i(\theta'_{i,k})\| \\ &\quad + \alpha_{i,k}(1 + p_i)\|g_i(\theta'_t) - g'_i(\theta'_t)\|.\end{aligned}$$

Combining these two case renders

$$\begin{aligned}\mathbb{E}\|\theta_{i,k+1} - \theta'_{i,k+1}\| &\leq (1 + \beta\alpha_{i,k})\mathbb{E}\|\theta_{i,k} - \theta'_{i,k}\| + 2\beta\alpha_{i,k}\mathbb{E}\|\theta_t - \theta'_t\| + \frac{2}{\alpha_{i,k}}n_i\mathbb{E}\|g_i(\theta_{i,k})\| \\ &\quad + \frac{2\alpha_{i,k}(1 + p_i)}{n_i}\mathbb{E}\|g_i(\theta_t)\|\end{aligned}$$

and unrolling it and using Lemma 13 gives

$$\begin{aligned}\mathbb{E}\|\theta_{i,K_i}-\theta'_{i,K_i}\| &\leq (1+2\beta\tilde{\alpha}_{i,t})e^{\beta\tilde{\alpha}_{i,t}}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2(1+p_i)}{n_i}\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}\mathbb{E}\|g_i(\theta_t)\| \\ &\quad+\frac{2\tilde{\alpha}_{i,t}}{n_i}e^{\beta\tilde{\alpha}_{i,t}}\big(\tilde{c}(\mathbb{E}\|\nabla R(\theta_t)\|+\sigma)+2LD_i\big) \qquad (19)\end{aligned}$$

where $\tilde{c}$ is an upper bound of $1+\beta\tilde{\alpha}_{i,t}(1+c)^{K_i+1}$, which is a constant and we use Lemma 13.

Combining (18) and (19) we have

$$\begin{aligned}\mathbb{E}\|\theta_{t+1}-\theta'_{t+1}\| &\leq \sum_{i=1}^m p_i(1+2\beta\tilde{\alpha}_{i,t})e^{\beta\tilde{\alpha}_{i,t}}\mathbb{E}\|\theta_t-\theta'_t\|+\frac{2}{n}\sum_{i=1}^m p_i\beta\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}\mathcal{E}_t \\ &\quad+\frac{2}{n}\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}\mathcal{E}_t+\frac{2}{n}\tilde{\alpha}_{i,t}e^{\beta\tilde{\alpha}_{i,t}}\big(\tilde{c}(\mathbb{E}\|\nabla R(\theta_t)\|+\sigma)+2LD_i\big). \qquad (20)\end{aligned}$$

Finally, under the choice of stepsize stated in the theorem, unrolling (20) over t and further using Theorem 1 together with the same techniques in the proof of Theorem 3, we complete the proof. □

D.3 Analysis for FedProx under non-convex losses

Lemma 14. *Suppose Assumptions 1,2 hold and assume that $\nabla^2_\theta l(\theta;z)\succ-\mu I$ with $\mu>0$. Considering FedProx with local update* (5)*, then for any $\eta_i\leq\frac{1}{\mu}$, we have for any $i\in[m]$*

$$\mathbb{E}\|\theta^i_{t+1}-\theta_t\|\leq\frac{\eta_i}{1-\eta_i\mu}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma),\quad\forall t=0,1,\ldots.$$

Proof. Recalling the local update (5) of FedProx and according to the first-order optimality condition, we have

$$\eta_i\nabla\hat{R}_{\mathcal{S}_i}(\theta^i_{t+1})+\theta^i_{t+1}-\theta_t=0.$$

Moreover, since the function $\eta_i\hat{R}_{\mathcal{S}_i}(\theta)+\frac{1}{2}\|\theta-\theta_t\|$ is $1-\eta_i\mu$-strongly-convex, we have

$$\|\theta^i_{t+1}-\theta_t\|\leq\frac{\eta_i}{1-\eta_i\mu}\|\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)\|$$

by combining the first-order optimality condition. Moreover, note that

$$\begin{aligned}\mathbb{E}\|\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)\| &\leq \mathbb{E}\|\nabla R_i(\theta_t)\|+\sigma \\ &\leq \mathbb{E}\|\nabla R(\theta_t)\|+\mathbb{E}\|\nabla R_i(\theta_t)-\nabla R(\theta_t)\|+\sigma \\ &\leq \mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma,\end{aligned}$$

where we use Lemma 1 and note

$$\mathbb{E}\|\nabla\hat{R}_{\mathcal{S}_i}(\theta_t)-\nabla R_i(\theta_t)\|\leq\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}\|\nabla l(\theta_t;z_{i,j})-\nabla R_i(\theta_t)\|\leq\sigma.$$

