RETHINKING CLIENT REWEIGHTING FOR SELFISH FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most federated learning (FL) algorithms aim to learn a model which achieves optimal overall performance across all clients. However, for some clients, the model obtained by conventional federated training may perform even worse than that obtained by local training. Therefore, for a stakeholder who only cares about the performance of a few *internal clients*, the outcome of conventional federated learning may be unsatisfactory. To this end, we study a new *selfish* variant of federated learning, in which the ultimate objective is to learn a model with optimal performance on internal clients *alone* instead of all clients. We further propose Variance Reduction Selfish Learning (VaRSeL), a novel algorithm that reweights the external clients based on variance reduction for learning a model desired in this setting. Within each round of federated training, it guides the model to update towards the direction favored by the internal clients. We give a convergence analysis for both strongly-convex and non-convex cases, highlighting its fine-tune effect. Finally, we perform extensive experiments on both synthesized and real-world datasets, covering image classification, language modeling, and medical image segmentation. Experimental results empirically justify our theoretical results and show the advantage of VaRSeL over related FL algorithms.

1 INTRODUCTION

Federated learning (FL) proposes a privacy-preserving scheme, which enables different clients to cooperatively learn a global model by integrating knowledge from multiple clients. Suppose there are Q clients $\{\mathcal{D}_n\}_{n=1}^{Q}$ in total. A general objective of FL is given by

$$\min_{f} \sum_{n=1}^{Q} \mathcal{L}_{\mathcal{D}_n}(f), \tag{1}$$

where f is the model, and $\mathcal{L}_{\mathcal{D}_n}$ is the generalization loss on client $\mathcal{D}_n$.

However, since the general objective (1) fairly optimizes the mean generalization loss across all clients, a model learned by (1) may not perform well on some individual clients. We implemented several state-of-the-art FL algorithms on FeTS 2021 (Pati et al., 2021), a real-world FL dataset for medical image segmentation, and discovered that *the federally learned model is not guaranteed to generalize on local client(s) at least as well as the locally learned model.* (see Figure 1)

In cross-silo FL (Kairouz et al., 2019), this issue becomes particularly serious since clients typically have sufficient data for training a decent local model. In this case, the clients may become reluctant to participate in federated training. This issue raises an important challenge for FL society as well as practitioners: *For a stakeholder who is only interested in the model's performance on a small portion of clients, is it possible to guarantee its benefits from federated learning?*

This is a common challenge in many practical cases. For instance, in a bank network, due to the limited size of internal datasets, a banking group wants to leverage more data and learn a better model through federated learning, while the banking group expects the final model to maximize the performance on its own internal banks, rather than overfitting to other external banks. As another example, to build an AI-assisted diagnostic system, a medical association hopes to integrate knowledge from the datasets of external partner hospitals through federated learning. As the system ultimately serves

its internal hospitals/patient populations, its performance should not be disturbed by external partner hospitals (Rieke et al., 2020). In these cases, to enhance the model perf the stakeholder would like to purchase gradient updates from external collaborators.

Selfish federated learning In this work, we resolve the aforementioned challenge by investigating a novel learning setting called *selfish federated learning* (Selfish-FL), where the stakeholder who only cares about the performance on a few internal clients can reap the benefits of cooperative learning with external clients. To formalize, the Q clients are partitioned into M *internal clients* and N *external clients* ($M + N = Q$, $M \ll N$). Though it is possible to incorporate the knowledge in all Q clients for model training, the stakeholder only focuses on the model's performance on the M internal clients.

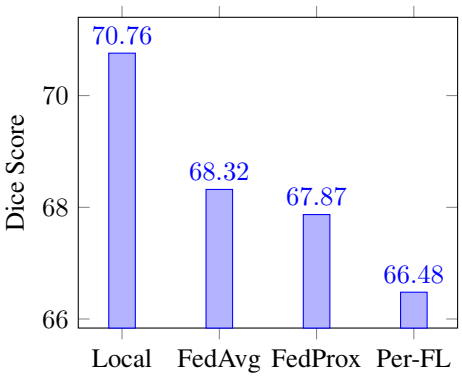

Figure 1: The Dice score of FL methods and local training on 1^{st} site in FeTS 2021 for brain tumor segmentation. FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020b), and Per-FL (Fallah et al., 2020) are different FL variants.

Comparisons with related setups Table 1 lists several FL settings. While centralized and decentralized FL differ in communication network, both of them are looking for a global consensus among all the clients (Wang et al., 2021; Koloskova et al., 2020). Personalized-FL allows each client to maintain a personalized model, but the losses of *all clients* are still considered in the overall objectives (Fallah et al., 2020; Tan et al., 2021). In Selfish-FL, only internal clients are the concern of the stakeholder, while the role of external clients is to provide additional information to further improve the model's performance on internal clients.

Table 1: Comparisons among related federated learning settings.

Learning Setting	Objective	Comments
Centralized-FL	$\sum_{n=1}^{M+N} \mathcal{L}_{\mathcal{D}_n}(f)$	Server has access to clients' updates $\{g_n\}_{n=1}^{M+N}$.
Decentralized-FL	$\sum_{n=1}^{M+N} \mathcal{L}_{\mathcal{D}_n}(f)$	Client n has access S_n and $\{f_k\}_{k\in\mathcal{N}(n)}$; f_n's finally converge to a global consensus.
Personalized-FL	$\sum_{n=1}^{M+N} \mathcal{L}_{\mathcal{D}_n}(f_n)$	f_n's may not converge to a global consensus.
Selfish-FL	$\sum_{i=1}^{M} \mathcal{L}_{\mathcal{D}_i}(f)$	**Internal clients** $1, \cdots, M$ **have access to datasets** $S_1, \cdots, S_M$ **and updates** $\{g_j\}_{j=M+1}^{M+N}$.

Our contributions The main contributions in this paper can be summarized as four aspects: First, we study a novel setting in federated learning, namely "Selfish-FL", which is the first work (to the best of our knowledge) considering biased objectives and internal/external partitions in FL setting.

Second, we propose a novel algorithm, called VaRSeL, for solving Selfish-FL problem. It leverages knowledge from external clients to improve the model via gradients variance reduction. This is a different perspective compared with previous loss-based clients reweighting strategy (Cho et al., 2020; Chen et al., 2020). Within each round of federated training, it guides the model to update towards the direction favored by the internal clients. Moreover, our strategy is free of hyperparameter-tuning.

Third, we present a convergence analysis of our reweighting strategy for this new setting, both in strongly-convex cases and non-convex cases. Based on the convergence result, we also observe its fine-tune effect.

Finally, we perform extensive experiments on both synthesized and real-world natural language and medical image analysis datasets. The experimental results demonstrate that our proposed VaRSeL

outperforms the classical FedAvg, related client selection methods, as well as the state-of-the-art for non-iid data (FedProx) and personalized FL algorithm (Per-FL) in terms of model accuracy.

2 RELATED WORKS

2.1 FEDERATED LEARNING AND CHALLENGES

Federated learning (FL) (Yang et al., 2019; Kairouz et al., 2019; Wang et al., 2021) provides a promising way of cooperative learning with privacy protection. In this subsection, we will briefly summarize a few important lines of FL (Some works may lie at intersection):

Relieving clients' non-iid Proximal terms (Li et al., 2020c), cluster-based methods (Briggs et al., 2020; Ghosh et al., 2020), client-drift regulations (Karimireddy et al., 2020; Acar et al., 2021) are some leading methods to address the heterogeneity of clients' data distributions. At a high level, all these methods can be summarized as pursuing a compromise solution that can work fairly well on all the clients. Therefore, even though these methods are very effective in classical FL frameworks, none of them fits well into our selfish-FL framework, because the internal clients are in nature reluctant to compromise with any external clients.

Personalized FL Many approaches have been proposed for achieving personalization in FL setting (Kairouz et al., 2019), *e.g.*, mixing the global model and user's local model, dividing the network into base and personalized layers. Recently, Per-FL (Fallah et al., 2020) linked personalized FL with meta-learning by building a good initial meta-model that can be updated effectively. Dinh et al. (2020) used Moreau envelopes as clients' regularization loss to achieve personalized FL and Zhang et al. (2020) proposed a flexible FL framework that allows clients to personalize to specific target data distributions. Readers can refer to comprehensive surveys in Li et al. (2020a); Tan et al. (2021). However in personalized FL, the losses of *all clients* are still considered in the overall objectives.

Alleviating communication cost Lin et al. (2017); Richtárik et al. (2021) provided several model compression schemes to reduce communication cost and proved the convergence of their compression methods. Liang et al. (2020); Li et al. (2021) proposed to learn a combination of local and global models, while only the global parts are synchronized. It's worth mentioning that our method is orthogonal to these communication-efficient methods and can be combined with theirs directly.

2.2 SAMPLE REWEIGHTING.

Our work is also related to sample reweighting, *i.e.*, assigning different weights to samples in model training to tackle dataset biases. There are two completely opposite strategies on how to set the weights $\mathbf{w}$ for external datasets. One line of works Freund et al. (1996); Malisiewicz et al. (2011); Cho et al. (2020) argues that the samples with higher losses should be given more weights, since these *difficult samples* can help sharpen the decision boundaries and increase the adversarial robustness. The other line of works Ghadikolaei et al. (2019); Song et al. (2021) suggests lowering their weights, since their high losses are attributed to training set biases or mislabels (*e.g.*, *nosiy samples*). In fact, this contradiction directly points to a fundamental, difficult problem in machine learning:

Dilemma. *How to distinguish the difficult samples from the noisy ones?*

To the best of our knowledge, the controversy continues, and people are still far from getting a clear resolution. To bypass the aforementioned dilemma, some researchers (Ren et al., 2018; Shu et al., 2019) used meta-learning approach to learn how to assign weights $\mathbf{w}$ adaptively. However, their method suffers from computationally expensive second derivatives estimation, and can hardly generalize to federated learning framework since it requires full-client participation throughout training.