Thus, we have

$$\mathbb{E}\|\theta^i_{t+1}-\theta_t\|\leq\frac{\eta_i}{1-\eta_i\mu}(\mathbb{E}\|\nabla R(\theta_t)\|+2LD_i+\sigma),$$

which completes the proof. □

Lemma 15. *Suppose the assumptions stated in Lemma 14 hold and consider FedProx with local update* (5)*. Then, for any $i\in[m]$ and $j\in[n_i]$, we have*

$$\mathbb{E}\|\nabla l(\theta^i_{t+1};z_{i,j})\|\leq(1+\frac{\beta\eta_i}{1-\eta_i\mu})(2LD_i+\mathbb{E}\|\nabla R(\theta_t)\|+\sigma),\quad\forall t=0,1,\ldots.$$

Proof. For any $i \in [m]$ and time t,

$$\begin{aligned}
\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j})\| &\leq \mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla R_i(\theta_{t+1}^i)\| + \mathbb{E}\|\nabla R_i(\theta_{t+1}^i)\| \\
&\leq \mathbb{E}\|\nabla R_i(\theta_{t+1}^i)\| + \sigma \\
&\leq \mathbb{E}\|\nabla R_i(\theta_t)\| + \mathbb{E}\|\nabla R_i(\theta_{t+1}^i) - \nabla R_i(\theta_t)\| + \sigma \\
&\leq \beta\mathbb{E}\|\theta_{t+1}^i - \theta_t\| + \mathbb{E}\|\nabla R_i(\theta_t) - \nabla R(\theta_t)\| + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma \\
&\leq (1 + \frac{\beta\eta_i}{1-\eta_i\mu})(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma),
\end{aligned}$$

where we use Lemma 1 and Lemma 14 in the last step. □

Lemma 16. *Suppose f is non-convex, whose eigenvalues of its Hessian $\nabla^2 f$ are lower bounded by $-\mu$ with $0 < \mu < 1$, i.e., $\nabla^2 f(x) \succeq -\mu I, \forall x$. Define the proximal operator by*

$$\text{prox}_f(x) := \arg\min_y f(y) + \frac{1}{2}\|y - x\|^2.$$

Then, for any x_1, x_2, we have

$$\|\text{prox}_f(x_1) - \text{prox}_f(x_2)\| \leq \frac{1}{1-\mu}\|x_1 - x_2\|.$$

Proof. Let $u_1 = \text{prox}_f(x_1)$ and $u_2 = \text{prox}_f(x_2)$. According to the first-order optimality condition, we have

$$\begin{aligned}
\nabla f(u_1) + u_1 - x_1 &= 0 \\
\nabla f(u_2) + u_2 - x_2 &= 0
\end{aligned}$$

Since $\nabla^2 f$ has eigenvalues greater than $-\mu$, we further have

$$\begin{aligned}
-\mu\|u_1 - u_2\|^2 &\leq \langle \nabla f(u_1) - \nabla f(u_2), u_1 - u_2\rangle \\
&= \langle x_1 - u_2 - (x_2 - u_2), u_1 - u_2\rangle \\
&= \langle x_1 - x_2, u_1 - u_2\rangle - \|u_1 - u_2\|^2
\end{aligned}$$

and hence

$$(1-\mu)\|u_1 - u_2\|^2 \leq \langle x_1 - x_2, u_1 - u_2\rangle \leq \|x_1 - x_2\|\|u_1 - u_2\|$$

which means

$$\|u_1 - u_2\| \leq \frac{1}{1-\mu}\|x_1 - x_2\|.$$

□

Theorem 11 (FedProx part of Theorem 4)**.** *Suppose Assumptions 1-3 hold and consider FedProx (Algorithm 3). Assume that all eigenvalues of the Hessian of $l(\cdot; z)$ are strictly greater than $-\mu$ with $\mu > 0$ for any z. With $\eta_i \leq \frac{\delta_t}{\mu}$ for $0 < \delta < 1$ being diminishing at the order of $\mathcal{O}(c/t)$ (where $c > 0$). Then,*

$$\epsilon_{gen} \leq \tilde{\mathcal{O}}\Big(\frac{T^c}{n}D_{max}\Big) + \mathcal{O}\Big((\Delta_0\tilde{D})^{\frac{1}{2}}\frac{T^{\frac{3}{4}+c}}{n} + \sqrt{\Delta_0}\frac{T^{\frac{1}{2}+c}}{n}\Big) + \tilde{\mathcal{O}}\Big(\frac{T^c}{n}\sigma\Big),$$

where $\Delta_0 := \mathbb{E}[R(\theta_0) - R(\theta^*)]$.

Proof. Denoting $\text{prox}_f(x) := \arg\min_y f(y) + \frac{1}{2}\|y - x\|^2$, we can rewrite the local update (5) as

$$\theta_{t+1}^i = \text{prox}_{\eta_i\hat{R}_{\mathcal{S}_i}}(\theta_t).$$

There are two different cases for local updates. For client j with $j \neq i$, we note $\hat{R}_{\mathcal{S}_j}(\cdot) = \hat{R}_{\mathcal{S}_j'}(\cdot)$ in the sense that there is no perturbation for client j. In this case, using Lemma 16 we obtain

$$\begin{aligned}
\|\theta_{t+1}^i - (\theta_{t+1}^i)'\| &= \|\text{prox}_{\eta_i\hat{R}_{\mathcal{S}_i}}(\theta_t) - \text{prox}_{\eta_i\hat{R}_{\mathcal{S}_i}}(\theta_t')\| \\
&\leq \frac{1}{1-\eta_i\mu}\|\theta_t - \theta_t'\| \\
&\leq \frac{1}{1-\delta}\|\theta_t - \theta_t'\|
\end{aligned}$$

For client i, we note that $\hat{R}_i(\cdot) - \hat{R}_i'(\cdot) = \frac{1}{n_i}(l(\cdot; z_{i,j}) - l(\cdot; z'_{i,j}))$, where $z'_{i,j}$ is the perturbed data point. And we also use $\hat{R}_i$ and $\hat{R}_i'$ to represent $\hat{R}_{\mathcal{S}_i}$ and $\hat{R}_{\mathcal{S}_i'}$ for simplicity. Then, we have

$$\begin{aligned}\theta_{t+1}^i &= \arg\min_\theta \eta_i \hat{R}_i(\theta) + \frac{1}{2}\|\theta - \theta_t\|^2 \\ (\theta_{t+1}^i)' &= \arg\min_\theta \eta_i \hat{R}_i'(\theta) + \frac{1}{2}\|\theta - \theta_t'\|^2.\end{aligned}$$

According to the first-order optimality condition, it yields

$$\begin{aligned}\theta_{t+1}^i - \theta_t &= -\eta_i \nabla \hat{R}_i(\theta_{t+1}^i) \\ (\theta_{t+1}^i)' - \theta_t' &= -\eta_i \nabla \hat{R}_i'((\theta_{t+1}^i)') \\ &= -\eta_i \nabla \hat{R}_i((\theta_{t+1}^i)') + \frac{\eta_i}{n_i}\left(\nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l((\theta_{t+1}^i)'; z'_{i,j})\right).\end{aligned}$$

Moreover, by the techniques used in Lemma 16,

$$\begin{aligned}\|(\theta_{t+1}^i)' - \theta_{t+1}^i\|^2 &\le \langle \theta_t' - \theta_t, (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle - \eta_i\langle \nabla \hat{R}_i((\theta_{t+1}^i)') - \nabla \hat{R}_i(\theta_{t+1}^i), (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle \\ &\quad + \frac{\eta_i}{n_i}\langle (\theta_{t+1}^i)' - \theta_{t+1}^i, \nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l(\theta_{t+1}^i; z'_{i,j})\rangle \\ &\le \langle \theta_t' - \theta_t, (\theta_{t+1}^i)' - \theta_{t+1}^i\rangle + \frac{\eta_i}{n_i}\langle (\theta_{t+1}^i)' - \theta_{t+1}^i, \nabla l((\theta_{t+1}^i)'; z_{i,j}) - \nabla l(\theta_{t+1}^i; z'_{i,j})\rangle \\ &\quad + \eta_i\mu\|(\theta_{t+1}^i)' - \theta_{t+1}^i\| \qquad (21)\end{aligned}$$