Client reweighting in FL In FL framework, since one only has access to limited training examples, clients instead of samples are reweighted to tackle their heterogeneity. Existing FL algorithms reweight clients using the sizes of their datasets (Li et al., 2020c), the numbers of their local steps (Wang et al., 2020), their local losses (Cho et al., 2020; Chen et al., 2020), etc.

As an inevitable issue in FL, communication costs usually restrict the number of clients one can access within each communication round (McMahan et al., 2017). Suppose client j is sampled with probability $\mathbf{w}_j \in [0,1]$, $M+1 \leq j \leq M+N$. Since the communication costs mainly depend on the number of sampled clients, existing algorithms (Lian et al., 2018; Wang et al., 2019) impose hard constraints on $\mathbf{w}_j$ as $\mathbf{1}^{\mathrm{T}}\mathbf{w} = \sum_{j=M+1}^{M+N} \mathbf{w}_j \leq C_b N = K$, where $C_b > 0$ is called the *communication budget*, to alleviate the communication bottleneck.

3 PROBLEM FORMULATION

In this section, we give a formal definition of the setting *selfish federated learning* (Selfish-FL).

Internal and external clients Let $\mathcal{X}$ denote the input space (*e.g.*, feature space) and $\mathcal{Y}$ the space of outputs (*e.g.*, label space). Each client characterizes a joint distribution $\mathcal{D} = P(X,Y)$ over $\mathcal{X} \times \mathcal{Y}$. Suppose there are $M+N$ clients $\{\mathcal{D}_n\}_{n=1}^{M+N}$, among which the first M clients are called the *internal clients* and the remaining N clients are called the *external clients* ($M \ll N$). We assume that internal clients are consistent and the difference between internal and external clients are expected to be larger than that between internal clients. To formalize, we have

$$\mathcal{D}_1 = \cdots = \mathcal{D}_M \triangleq \mathcal{D}_{\mathcal{I}};$$
$$\mathcal{D}_j \neq \mathcal{D}_{\mathcal{I}},\ M+1 \leq j \leq M+N.$$

In practice, the distributions of clients are not observed directly, but given in the form of sample sets $S_n \sim \mathcal{D}_n$, for $1 \leq n \leq M+N$. The stakeholder has *direct access* to the raw data in $S_1, \ldots, S_M$, but can only query external clients in a *privacy-preserving* way as in typical FL scenario.

Ultimate objective While conventional federated learning setting usually aims to minimize the generalization loss across all clients, the objective in Selfish-FL is to minimize the generalization loss on internal clients $\mathcal{D}_1, \cdots, \mathcal{D}_M$ alone:

$$\min_{\theta}\ l(\theta) := \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \mathcal{L}(f_\theta(x), y), \tag{2}$$

where f_θ denotes a model f parameterized by θ, and $\mathcal{L}$ denotes an appropriate loss function. Although the stakeholder can simply learn θ by local training on datasets of internal clients, we will show both in theory and in experiments that it is probable to improve the estimation of θ by leveraging additional knowledge from external clients without sharing raw data.

4 VARSEL: VARIANCE REDUCTION SELFISH LEARNING

4.1 REWEIGHTING VIA VARIANCE REDUCTION

In Selfish-FL, we incorporate the knowledge of external clients to further improve the performance of model f_θ on internal clients. Since external clients are of uneven qualities, it's natural to assign them different weights and formulate a *surrogate objective* as

$$l(\theta; \mathbf{w}) = \frac{\sum_{i=1}^{M} l_i(\theta) + \sum_{j=M+1}^{M+N} \mathbf{w}_j l_j(\theta)}{M + \mathbf{1}^{\mathrm{T}}\mathbf{w}}, \tag{3}$$

where $\mathbf{w} \in [0,1]^N$, $\mathbf{1}^{\mathrm{T}}\mathbf{w} \leq K$, and $l_n(\theta) = \frac{1}{|S_n|}\sum_{(x,y)\in S_n} \mathcal{L}(f_\theta(x), y)$ is the empirical loss on client n. As a result, how to set the weights $\mathbf{w}$ is a key issue in designing Selfish-FL algorithms. In this paper, to bypass the dilemma in section 2.2, we consider the client reweighting strategy from a different perspective compared with previous works:

> **Strategy.** *We're setting* $\mathbf{w}$ *neither to reduce losses nor to increase losses, but to make our surrogate objective (Eq.* (3)*) simulate the ultimate objective (Eq.* (2)*).*

Since the majority of optimizers applied in deep learning are first-order methods (*e.g.*, SGD, Adam) (Wang et al., 2021), we consider making the estimated gradient as *precise* as possible. Specifically,

the goal is to reduce the variance (denoted as $\Phi(\mathbf{w};\theta)$) between the estimated, surrogate gradient $\nabla l(\theta;\mathbf{w})$ and the unseen, ultimate gradient $\nabla l(\theta)$[1]:

$$\begin{aligned}\Phi(\mathbf{w};\theta) &= \mathbb{E}\|\nabla l(\theta;\mathbf{w}) - \nabla l(\theta)\|^2 \\ &= \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\cdot\mathbb{E}\Big\|M(\nabla l_{\mathcal{I}}(\theta)-\nabla l(\theta)) + \sum_{j=M+1}^{M+N}\mathbf{w}_j(\nabla l_j(\theta)-\nabla l(\theta))\Big\|^2,\end{aligned}$$

where $l_{\mathcal{I}}(\theta) = \frac{1}{M}\sum_{i=1}^{M} l_i(\theta)$. Since the internal gradients are unbiased estimators of the ultimate gradients, we have

$$\mathbb{E}[\nabla l_i(\theta)] = \mathbb{E}[\nabla l_{\mathcal{I}}(\theta)] = \nabla l(\theta), \quad 1 \le i \le M.$$

Thus, the variance $\Phi(\mathbf{w};\theta)$ can be simplified as follows:

$$\Phi(\mathbf{w};\theta) = \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\left[\sum_{i=1}^{M}\mathbb{E}\|\nabla l_i(\theta)-\nabla l(\theta)\|^2 + \mathbb{E}\Big\|\sum_{j=M+1}^{M+N}\mathbf{w}_j\big(\nabla l_j(\theta)-\nabla l(\theta)\big)\Big\|^2\right]. \quad (4)$$

The optimal weights $\mathbf{w}_*$ are thus chosen to minimizes the variance in Equation (4):

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \Phi(\mathbf{w};\theta).$$

In this way, we can head towards the optimal direction for the ultimate objective.

4.2 Convergence analysis

In this section, we give a convergence analysis for our reweighting strategy. Our proofs mainly follow Li et al. (2020c), and we refer readers to Appendix A and B for the details of the proofs.

As in Schmidt & Roux (2013); Li et al. (2020b); Vaswani et al. (2019), we assume bounded dissimilarity on internal datasets. For external datasets, we take a weaker assumption by allowing non-vanishing noises (ν_j) near the optimal point. In general, our assumptions are weaker than the constantly-bounded noise assumptions since the noises are allowed to grow with the gradient norms.

Assumption 1 *For internal datasets, we assume: for some constant $\sigma > 0$,*

$$\mathbb{E}\|\nabla l_{\mathcal{I}}(\theta) - \nabla l(\theta)\|^2 \le \sigma\|\nabla l(\theta)\|^2;$$

For external datasets, we take weaker assumptions: for each $M+1 \le j \le M+N$, for some constants $\kappa_j, \nu_j > 0$,

$$\mathbb{E}\|\nabla l_j(\theta) - \nabla l(\theta)\|^2 \le \kappa_j\|\nabla l(\theta)\|^2 + \nu_j.$$

Lemma 1. *Under Assumption 1, there exist constants $A, B > 0$, such that*

$$\Phi(\mathbf{w}_*;\theta) \le A\|\nabla l(\theta)\|^2 + B.$$

To ensure convergence, we further assume that the objective functions are well-conditioned.

Assumption 2a *The ultimate objective function is L-smooth ($L > 0$):*

$$\|\nabla l(\theta_1) - \nabla l(\theta_2)\| \le L\|\theta_1 - \theta_2\|.$$

Assumption 2b *The empirical loss functions are L-smooth ($L > 0$):*

$$\|\nabla l_n(\theta_1) - \nabla l_n(\theta_2)\| \le L\|\theta_1 - \theta_2\|, \quad 1 \le n \le M+N.$$

Assumption 3 *The ultimate objective is μ-strongly convex ($\mu > 0$):*

$$\|\nabla l(\theta_1) - \nabla l(\theta_2)\| \ge \mu\|\theta_1 - \theta_2\|.$$

[1]The "variance reduction" here is essentially different from that in conventional FL. In conventional FL, it reduces the variance between all clients to speedup the global convergence; in selfish-FL, it can be seen as reducing the variance between internal and external clients, thus alleviating the disturbance of external noises.

ASSUMPTION 4 *The estimated gradient norms are uniformly bounded:*

$$\mathbb{E}\|\nabla l_n(\theta)\|^2 \leq G^2, \quad 1 \leq n \leq M+N.$$

To gain some insights, we start our analysis from a well-conditioned case and assume local steps $\tau = 1$. Using the bound of Lemma 1, we can derive Theorem 2 on convergence rate. The first term on the right hand side of Eq. (5) is called *optimization term*, and the second is called *noise term*. The convergence rate is bounded by the sum of the two terms.

Theorem 2 (Basic Convergence). *Let $\theta^{t+1} = \theta^t - \eta\nabla l(\theta^t; \mathbf{w}_*^t)$, $0 \leq t \leq T-1$. Under Assumption 1, Assumption 2a and Assumption 3, for $\eta < \frac{\frac{1}{L}-\frac{A}{\mu}}{2\sqrt{A}}$, we have*

$$\mathbb{E}\|\theta^* - \theta^t\|^2 \leq \underbrace{\left(1-\eta\Xi_1\right)^t \mathbb{E}\|\theta^* - \theta^0\|^2}_{\textit{optimization term}} + \underbrace{\Xi_2 B}_{\textit{noise term}}, \tag{5}$$

where Ξ_1 and Ξ_2 are constants,

The basic convergence result (Eq. (5)) provides useful insights for model training. At the beginning of training, due to random initialization, the optimization term can be arbitrarily large. But as T increases, the optimization term decays in linear rate. Consequently, the dominant term will be gradually shifted to noise term, which is mainly determined by the noises in external datasets. Based on this observation, when training progresses to later stage, in order to get more precise gradients, it will gradually switch back to internal datasets and fine-tune the trained model at the end of learning.

Taking multiple steps between communication rounds is an entrenched convention in communication efficient federated learning (McMahan et al., 2017; Kairouz et al., 2019). Hence we give a more general convergence analysis, in which local steps $\tau \geq 1$. We formalize it as follows:

$$\begin{aligned}
&\theta_n^t = \theta^t, \\
&\theta_n^{t+\frac{k+1}{\tau+1}} = \theta_n^{t+\frac{k}{\tau+1}} - \eta\nabla l_n(\theta_n^{t+\frac{k}{\tau+1}}), \quad 1 \leq n \leq M+N, \text{ for } 0 \leq k \leq \tau-1, \\
&\theta^{t+1} = \overline{\theta^{t+\frac{\tau}{\tau+1}}}_{\mathbf{w}},
\end{aligned}$$

where $\overline{\theta}_{\mathbf{w}} = \frac{\sum_{i=1}^{M}\theta_i + \sum_{j=M+1}^{M+N}\mathbf{w}_j\theta_j}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}$.

Theorem 3 (Main Convergence). *For $\tau \geq 1$, under Assumption 1, Assumption 2a, Assumption 2b and Assumption 4, for $\eta < \frac{2(1-\sqrt{A})}{L(A+\sqrt{A})}$, we have*

$$\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left\|\nabla l\left(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}_*^t}\right)\right\|^2 \leq \frac{l(\theta^0)-l(\theta^*)}{\eta\Gamma_3\tau T} + \frac{\Gamma_1\Gamma_2}{\Gamma_3\tau},$$

where Γ_1, Γ_2 and Γ_3 are constants.

Algorithm 1 Variance **R**eduction **Se**lfish **L**earning (**VaRSeL**)

Input: learning rate η, communication rounds T.

Output: final model θ^{final}.

1: Initialize model θ^0.
2: **for** $t = 0, 1 \cdots, T-1$ **do**
3: Compute internal gradients $\nabla l_1(\theta^t), \cdots, \nabla l_M(\theta^t)$, and then $\nabla l_{\mathcal{I}}(\theta^t)$.
4: Broadcast $(\theta^t, \nabla l_{\mathcal{I}}(\theta^t))$ to clients $[M+1, M+N]$.
5: **for** $j \in [M+1, M+N]$ **in parallel do**
6: Compute gradient update $\nabla l_j(\theta^t)$.
7: Send back $\|\nabla l_j(\theta^t) - \nabla l_{\mathcal{I}}(\theta^t)\|$.
8: **end for**
9: Solve $\tilde{\mathbf{w}}$ according to Equation (8).
10: Sample a subset of clients $\mathcal{S}^t$ according to $\tilde{\mathbf{w}}$.
11: **for** $j \in \mathcal{S}^t$ **in parallel do**
12: Send back $\nabla l_j(\theta^t)$.
13: **end for**
14: Solve $\mathbf{w}_*$ according to Equation (4).
15: $\Delta^t = \frac{M\nabla l_{\mathcal{I}}(\theta^t) + \mathbf{w}_*^{\mathrm{T}}\left[\nabla l_j(\theta^t)\right]_{j\in\mathcal{S}^t}}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}_*}$
16: $\theta^{t+1} := \theta^t - \eta\Delta^t$
17: **end for**
18: **return** $\theta^{\text{final}} := \theta^t$

4.3 VARIANCE REDUCTION SELFISH LEARNING

Based on our above reweighting strategy, we design an algorithm, called **Va**riance **R**eduction **Se**lfish **L**earning (VaRSeL). It is free of hyperparameter-tuning, and moreover, as shown in experiments, it will automatically fine-tune the model when learning progresses to later stage. VaRSeL is not exactly the reweighting strategy in section 4.1, but an approximation algorithm to the reweighting strategy,

The main idea of VaRSeL is to select the external collaborators wisely within each communication round. Within iteration t, suppose the central server has collected the gradient updates on internal datasets. Then we can estimate the ultimate gradient unbiasedly:

$$\mathbb{E}[\nabla l_{\mathcal{I}}(\theta^t)] = l(\theta^t). \tag{6}$$

Communication bottleneck is a major obstacle that prevents us from accurately solving the optimal weights $\mathbf{w}_*$ in Equation (4). To this end, we propose a heuristic:

$$\begin{aligned}\Phi(\mathbf{w};\theta) &= \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\left[\sum_{i=1}^{M}\mathbb{E}\|\nabla l_i(\theta)-\nabla l(\theta)\|^2 + \mathbb{E}\Big\|\sum_{j=M+1}^{M+N}\mathbf{w}_j\big(\nabla l_j(\theta)-\nabla l(\theta)\big)\Big\|^2\right] \\ &\le \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\left[\sum_{i=1}^{M}\mathbb{E}\|\nabla l_i(\theta)-\nabla l(\theta)\|^2 + \mathbf{1}^{\mathrm{T}}\mathbf{w}\sum_{j=M+1}^{M+N}\mathbf{w}_j\underbrace{\mathbb{E}\Big\|\nabla l_j(\theta)-\nabla l(\theta)\Big\|^2}_{\text{communicated term}}\right] \triangleq \tilde{\Phi}(\mathbf{w};\theta).\end{aligned} \tag{7}$$

Before the external clients send back their updates, we first estimate an *approximate weight*:

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \tilde{\Phi}(\mathbf{w};\theta). \tag{8}$$

This approximate weight tightens the upper bound in (7) and can be computed very efficiently, since only a few bytes of communication are required. Then, we sample clients according to $\tilde{\mathbf{w}}$, and after receiving their updates, we can solve a more accurate $\mathbf{w}_*$ according to Equation (4).

The pseudo-code of VaRSeL is given in Algorithm 1. This is a simple implementation when local steps τ equals 1. In fact, τ is allowed to be greater than 1, and we almost don't modify any other parts of the algorithm, except to replace the gradient update with the accumulated model update.

5 EXPERIMENTS

In this section, we present experimental results on both synthesized benchmark datasets and real-world datasets. For synthesized benchmark experiments, we mainly use them to support the claims made in our paper. For real-world dataset experiments, we show the real-world performance of our method and its higher accuracy over the other competitors.

We compared VaRSeL with seven existing methods mentioned in section 2:

- **Baseline methods**, including *Local Training* (*i.e.*, training the model with internal datasets) and the original *FedAvg* (McMahan et al., 2017);
- **Non-iid relief methods**, *FedProx* (Li et al., 2020b);
- **Reweighting methods**, including *Skew* (Cho et al., 2020), *SL* (Song et al., 2021), and *OCS* (Chen et al., 2020);
- **Personalized FL methods**, the MAML-based *Per-FL* (Fallah et al., 2020).

5.1 EXPERIMENTS ON SYNTHESIZED BENCHMARK DATASETS

Dataset and settings We conducted experiments on three benchmark datasets, Fashion-MNIST, EMNIST, and CIFAR-10 (Xiao et al., 2017; Cohen et al., 2017; Krizhevsky, 2009) to demonstrate the robustness and fine-tune effect of our methods. The datasets were preprocessed in similar ways as McMahan et al. (2017). We used a logistic model for Fashion-MNIST, a 2NN model for EMNIST, and a CNN model for CIFAR-10. For model training, we used cross entropy loss and SGD optimizer, setting batch size $b = 32$. Every result is taken from the average of 3 independent runs. More details about data preprocessing and model architecture can be found in Appendix C.

Table 2: Test accuracy on synthesized datasets with different learning rate lr and local step τ.

Methods	Fashion-MNIST			EMNIST			CIFAR-10			
	$0.1, 3$	$0.3, 3$	$0.3, 5$	$0.3, 3$	$0.1, 5$	$0.3, 5$	$.03, 3$	$0.1, 3$	$.01, 5$	$.03, 5$
Local Training	92.94	92.66	92.80	76.57	77.50	78.89	73.5	74.0	74.0	75.0
FedAvg (McMahan et al., 2017)	78.53	84.35	88.09	76.49	73.53	79.51	67.0	70.5	66.5	69.5
FedProx (Li et al., 2020b)	88.37	89.34	90.86	76.49	73.53	79.32	66.5	71.5	66.5	69.5
Skew (Cho et al., 2020)	83.38	85.04	88.92	72.90	72.83	77.81	76.5	**78.8**	75.0	76.2
SL (Song et al., 2021)	85.32	89.20	90.17	78.94	76.54	81.02	77.2	74.0	78.0	73.8
OCS (Chen et al., 2020)	81.30	86.43	85.04	70.72	69.15	77.18	72.0	52.0	75.0	68.8
Per-FL (Fallah et al., 2020)	77.56	82.13	87.40	74.72	72.39	78.85	67.0	71.0	64.5	69.5
VaRSeL (Ours)	**95.01**	**96.88**	**94.88**	**80.65**	**80.18**	**82.61**	**81.5**	**78.8**	**80.0**	**80.5**

Robustness We've tested the robustness of VaRSeL under different learning rates and local steps. Learning rates range from $0, 3, 0.1, 0.03, 0.01$, and local steps range from $1, 3, 5$. As shown in Table 2, our method outperforms the best accuracy of other competitors by at least 3.94% in Fashion-MNIST, 1.59% in EMNIST, and 2.7% in CIFAR-10.