which further implies by symmetry of $z_{i,j}$ and $z'_{i,j}$,

$$\begin{aligned}\|(\theta_{t+1}^i)' - \theta_{t+1}^i\| &\le \frac{1}{1-\eta_i\mu}\|\theta_t' - \theta_t\| + \frac{\eta_i}{n_i}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla l(\theta_{t+1}^i; z'_{i,j})\| \\ &\le \frac{1}{1-\delta}\|\theta_t' - \theta_t\| + \frac{\eta_i}{n_i}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla l(\theta_{t+1}^i; z'_{i,j})\|\end{aligned}$$

where we also use Cauchy-Schwatz inequality.

Combining two cases gives

$$\begin{aligned}\mathbb{E}\|\theta_{t+1} - \theta_{t+1}'\| &\le \sum_{j=1}^m p_j \mathbb{E}\|\theta_{t+1}^j - (\theta_{t+1}^j)'\| \\ &\le \frac{1}{1-\delta}\mathbb{E}\|\theta_t - \theta_t'\| + \frac{\eta_i}{n}\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j}) - \nabla l(\theta_{t+1}^i; z'_{i,j})\|, \\ &\le \frac{1}{1-\delta}\mathbb{E}\|\theta_t - \theta_t'\| + \frac{2\eta_i}{n}\mathbb{E}\|\nabla l(\theta_{t+1}^i; z_{i,j})\| \\ &\le \frac{1}{1-\delta}\mathbb{E}\|\theta_t - \theta_t'\| + \frac{2\eta_i}{n}(1 + \frac{\beta\delta}{(1-\delta)\mu})(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma),\end{aligned}$$

where we use Lemma 15 in the last step. Define $\tau := \frac{\delta}{1-\delta}$. Then, unrolling it over t we obtain

$$\mathbb{E}\|\theta_T - \theta_T'\| \le T^c \frac{2}{n}\sum_{t=0}^{T-1}\eta_i(1 + \beta\tau/\mu)(2LD_i + \mathbb{E}\|\nabla R(\theta_t)\| + \sigma). \qquad (22)$$

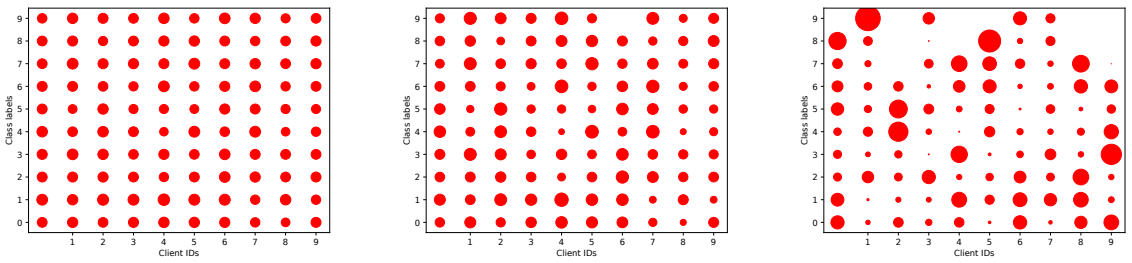
Finally, based on (22), combining Theorem 1 and using the proof techniques in Theorem 3, we complete the proof. □

E Additional experiments

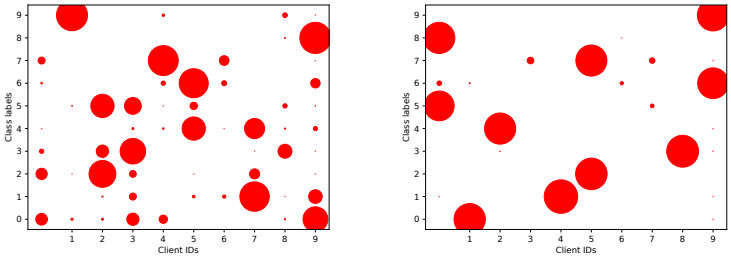
The implementation of the experiments in this section and Section 6 are based on [46] and can be found through the following link: https://github.com/fedcodexx/Generalization-of-Federated-Learning.