The fine-tune effect As indicated in our basic convergence analysis, the external weights $\mathbf{w}$ should decay gradually. And in VaRSeL, the weights $\mathbf{w}_*$ is computed adaptively within each iteration, so the fine-tuning will take effect automatically without any human intervention or additional hyper-parameters. In Figure 2, we plot the sum of external weights ($\mathbf{1}^{\mathrm{T}}\mathbf{w}$) in our synthesized experiments, as an indicator of the fine-tune effect. It can be observed that the external weights gradually decrease during the training process, which is consistent with our theoretical analysis.

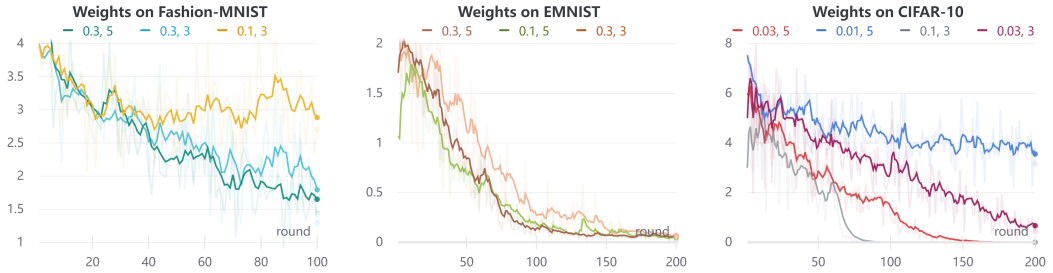

Figure 2: The automatic fine-tune effect of VaRSeL on Fashion-MNIST, EMNIST and CIFAR-10.

5.2 Experiments on Shakespeare dataset

Dataset and settings We used an FL dataset derived from *The Complete Works of William Shakespeare* (Shakespeare, 2007). There are altogether 715 clients, each of whom corresponds to a role in one of the Shakespeare's six classic plays—'*All's Well That Ends Well*', '*Much Ado About Nothing*', '*Pericles, Prince of Tyre*', '*The First Part of King Henry the Fourth*', '*The Taming of the Shrew*', and '*The Tragedy of King Lear*'. The clients' datasets are comprised of their lines in the play. Finally, we designate the roles in one of the six plays as internal clients, and those in the remaining five plays as external clients.

Details of implementation and results Following McMahan et al. (2017), we used a 2-layer LSTM for the prediction of next character. For model training, we employed cross entropy loss, learning rate $\eta \in [0.5, 5.0]$, local steps $\tau \in \{1, 2, 3\}$, and batch size $b = 4$. We took the average results of three independent runs of 200 communication rounds. As shown in Table 3, the average score of VaRSeL outperforms other competitors by at least 0.89%.

5.3 Experiments On medical image dataset

Dataset and settings Tumor segmentation is one of challenging tasks (Hesamian et al., 2019) due to the limitation annotations, the imbalance of data distribution, and the heavy noises in different image scans. In this subsection, we show the real-world performance of our method and other competitors on FeTS 2021 (Pati et al., 2021), a real-world brain tumor segmentation dataset composed of clinically acquired, magnetic resonance imaging (MRI) scans from 17 different medical sites.

Table 3: Prediction accuracy on the plays of Shakespeare.

Methods	*AWTEW*	*MAAN*	*PPT*	*TFPKHF*	*TTS*	*TTKL*	Acc_{mean}
Local Training	45.39	45.23	47.41	45.81	43.43	43.53	45.13
FedAvg (McMahan et al., 2017)	45.52	44.59	45.21	43.61	43.99	44.61	44.59
FedProx (Li et al., 2020b)	45.40	44.49	46.13	44.27	42.99	44.51	44.63
Skew (Cho et al., 2020)	45.33	**46.76**	47.13	45.86	43.89	38.10	44.51
SL (Song et al., 2021)	44.88	42.56	46.25	38.61	42.05	43.27	42.94
OCS (Chen et al., 2020)	45.28	44.23	47.22	45.63	43.84	40.65	44.47
Per-FL (Fallah et al., 2020)	42.20	41.44	42.82	40.48	41.42	42.60	41.83
VaRSeL (Ours)	**45.83**	46.52	**47.96**	**46.27**	**44.48**	**45.03**	**46.02**

Table 4: Dice scores of different methods (95% C.I.) for whole brain tumor segmentation on FeTS 2021 by taking different site data as internal sites.

Methods	Score on 1^{st} Site	Score on 6^{th} Site	Score on 16^{th} Site	$\text{Score}_{\text{mean}}$
Local Training	70.76 ± 0.14	58.12 ± 0.10	53.83 ± 0.34	60.90
FedAvg (McMahan et al., 2017)	68.32 ± 0.83	58.01 ± 0.57	56.95 ± 0.47	61.90
FedProx (Li et al., 2020b)	67.87 ± 0.62	58.48 ± 0.57	56.43 ± 0.31	60.92
Skew (Cho et al., 2020)	68.20 ± 0.08	53.84 ± 0.29	56.32 ± 0.88	59.45
SL (Song et al., 2021)	71.67 ± 0.04	57.42 ± 0.01	58.86 ± 0.23	62.65
OCS (Chen et al., 2020)	69.12 ± 0.09	59.95 ± 0.83	57.33 ± 0.66	62.13
Per-FL (Fallah et al., 2020)	66.48 ± 0.02	55.96 ± 0.10	56.64 ± 0.25	59.69
VaRSeL (Ours)	$\mathbf{72.06 \pm 0.04}$	$\mathbf{61.06 \pm 0.02}$	$\mathbf{61.60 \pm 0.83}$	**64.91**

To form the internal and external clients, we took one of the three largest sites (1^{st}, 6^{th}, or 16^{th}) as the internal site, and the rest as the external ones. Furthermore, for each setting, the chosen internal site was partitioned equally into 3 parts, which are distributed to 3 internal clients respectively to simulate more internal clients. The MRI scans include various modalities (Bakas et al., 2018), say a) native (T1), b) post-contrast T1-weighted (T1Gd), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). For whole brain tumor segmentation, T2-FLAIR is reported to better identify the malignant features (Zeineldin et al., 2020), To highlight the nature of internal/external partition, the internal datasets are all organized in T2-FLAIR, while every external dataset randomly picks up from one modality to enlarge the distribution shift of different datasets.

Details of implementation and results We used the 2D U-Net (Ronneberger et al., 2015) with instance normalization (Ulyanov et al., 2016). We took the following measures to ensure fair competitions between all the methods: (i) Try gradient descent with three different learning rates $\eta = 0.3, 0.1, 0.03$; (ii) Try three different local steps $\tau = 1, 3, 5$; (iii) For the best learning rate and local steps, we take the average results of three independent runs of 60 communication rounds. The experimental results are summarized in Table 4. It is observed that our VaRSeL outperforms local training on all three sites, show the effectiveness of incorporating external sites to boost the performance of internal sites. Also, the average score of VaRSeL largely outperforms other competitors by over 2.26%, showing the effectiveness of our proposed framework.

6 Conclusions and Remarks

This work proposes a novel setting called "selfish federated learning" that cooperatively optimizes a biased objective towards internal distributions in a communication-efficient and heterogeneous privacy-preserving way. The central idea is to reweight the external datasets via a variance-reduction approach, based on which we develop "VaRSeL" (Algorithm 1). We provide convergence analysis for our reweighting strategy, and highlight its fine-tune effect. In synthesized and realistic experiments, we support the theoretical results made in our paper, and demonstrate its better performance compared to alternative methods. Our code is publicly available and we believe that VaRSeL can find a wide range of applications, such as healthcare (Rieke et al., 2020).

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

S Bakas, M Reyes, A Jakab, S Bauer, M Rempfler, A Crimi, RT Shinohara, C Berger, SM Ha, M Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. corr abs/1811.02629 (2018), 2018.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2020.

Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Canh Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pp. 148–156. Citeseer, 1996.

Hossein Shokri Ghadikolaei, Hadi Ghauch, Carlo Fischione, and Mikael Skoglund. Learning and data selection in big datasets. In *International Conference on Machine Learning*, pp. 2191–2200. PMLR, 2019.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *arXiv preprint arXiv:2006.04088*, 2020.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

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

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems*, 2020b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020c.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pp. 3043–3052. PMLR, 2018.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, pp. 89–96. IEEE, 2011.

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

Sarthak Pati, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Jason Martin, Shadi Albarqouni, et al. The federated tumor segmentation (fets) challenge, 2021.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *arXiv preprint arXiv:2106.05203*, 2021.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

William Shakespeare. *The complete works of William Shakespeare*. Wordsworth Editions, 2007.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: learning an explicit mapping for sample weighting. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 1919–1930, 2019.

Youyi Song, Lequan Yu, Baiying Lei, Kup-Sze Choi, and Jing Qin. Selective learning from external data for ct image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 420–430. Springer, 2021.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *arXiv preprint arXiv:2103.00710*, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR, 2019.

Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. Matcha: Speeding up decentralized sgd via matching decomposition sampling. In *2019 Sixth Indian Control Conference (ICC)*, pp. 299–300. IEEE, 2019.

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.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Ramy A Zeineldin, Mohamed E Karar, Jan Coburger, Christian R Wirtz, and Oliver Burgert. Deepseg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance flair images. *International journal of computer assisted radiology and surgery*, 15(6):909–920, 2020.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2020.

APPENDIX

Roadmap of Appendix The Appendix is organized as follows. We provide the theoretical proof of basic convergence results in Section A and main convergence results in Section B. The details of experimental setting (*e.g.*, model architectures and dataset preprocessing) are in Section C.

A PROOF OF BASIC CONVERGENCE RESULTS

A.1 PRELIMINARIES

Proposition 4. *For $\gamma > 0$, $2\langle \mathbf{u}, \mathbf{v}\rangle \leq \gamma\|\mathbf{u}\|^2 + \frac{1}{\gamma}\|\mathbf{v}\|^2$.*

Proof. According to Cauchy-Schwarz inequality, we have

$$2\langle \mathbf{u}, \mathbf{v}\rangle = 2\langle \sqrt{\gamma}\mathbf{u}, \frac{1}{\sqrt{\gamma}}\mathbf{v}\rangle \leq 2\|\sqrt{\gamma}\mathbf{u}\| \cdot \|\frac{1}{\sqrt{\gamma}}\mathbf{v}\| \leq \gamma\|\mathbf{u}\|^2 + \frac{1}{\gamma}\|\mathbf{v}\|^2.$$

□

Proposition 5. *Let $\sum_{i=1}^N p_i = 1$, $p_i > 0$ for $1 \leq i \leq N$. Then*

$$\mathbb{E}\|\sum_{i=1}^N p_i \mathbf{x}_i\|^2 \leq \sum_{i=1}^N p_i \mathbb{E}\|\mathbf{x}_i\|^2.$$

Proof. This follows from Jensen's inequality and the convexity of $\mathbb{E}\|\mathbf{x}\|^2$. □

Lemma 6. *For $\theta \in \Theta_{\mathcal{H}}$, we have*

$$l(\theta^*) - l(\theta) \leq -\frac{1}{2L}\|\nabla l(\theta)\|^2.$$

Proof. Let $\theta' = \theta - \frac{1}{L}\nabla l(\theta)$. Then

$$\begin{aligned} l(\theta') &\leq l(\theta) + \langle \nabla l(\theta), \theta' - \theta\rangle + \frac{L}{2}\|\theta' - \theta\|^2 \\ &\leq l(\theta) + \langle \nabla l(\theta), -\frac{1}{L}\nabla l(\theta)\rangle + \frac{L}{2}\|\frac{1}{L}\nabla l(\theta)\|^2 \\ &= l(\theta) - \frac{1}{2L}\|\nabla l(\theta)\|^2. \end{aligned}$$

Hence

$$l(\theta^*) - l(\theta) \leq l(\theta') - l(\theta) \leq -\frac{1}{2L}\|\nabla l(\theta)\|^2.$$

□

Lemma 7. *Let $\{a_n\}$ be the sequence of non-negative real numbers. If $0 < \alpha < 1$, $\beta > 0$, and*

$$a_{n+1} \leq (1-\alpha)a_n + \beta,$$

then

$$a_n \leq (1-\alpha)^n a_0 + \frac{\beta}{\alpha}.$$

Proof. Since

$$a_n - \frac{\beta}{\alpha} \leq (1-\alpha)(a_n - \frac{\beta}{\alpha}),$$

we can simply prove by induction that

$$a_n - \frac{\beta}{\alpha} \leq (1-\alpha)^n[a_0 - \frac{\beta}{\alpha}],$$

and then

$$\begin{aligned} a_n &\leq (1-\alpha)^n a_0 + [1-(1-\alpha)^n]\frac{\beta}{\alpha} \\ &\leq (1-\alpha)^n a_0 + \frac{\beta}{\alpha}. \end{aligned}$$

□

A.2 PROOF OF LEMMA 1

Proof. We have:

$$\begin{aligned} &\Phi(\mathbf{w};\theta) \\ &= \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\left[\sum_{i=1}^{M}\mathbb{E}\|\nabla l_i(\theta)-\nabla l(\theta)\|^2 + \mathbb{E}\Big\|\sum_{j=M+1}^{M+N}\mathbf{w}_j\big(\nabla l_j(\theta)-\nabla l(\theta)\big)\Big\|^2\right] \\ &\leq \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\left[M^2\sigma\|\nabla l(\theta)\|^2 + \mathbb{E}\Big\|\sum_{j=M+1}^{M+N}\mathbf{w}_j\big(\nabla l_j(\theta)-\nabla l(\theta)\big)\Big\|^2\right] \qquad (9) \\ &= \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \mathbb{E}\big[(\Sigma^{\mathrm{T}}\mathbf{w})^{\mathrm{T}}(\Sigma^{\mathrm{T}}\mathbf{w})\big]\Big) \\ &= \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \mathbb{E}\big[\mathbf{w}^{\mathrm{T}}(\Sigma\Sigma^{\mathrm{T}})\mathbf{w}\big]\Big), \end{aligned}$$

where

$$\Sigma = \begin{bmatrix} \nabla l_{M+1}(\theta)-\nabla l(\theta) \\ \vdots \\ \nabla l_{M+N}(\theta)-\nabla l(\theta) \end{bmatrix} \in \mathbb{R}^{N\times d}.$$

Since the right hand side of Equation (9) is an upper bound of $\Phi(\mathbf{w};\theta)$, and we want to make this upper bound tight by taking its minimum (w.r.t. $\tilde{\mathbf{w}}$):

$$\tilde{\mathbf{w}} = \underset{\mathbf{w}}{argmin}\ \frac{1}{(M+\mathbf{1}^{\mathrm{T}}\mathbf{w})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \mathbb{E}\big[\mathbf{w}^{\mathrm{T}}(\Sigma\Sigma^{\mathrm{T}})\mathbf{w}\big]\Big).$$

Clearly, whatever $\tilde{\mathbf{w}}$ you take, $\mathrm{RHS}_{(9)}$ will always be no less than $\Phi(\mathbf{w}_*;\theta)$.

Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ denote all the N eigenvalues of $\Sigma\Sigma^{\mathrm{T}}$. Then

$$N\lambda_1 \leq \sum_{j=M+1}^{M+N}\lambda_j = \mathrm{trace}\big(\Sigma\Sigma^{\mathrm{T}}\big) = \sum_{j=M+1}^{M+N}\mathbb{E}\|\nabla l_j(\theta)-\nabla l(\theta)\|^2. \qquad (10)$$

Let $\mathbf{e}$ be the corresponding unit eigenvector (*i.e.*$\mathbf{e}^{\mathrm{T}}\mathbf{e}=1$) of λ_1. Then, in order to minimize the $\mathrm{RHS}_{(9)}$, $\tilde{\mathbf{w}}$ and $\mathbf{e}$ must be collinear vectors. Therefore, we can assume that

$$\tilde{\mathbf{w}} = \omega\mathbf{e}.$$

Substituting this into $\mathrm{RHS}_{(9)}$, we have

$$\begin{aligned} &\frac{1}{(M+\mathbf{1}^{\mathrm{T}}\tilde{\mathbf{w}})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \mathbb{E}\big[\tilde{\mathbf{w}}^{\mathrm{T}}(\Sigma\Sigma^{\mathrm{T}})\tilde{\mathbf{w}}\big]\Big) \\ &= \frac{1}{(M+\omega\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \omega^2\mathbb{E}\big[\mathbf{e}^{\mathrm{T}}(\Sigma\Sigma^{\mathrm{T}})\mathbf{e}\big]\Big) \\ &= \frac{1}{(M+\omega\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}\Big(M^2\sigma\|\nabla l(\theta)\|^2 + \lambda_1\omega^2\Big) \triangleq h(\omega) \end{aligned}$$

Take derivative with respect to ω:

$$\frac{\partial h}{\partial \omega} = \frac{2\lambda_1\omega(M+\omega\mathbf{1}^{\mathrm{T}}\mathbf{e})^2 - \Big(M^2\sigma\|\nabla l(\theta)\|^2 + \lambda_1\omega^2\Big)\cdot 2\mathbf{1}^{\mathrm{T}}\mathbf{e}(M+\omega\mathbf{1}^{\mathrm{T}}\mathbf{e})}{(M+\omega\mathbf{1}^{\mathrm{T}}\mathbf{e})^4} = 0.$$

Thus,

$$\omega = \lambda_1^{-1} \cdot \mathbf{1}^{\mathrm{T}}\mathbf{e} M\sigma\|\nabla l(\theta)\|^2,$$

and then

$$\begin{aligned} h(\omega) &= \frac{\sigma\|\nabla l(\theta)\|^2}{1+\lambda_1^{-1}(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2\sigma\|\nabla l(\theta)\|^2} \\ &\overset{(10)}{\leq} \frac{\sigma\|\nabla l(\theta)\|^2}{1+\frac{N(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2\sigma\|\nabla l(\theta)\|^2}{\sum_{j=M+1}^{M+N}\mathbb{E}\|\nabla l_j(\theta)-\nabla l(\theta)\|^2}} \\ &= \frac{1}{\frac{1}{\sigma\|\nabla l(\theta)\|^2}+\frac{N(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}{\sum_{j=M+1}^{M+N}\mathbb{E}\|\nabla l_j(\theta)-\nabla l(\theta)\|^2}}. \end{aligned}$$