In addition to the experimental setup outlined in Section 6, we explore an alternative approach for distributing the MNIST dataset to clients, exhibiting varying levels of data heterogeneity. As explained in [40], we generate disjoint heterogeneous training data for clients by leveraging the Dirichlet distribution. The value of α in the Dirichlet distribution controls the level of data heterogeneity. A smaller value of α means a higher probability of clients possessing examples from a single randomly selected class. In particular, setting $\alpha = 100$ mimics identical local data distributions. We compare heterogeneity levels with $\alpha = 100, 10, 1, 0.1, 0.01$. Figure 2 presents the generated number of samples per class assigned to each client across varying values of α. This depiction highlights different levels of data heterogeneity among the clients. In each setting, we implement the algorithms using the same way outlined in Section 6. Figure 3 presents the obtained results. These results align closely with those in Section 6. We highlight that the results in Figure 3 visually verify the theoretical findings from Section 5. In particular, increasing data heterogeneity leads to larger generalization errors.

Moreover, we run additional experiments to show how other parameters like stepsizes affect generalization errors. We implement the three algorithms using the experimental setup in Section 6 with $\rho = 0.8$. Figure 4 presents the obtained results. These results show that larger stepsizes lead to larger generalization errors for each algorithm, which verifies our theoretical findings in Section 5.

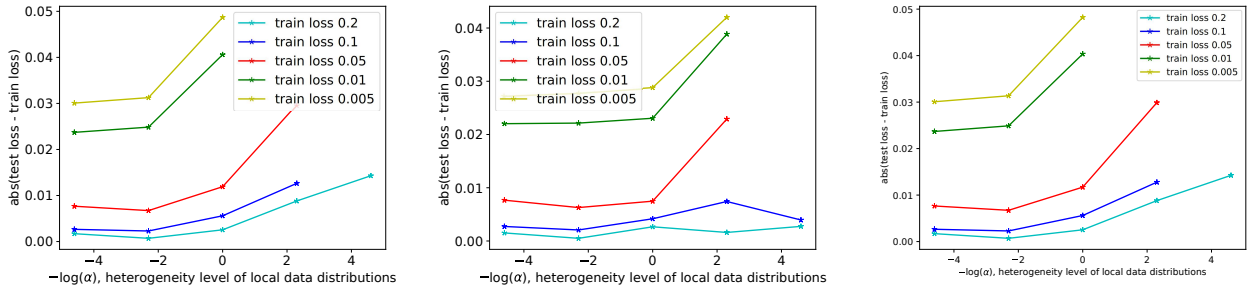


(a) $\alpha = 100$ (b) $\alpha = 10$ (c) $\alpha = 1$



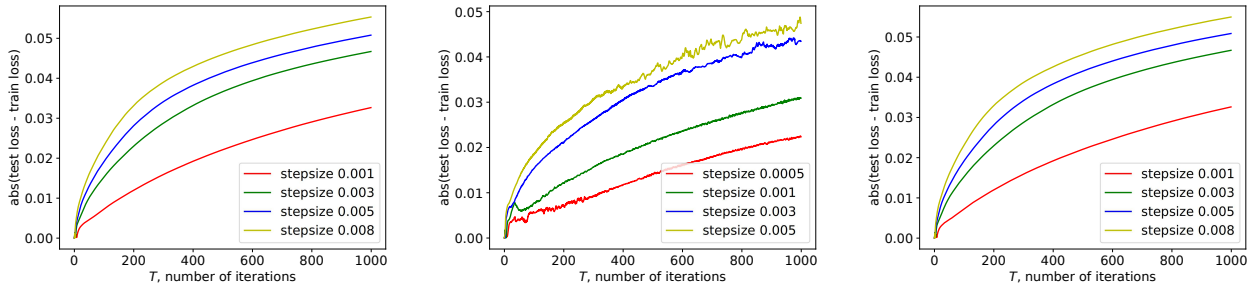
(d) $\alpha = 0.1$ (e) $\alpha = 0.01$

Figure 2: Number of samples per class assigned to each client (indicated by dot sizes), for different Dirichlet distribution α values with larger α corresponding to higher heterogeneity.



(a) FedAvg (b) SCAFFOLD (c) FedProx

Figure 3: Generalization errors of FedAvg, SCAFFOLD, and FedProx (local datasets by Dirichlet distribution).



(a) FedAvg (b) SCAFFOLD (c) FedProx

Figure 4: Generalization errors of FedAvg, SCAFFOLD, and FedProx for different stepsizes.