According to Assumption 1, we further have

$$\begin{aligned} h(\omega) &\leq \frac{1}{\frac{1}{\sigma\|\nabla l(\theta)\|^2}+\frac{N(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}{\sum_{j=M+1}^{M+N}\mathbb{E}\|\nabla l_j(\theta)-\nabla l(\theta)\|^2}} \\ &\leq \frac{1}{\frac{1}{\sigma\|\nabla l(\theta)\|^2}+\frac{N(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}{(\sum_{j=M+1}^{M+N}\kappa_j)\|\nabla l(\theta)\|^2+\sum_{j=M+1}^{M+N}\nu_j}}. \end{aligned}$$

Further, due to *AM-HM* inequality, we have

$$\begin{aligned} h(\omega) &\leq \frac{1}{\frac{M}{M\sigma\|\nabla l(\theta)\|^2}+\frac{N(\mathbf{1}^{\mathrm{T}}\mathbf{e})^2}{(\sum_{j=M+1}^{M+N}\kappa_j)\|\nabla l(\theta)\|^2+\sum_{j=M+1}^{M+N}\nu_j}} \\ &\leq \frac{M\sigma\|\nabla l(\theta)\|^2+(\mathbf{1}^{\mathrm{T}}\mathbf{e})^{-2}\left[(\sum_{j=M+1}^{M+N}\kappa_j)\|\nabla l(\theta)\|^2+\sum_{j=M+1}^{M+N}\nu_j\right]}{(M+N)^2} \end{aligned}$$

Therefore, $\Phi(\mathbf{w}_*;\theta)$ can be finally bounded as follows:

$$\begin{aligned} \Phi(\mathbf{w}_*;\theta) \leq h(\omega) &\leq \frac{M\sigma\|\nabla l(\theta)\|^2+(\mathbf{1}^{\mathrm{T}}\mathbf{e})^{-2}\left[(\sum_{j=M+1}^{M+N}\kappa_j)\|\nabla l(\theta)\|^2+\sum_{j=M+1}^{M+N}\nu_j\right]}{(M+N)^2} \\ &= \underbrace{\frac{M\sigma+(\mathbf{1}^{\mathrm{T}}\mathbf{e})^{-2}\sum_{j=M+1}^{M+N}\kappa_j}{(M+N)^2}}_{A}\|\nabla l(\theta)\|^2+\underbrace{\frac{(\mathbf{1}^{\mathrm{T}}\mathbf{e})^{-2}\sum_{j=M+1}^{M+N}\nu_j}{(M+N)^2}}_{B} \\ &\triangleq A\|\nabla l(\theta)\|^2+B. \end{aligned} \tag{11}$$

□

Comments. From Formula (11), we can observe that B is mainly determined by $\sum_{j=M+1}^{M+N}\nu_j$, *i.e.* the noises in the external datasets.

A.3 Proof of Theorem 2

Proof. Due to the μ-strongly convexity, we have

$$\begin{aligned} l(\theta^*) &\geq l(\theta^t)+\langle\nabla l(\theta^t),\theta^*-\theta^t\rangle+\frac{\mu}{2}\|\theta^*-\theta^t\|^2 \\ &= l(\theta^t)+\langle\nabla l(\theta^t;\mathbf{w}),\theta^*-\theta^t\rangle-\langle\nabla l(\theta^t;\mathbf{w})-\nabla l(\theta^t),\theta^*-\theta^t\rangle+\frac{\mu}{2}\|\theta^*-\theta^t\|^2 \\ &\geq l(\theta^t)+\langle\nabla l(\theta^t;\mathbf{w}),\theta^*-\theta^t\rangle+\frac{\mu}{2}\|\theta^*-\theta^t\|^2 \\ &\quad-\frac{1}{2}\left[\gamma_1\|\nabla l(\theta^t;\mathbf{w})-\nabla l(\theta^t)\|^2+\frac{1}{\gamma_1}\|\theta^*-\theta^t\|^2\right] \\ &= l(\theta^t)+\langle\nabla l(\theta^t;\mathbf{w}),\theta^*-\theta^t\rangle+(\frac{\mu}{2}-\frac{1}{2\gamma_1})\|\theta^*-\theta^t\|^2-\frac{\gamma_1}{2}\|\nabla l(\theta^t;\mathbf{w})-\nabla l(\theta^t)\|^2, \end{aligned} \tag{12}$$

where $\gamma_1 > 0$.

Since $\theta^{t+1} = \theta^t - \eta \nabla l(\theta^t; \mathbf{w})$, it yields

$$\begin{aligned}
\langle \nabla l(\theta^t; \mathbf{w}), \theta^* - \theta^t \rangle &= -\frac{1}{\eta} \langle \theta^{t+1} - \theta^t, \theta^* - \theta^t \rangle \\
&= -\frac{1}{2\eta} \left[\|\theta^{t+1} - \theta^t\|^2 + \|\theta^* - \theta^t\|^2 - \|\theta^* - \theta^{t+1}\|^2 \right] \\
&= -\frac{1}{2\eta} \left[\eta^2 \|\nabla l(\theta^t; \mathbf{w})\|^2 + \|\theta^* - \theta^t\|^2 - \|\theta^* - \theta^{t+1}\|^2 \right] \\
&= -\frac{\eta}{2} \|\nabla l(\theta^t; \mathbf{w})\|^2 - \frac{1}{2\eta} \|\theta^* - \theta^t\|^2 + \frac{1}{2\eta} \|\theta^* - \theta^{t+1}\|^2.
\end{aligned} \tag{13}$$

Based on Lemma 1, we have

$$\mathbb{E} \|\nabla l(\theta^t; \mathbf{w}_*) - \nabla l(\theta^t)\|^2 \leq A \mathbb{E} \|\nabla l(\theta^t)\|^2 + B, \tag{14}$$

and then

$$\begin{aligned}
\mathbb{E} \|\nabla l(\theta^t; \mathbf{w}_*)\|^2 &= \mathbb{E} \|\nabla l(\theta^t) + \nabla l(\theta^t; \mathbf{w}_*) - \nabla l(\theta^t)\|^2 \\
&\leq \gamma_2 \mathbb{E} \|\nabla l(\theta^t)\|^2 + \frac{1}{\gamma_2} \mathbb{E} \|\nabla l(\theta^t; \mathbf{w}_*) - \nabla l(\theta^t)\|^2 \\
&\leq \gamma_2 \mathbb{E} \|\nabla l(\theta^t)\|^2 + \frac{1}{\gamma_2} \left(A \mathbb{E} \|\nabla l(\theta^t)\|^2 + B \right) \\
&= (\gamma_2 + \frac{A}{\gamma_2}) \mathbb{E} \|\nabla l(\theta^t)\|^2 + \frac{B}{\gamma_2},
\end{aligned}$$

where $\gamma_2 > 0$.

Substituting it into Equation (13), we have

$$\begin{aligned}
&\mathbb{E} \langle \nabla l(\theta^t; \mathbf{w}_*), \theta^* - \theta^t \rangle \\
&\geq -\frac{\eta}{2} \left[(\gamma_2 + \frac{A}{\gamma_2}) \mathbb{E} \|\nabla l(\theta^t)\|^2 + \frac{B}{\gamma_2} \right] - \frac{1}{2\eta} \mathbb{E} \|\theta^* - \theta^t\|^2 + \frac{1}{2\eta} \mathbb{E} \|\theta^* - \theta^{t+1}\|^2.
\end{aligned} \tag{15}$$

Then, substituting Formula (14), (15) into Inequality (12), we have

$$\begin{aligned}
\mathbb{E}[l(\theta^*)] \geq \mathbb{E}[l(\theta^t)] &- \frac{\eta}{2} \left[(\gamma_2 + \frac{A}{\gamma_2}) \mathbb{E} \|\nabla l(\theta^t)\|^2 + \frac{B}{\gamma_2} \right] - \frac{1}{2\eta} \mathbb{E} \|\theta^* - \theta^t\|^2 + \frac{1}{2\eta} \mathbb{E} \|\theta^* - \theta^{t+1}\|^2 \\
&+ (\frac{\mu}{2} - \frac{1}{2\gamma_1}) \mathbb{E} \|\theta^* - \theta^t\|^2 - \frac{\gamma_1}{2} \left(A \mathbb{E} \|\nabla l(\theta^t)\|^2 + B \right).
\end{aligned}$$

Rewrite the above formula as:

$$\begin{aligned}
\mathbb{E} \|\theta^* - \theta^{t+1}\|^2 \leq &\left(1 - (\eta\mu - \frac{\eta}{\gamma_1})\right) \mathbb{E} \|\theta^* - \theta^t\|^2 + 2\eta \mathbb{E} \left[l(\theta^*) - l(\theta^t) \right] \\
&+ (A\eta\gamma_1 + \frac{A\eta^2}{\gamma_2} + \eta^2\gamma_2) \mathbb{E} \|\nabla l(\theta^t)\|^2 + (\frac{\eta^2}{\gamma_2} + \eta\gamma_1) B,
\end{aligned}$$

and then, it follows from Lemma 6 that

$$\begin{aligned}
\mathbb{E} \|\theta^* - \theta^{t+1}\|^2 &\leq \left(1 - (\eta\mu - \frac{\eta}{\gamma_1}\right) \mathbb{E} \|\theta^* - \theta^t\|^2 - \frac{\eta}{L} \mathbb{E} \|\nabla l(\theta^t)\|^2 \\
&\quad + (A\eta\gamma_1 + \frac{A\eta^2}{\gamma_2} + \eta^2\gamma_2) \mathbb{E} \|\nabla l(\theta^t)\|^2 + (\frac{\eta^2}{\gamma_2} + \eta\gamma_1) B \\
&\leq \left(1 - (\eta\mu - \frac{\eta}{\gamma_1})\right) \mathbb{E} \|\theta^* - \theta^t\|^2 + (\frac{\eta^2}{\gamma_2} + \eta\gamma_1) B,
\end{aligned}$$

where $\frac{1}{L} \geq A\gamma_1 + \frac{A\eta}{\gamma_2} + \eta\gamma_2$ and $\mu > \frac{1}{\gamma_1}$.[2]

[2]It's easy to verify that γ_1 and γ_2 exist when $\eta < \frac{\frac{1}{L} - \frac{A}{\mu}}{2\sqrt{A}}$.

Finally, according to Lemma 7, we have

$$\mathbb{E}\|\theta^* - \theta^t\|^2 \leq \big(1 - (\eta\mu - \frac{\eta}{\gamma_1})\big)^t \mathbb{E}\|\theta^* - \theta^0\|^2 + \frac{\frac{\eta}{\gamma_2} + \gamma_1}{\mu - \frac{1}{\gamma_1}} B.$$

□

B PROOF OF MAIN CONVERGENCE RESULTS

B.1 PRELIMINARIES

To simplify, we introduce the following notations:

$$\theta_{\mathcal{I}} = \frac{1}{M}\sum_{i=1}^{M}\theta_i,$$
$$\overline{\theta}_{\mathbf{w}} = \frac{\sum_{i=1}^{M}\theta_i + \sum_{j=M+1}^{M+N} w_j\theta_j}{M + \mathbf{1}^{\mathrm{T}}\mathbf{w}} = \frac{M\theta_{\mathcal{I}} + \sum_{j=M+1}^{M+N} w_j\theta_j}{M + \mathbf{1}^{\mathrm{T}}\mathbf{w}}.$$

Thus,

$$\theta^{t+1} = \overline{\theta^{t+\frac{\tau}{\tau+1}}}_{\mathbf{w}}.$$

Lemma 8. *For* $0 \leq k \leq \tau - 1$,

$$\mathbb{E}\left\|\theta_i^{t+\frac{k}{\tau+1}} - \overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\right\|^2 \leq 4\eta^2 k^2 G^2. \tag{16}$$

Proof. Prove by induction. Inequality 16 holds for $k = 0$. Assume that it also holds for $k-1$, $0 < k < \tau$. Then,

$$\begin{aligned}
&\mathbb{E}\left\|\theta_i^{t+\frac{k}{\tau+1}} - \overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\right\| \\
&= \mathbb{E}\left\|\theta_i^{t+\frac{k}{\tau+1}} - \theta_i^{t+\frac{k-1}{\tau+1}} + \theta_i^{t+\frac{k-1}{\tau+1}} - \overline{\theta^{t+\frac{k-1}{\tau+1}}}_{\mathbf{w}} + \overline{\theta^{t+\frac{k-1}{\tau+1}}}_{\mathbf{w}} - \overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\right\| \\
&\leq \mathbb{E}\left\|\theta_i^{t+\frac{k}{\tau+1}} - \theta_i^{t+\frac{k-1}{\tau+1}}\right\| + \mathbb{E}\left\|\theta_i^{t+\frac{k-1}{\tau+1}} - \overline{\theta^{t+\frac{k-1}{\tau+1}}}_{\mathbf{w}}\right\| + \mathbb{E}\left\|\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}} - \overline{\theta^{t+\frac{k-1}{\tau+1}}}_{\mathbf{w}}\right\| \\
&\leq \eta G + 2\eta(k-1)G + \eta G \\
&= 2\eta k G.
\end{aligned}$$

Hence, $\mathbb{E}\left\|\theta_i^{t+\frac{k}{\tau+1}} - \overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\right\|^2 \leq 4\eta^2 k^2 G^2$. □

B.2 PROOF OF THEOREM 3

Proof. According to Assumption 2a, we have

$$
\begin{aligned}
&\mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k+1}{\tau+1}}}\mathbf{w}\big)\Big] \\
&= \mathbb{E}\Big[l\Big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w} - \frac{\eta}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\big(\sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\big)\Big)\Big] \\
&\le \mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big] \\
&\quad \underbrace{-\frac{\eta}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\mathbb{E}\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\Big\rangle}_{T_1} \\
&\quad \underbrace{+\frac{L\eta^2}{2}\mathbb{E}\Big\|\frac{1}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\big(\sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\big)\Big\|^2}_{T_2} \\
&\triangleq \mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big] + T_1 + T_2.
\end{aligned} \tag{17}
$$

Then we are to bound the T_1 and T_2 in Equation 17 respectively.

For T_1, we have

$$
\begin{aligned}
T_1 &= -\frac{\eta}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\mathbb{E}\Big[\sum_{i=1}^{M}\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \nabla l_i(\theta_i^{t+\frac{k}{\tau+1}})\Big\rangle + \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\Big\rangle\Big] \\
&= -\frac{\eta}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\mathbb{E}\Big[(M+\sum_{j=M+1}^{M+N}\mathbf{w}_j)\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big\rangle \\
&\qquad - \sum_{i=1}^{M}\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_i(\theta_i^{t+\frac{k}{\tau+1}})\Big\rangle \\
&\qquad - \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\Big\rangle\Big] \\
&= -\eta\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big\|^2 + \frac{\eta}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\mathbb{E}\Big\langle\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big), \\
&\qquad \sum_{i=1}^{M}\Big(\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_i(\theta_i^{t+\frac{k}{\tau+1}})\Big) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big(\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)\Big)\Big\rangle,
\end{aligned}
$$

and further, since $\langle\mathbf{u}_1, \mathbf{v}_1\rangle \le \gamma_1\|\mathbf{u}_1\|^2 + \frac{1}{4\gamma_1}\|\mathbf{v}_1\|^2$, it yields

$$
\begin{aligned}
T_1 &\le (-\eta+\gamma_1\eta)\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big\|^2 \\
&\quad + \frac{\eta}{4\gamma_1}\mathbb{E}\Big\|\frac{\sum_{i=1}^{M}\Big(\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_i(\theta_i^{t+\frac{k}{\tau+1}})\Big) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big(\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)\Big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\Big\|^2 \\
&= (-\eta+\gamma_1\eta)\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big\|^2 \\
&\quad \underbrace{+\frac{\eta}{4\gamma_1}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big) - \frac{\sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\Big\|^2}_{T_3} \\
&\triangleq (-\eta+\gamma_1\eta)\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}\mathbf{w}\big)\Big\|^2 + \frac{\eta}{4\gamma_1}T_3,
\end{aligned} \tag{18}
$$

where $0 < \gamma_1 < 1$.

For T_2, we have

$$
\begin{aligned}
T_2 &= \frac{L\eta^2}{2}\mathbb{E}\Big\|\frac{1}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\big(\sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}})\big)\Big\|^2 \\
&= \frac{L\eta^2}{2}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big) \\
&\qquad + \frac{\sum_{i=1}^{M}\Big(\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) - \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big(\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}}) - \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\Big\|^2
\end{aligned}
$$

and further, since $\|\mathbf{u}_2+\mathbf{v}_2\|^2 = \|\mathbf{u}_2\|^2 + \|\mathbf{v}_2\|^2 + 2\langle\mathbf{u}_2,\mathbf{v}_2\rangle \le (1+\gamma_2)\|\mathbf{u}_2\|^2 + (1+\frac{1}{\gamma_2})\|\mathbf{v}_2\|^2$, it yields

$$
\begin{aligned}
T_2 &\le \frac{(1+\gamma_2)L\eta^2}{2}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big\|^2 + \frac{(1+\frac{1}{\gamma_2})L\eta^2}{2} \\
&\qquad \cdot \mathbb{E}\Big\|\frac{\sum_{i=1}^{M}\Big(\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) - \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\Big(\nabla l_j(\theta_j^{t+\frac{k}{\tau+1}}) - \nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\Big\|^2 \\
&= \frac{(1+\gamma_2)L\eta^2}{2}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big\|^2 \\
&\qquad + \frac{(1+\frac{1}{\gamma_2})L\eta^2}{2}\underbrace{\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big) - \frac{\sum_{i=1}^{M}\nabla l_i(\theta_i^{t+\frac{k}{\tau+1}}) + \sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\Big\|^2}_{T_3^{(t+\frac{k}{\tau+1})}} \\
&\triangleq \frac{(1+\gamma_2)L\eta^2}{2}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big\|^2 + \frac{(1+\frac{1}{\gamma_2})L\eta^2}{2}T_3^{(t+\frac{k}{\tau+1})},
\end{aligned}
\tag{19}
$$

where $\gamma_2 > 0$.

Therefore, putting Equation 17, 18, 19 together, we have

$$
\begin{aligned}
\mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k+1}{\tau+1}}}_{\mathbf{w}}\big)\Big] &= \mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big] + T_1 + T_2 \\
&\le \mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big] - \Big((1-\gamma_1)\eta - \frac{(1+\gamma_2)L}{2}\eta^2\Big)\cdot\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big\|^2 \\
&\qquad + \Big(\frac{\eta}{4\gamma_1} + \frac{(1+\frac{1}{\gamma_2})L\eta^2}{2}\Big)\cdot T_3^{(t+\frac{k}{\tau+1})}.
\end{aligned}
\tag{20}
$$

Let $\Gamma_0 = (1-\gamma_1) - \frac{(1+\gamma_2)L}{2}\eta$, and $\Gamma_1 = \frac{1}{4\gamma_1} + \frac{(1+\frac{1}{\gamma_2})L\eta}{2}$. Summing over k from 0 to $\tau-1$ in Equation 20, we have

$$
\begin{aligned}
\sum_{k=0}^{\tau-1}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)\Big\|^2 &\le \frac{\mathbb{E}\Big[l\big(\overline{\theta^{t}}_{\mathbf{w}}\big)\Big] - \mathbb{E}\Big[l\big(\overline{\theta^{t+\frac{\tau}{\tau+1}}}_{\mathbf{w}}\big)\Big]}{\eta\Gamma_0} + \frac{\Gamma_1}{\Gamma_0}\sum_{k=0}^{\tau-1}T_3^{(t+\frac{k}{\tau+1})} \\
&= \frac{\mathbb{E}\Big[l\big(\theta^{t}\big)\Big] - \mathbb{E}\Big[l\big(\theta^{t+1}\big)\Big]}{\eta\Gamma_0} + \frac{\Gamma_1}{\Gamma_0}\sum_{k=0}^{\tau-1}T_3^{(t+\frac{k}{\tau+1})}
\end{aligned}
\tag{21}
$$

Then, we are to bound $\sum_{k=0}^{\tau-1} T_3^{(t+\frac{k}{\tau+1})}$:

$$
\begin{aligned}
T_3^{(t+\frac{k}{\tau+1})} &= \mathbb{E}\left\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big) - \frac{\sum_{i=1}^{M}\nabla l_i\big(\theta_i^{t+\frac{k}{\tau+1}}\big)+\sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\right\|^2 \\
&\le (1+\gamma_3)\mathbb{E}\left\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big) - \frac{\sum_{i=1}^{M}\nabla l_i\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)+\sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\right\|^2 \\
&\quad + (1+\frac{1}{\gamma_3})\mathbb{E}\left\|\frac{\sum_{i=1}^{M}\Big(\nabla l_i\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)-\nabla l_i\big(\theta_i^{t+\frac{k}{\tau+1}}\big)\Big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\right. \\
&\qquad\qquad \left. + \frac{\sum_{j=M+1}^{M+N}\mathbf{w}_j\Big(\nabla l_j\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)-\nabla l_j\big(\theta_j^{t+\frac{k}{\tau+1}}\big)\Big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\right\|^2 \\
&\le (1+\gamma_3)\mathbb{E}\left\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big) - \frac{\sum_{i=1}^{M}\nabla l_i\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)+\sum_{j=M+1}^{M+N}\mathbf{w}_j\nabla l_j\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}\big)}{M+\mathbf{1}^{\mathrm{T}}\mathbf{w}}\right\|^2 \\
&\quad + (1+\frac{1}{\gamma_3})\cdot 4L^2\eta^2k^2G^2 \\
&= (1+\gamma_3)\Phi(\mathbf{w};\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}}) + (4+\frac{4}{\gamma_3})L^2\eta^2k^2G^2 \\
&\le (1+\gamma_3)A\big\|\nabla l(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}})\big\|^2 + (1+\gamma_3)B + (4+\frac{4}{\gamma_3})L^2\eta^2k^2G^2,
\end{aligned}
$$

and then

$$
\begin{aligned}
\sum_{k=0}^{\tau-1} T_3^{(t+\frac{k}{\tau+1})} \le \sum_{k=0}^{\tau-1}&(1+\gamma_3)A\big\|\nabla l(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}_*^t})\big\|^2 \\
&+ \underbrace{(1+\gamma_3)\tau B + (\frac{2}{3}+\frac{2}{3\gamma_3})L^2\eta^2G^2(2\tau^3-3\tau^2+\tau)}_{\Gamma_2}.
\end{aligned}
\tag{22}
$$

Substituting Formula (22) into Inequality (21), we have

$$
\eta\underbrace{\big(\Gamma_0-(1+\gamma_3)A\Gamma_1\big)}_{\Gamma_3}\sum_{k=0}^{\tau-1}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}_*^t}\big)\Big\|^2 \le \mathbb{E}\Big[l\big(\theta^t\big)\Big]-\mathbb{E}\Big[l\big(\theta^{t+1}\big)\Big]+\eta\Gamma_1\Gamma_2,
$$

or

$$
\sum_{k=0}^{\tau-1}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}_*^t}\big)\Big\|^2 \le \frac{\mathbb{E}\Big[l\big(\theta^t\big)\Big]-\mathbb{E}\Big[l\big(\theta^{t+1}\big)\Big]}{\eta\Gamma_3}+\frac{\Gamma_1\Gamma_2}{\Gamma_3}.
$$

Summing over t from 0 to $T-1$ and dividing by τT, we have:[3]

$$
\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=0}^{\tau-1}\mathbb{E}\Big\|\nabla l\big(\overline{\theta^{t+\frac{k}{\tau+1}}}_{\mathbf{w}_*^t}\big)\Big\|^2 \le \frac{l\big(\theta^0\big)-l\big(\theta^*\big)}{\eta\Gamma_3\tau T}+\frac{\Gamma_1\Gamma_2}{\Gamma_3\tau}.
$$

□

C Experimental Details

C.1 Model Architectures

In synthesized experiments, we mainly follow the model architectures used in McMahan et al. (2017). The logistic model is a 784×10 fully connected layer. The 2NN model consists of a

[3] It's easy to verify that the constants in the derivation exist when $\eta < \frac{2(1-\sqrt{A})}{L(A+\sqrt{A})}$.

784×200 hidden layer, another 200×200 hidden layer, and a final 200×47 output layer. The CNN model consists of a convolution layer with 5×5 kernel and 32 channels, another convolution layer with same kernel size and 64 channels, a fully connected layer with 512 units with ReLU activation, and a final 512×10 output layer. For all three model outputs, we use cross entropy loss.

In Shakespeare experiment, we use a 8-dimension embedding, a 2-layer character LSTM with 256 hidden units, and a final 256×90 output layer. We use cross entropy loss that ignores empty words in the truncated sentence.

In FeTS2021 experiment, we follow the U-Net structure in Ronneberger et al. (2015). There are three slight modifications: (1) The input image is resized to 120×120; (2) We halves the channel sizes to $(32, 64, 128, 256, 512, 256, 128, 64, 32)$; (3) We replace the batch normalization layer with instance normalization layer.

C.2 Dataset Preprocessing

In synthesized experiments, we mainly follow the preprocessing methods suggested in McMahan et al. (2017). The Fashion-MNIST dataset is split into 120 shards, each of which has 500 samples of one single digit. Assign each of the 60 clients two shards. Two of the clients are chosen as internal clients. The EMNIST-Balanced dataset is split into 600 shards, each of which has 180 samples of one single category. Assign each client 24 shards with continuous category indices. Two of the clients are chosen as internal clients. The CIFAR-10 dataset is partitioned in the same way as Fashion-MNIST. For all three synthesized experiments, the communication threshold K is set to 10.

In Shakespeare experiments, there are altogether 715 clients, each of whom corresponds to one role in one of the six classic plays of Shakespeare. Then all the clients are naturally classified into 6 groups, according to which of the six plays they belong to. The clients' datasets are formed by their lines in the play. To simplify the training, we truncate each line into segments of 80 characters. Each time we designate the clients in one group as internal clients, and the remaining clients as external clients. The communication threshold K is set to 50.

In FeTS2021 experiments, as shown in Figure 3, there are altogether 17 medical sites. There are only 341 different volumes, so we only choose the three largest sites (numbered 1, 6 and 16) as internal sites in our experiments. We did not choose the remaining sites as internal sites since they are too small to separate a test set of appropriate size. Further, we divide the chosen internal site into three balanced parts to form the group of internal clients. Moreover, to simplify the training, all client samples are resized to 120×120 and all the 3 different lesion labels are merged for whole brain tumor segmentation. Finally, we set the communication threshold K to 4.

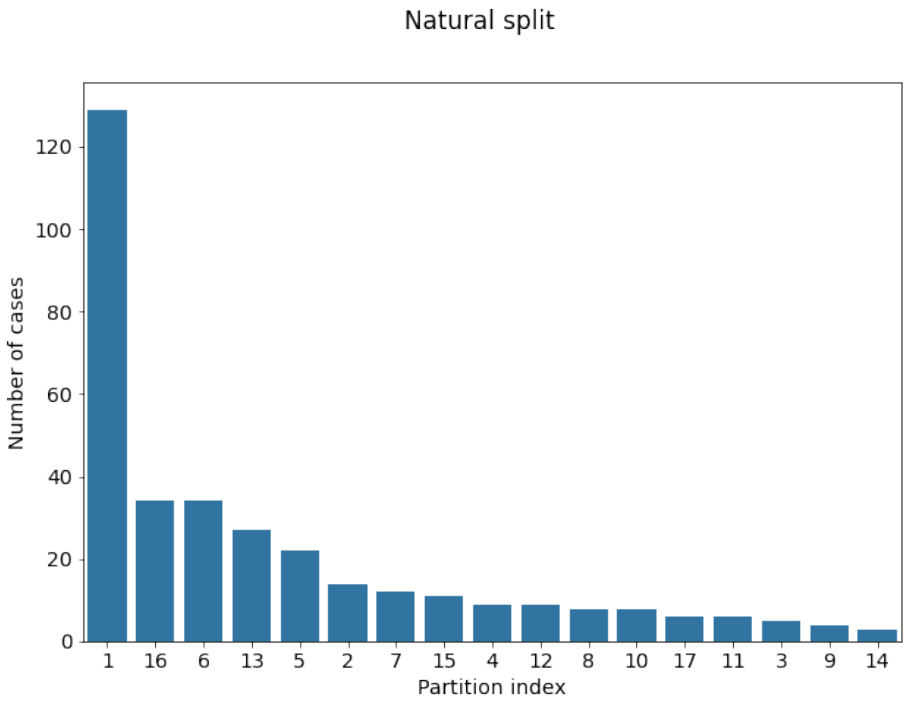

Figure 3: The natural partition of 17 medical sites in FeTS2021. The figure is downloaded from https://fets-ai.github.io/Challenge/data/.