

SUPPLEMENTARY MATERIAL: ADAPTIVE PERSONALIZED FEDERATED LEARNING

TABLE OF CONTENTS

A Additional Related Work	12
B Additional Experimental Results	14
B.1 Datasets	14
B.2 Additional Results	15
C Discussions and Extensions	15
D Proof of Generalization Bound	19
E Proof of Convergence Rate in Convex Setting	21
E.1 Proof without Sampling	22
E.1.1 Proof of Useful Lemmas	23
E.1.2 Proof of Theorem 4	26
E.1.3 Proof of Theorem 5	26
E.2 Proof of Convergence of APFL with Sampling	30
E.2.1 Proof of Useful Lemmas	31
E.2.2 Proof of Theorem 6	33
E.2.3 Proof of Theorem 2	34
F Convergence Rate without Assumption on α_i	37
G Proof of Convergence Rate in Nonconvex Setting	39
G.1 Proof of Technical Lemmas	39
G.2 Proof of Theorem 8	44
G.3 Proof of Theorem 3	44

A ADDITIONAL RELATED WORK

The number of research in federated learning is proliferating during the past few years. In federated learning, the main objective is to learn a global model that is good enough for yet to be seen data and has fast convergence to a local optimum. This indicates that there are several uncanny resemblances between federated learning and meta-learning approaches (Finn et al., 2017; Nichol et al., 2018). However, despite this similarity, meta-learning approaches are mainly trying to learn multiple models, personalized for each new task, whereas in most federated learning approaches, the main focus is on the single global model. As discussed by Kairouz et al. (2019), the gap between the performance of global and personalized models shows the crucial importance of personalization in federated learning. Several different approaches are trying to personalize the global model, primarily focusing on optimization error, while the main challenge with personalization is during the inference time. Some of these works on the personalization of models in a decentralized setting can be found in Vanhaesebrouck et al. (2017); Almeida & Xavier (2018), where in addition to the optimization error, they have network constraints or peer-to-peer communication limitation (Bellet et al., 2017; Zantedeschi et al., 2019). In general, as discussed by Kairouz et al. (2019), there are three significant categories of personalization methods in federated learning, namely, local fine-tuning, multi-task learning, and contextualization. Yu et al. (2020) argue that the global model learned by federated learning, especially with having differential privacy and robust learning objectives, can hurt the performance of many clients. They indicate that those clients can obtain a better model by using only their own data. Hence, they empirically show that using these three approaches can boost the performance of those clients. In addition to these three, there is also another category that fits the most to our proposed approach, which is mixing the global and local models.

Local fine-tuning: The dominant approach for personalization is local fine-tuning, where each client receives a global model and tune it using its own local data and several gradient descent steps. This approach is predominantly used in meta-learning methods such as MAML by [Finn et al. \(2017\)](#) or domain adaptation and transfer learning ([Ben-David et al., 2010](#); [Mansour et al., 2009](#); [Pan & Yang, 2009](#)). [Jiang et al. \(2019\)](#) discuss the similarity between federated learning and meta-learning approaches, notably the Reptile algorithm by [Nichol et al. \(2018\)](#) and FedAvg, and combine them to personalize local models. They observed that federated learning with a single objective of performance of the global model could limit the capacity of the learned model for personalization. In [Khodak et al. \(2019\)](#), authors using online convex optimization to introduce a meta-learning approach that can be used in federated learning for better personalization. [Fallah et al. \(2020\)](#) borrow ideas from MAML to learn personalized models for each client with convergence guarantees. Similar to fine-tuning, they update the local models with several gradient steps, but they use second-order information to update the global model, like MAML. Another approach adopted for deep neural networks is introduced by [Arivazhagan et al. \(2019\)](#), where they freeze the base layers and only change the last “personalized” layer for each client locally. The main drawback of local fine-tuning is that it minimizes the optimization error, whereas the more important part is the generalization performance of the personalized model. In this setting, the personalized model is prone to overfit.

Multi-task learning: Another view of the personalization problem is to see it as a multi-task learning problem similar to [Smith et al. \(2017\)](#). In this setting, optimization on each client can be considered as a new task; hence, the approaches of multi-task learning can be applied. One other approach, discussed as an open problem in [Kairouz et al. \(2019\)](#), is to cluster groups of clients based on some features such as region, as similar tasks, similar to one approach proposed by [Mansour et al. \(2020\)](#).

Contextualization: An important application of personalization in federated learning is using the model under different contexts. For instance, in the next character recognition task in [Hard et al. \(2018\)](#), based on the context of the use case, the results should be different. Hence, we need a personalized model on one client under different contexts. This requires access to more features about the context during the training. Evaluation of the personalized model in such a setting has been investigated by [Wang et al. \(2019\)](#), which is in line with our approach in experimental results in Section 5. [Liang et al. \(2020\)](#) propose to directly learn the feature representation locally, and train the discriminator globally, which reduces the effect of data heterogeneity and ensures the fair learning.

Personalization via model regularization: Another significant trial for personalization is model regularization. There are several studies to introduce different personalization approaches for federated learning by regularize the difference between the global and local models. [Hanzely & Richtárik \(2020\)](#) try to introduce a new formulation for federated learning where they add the regularization term on the distance of local and global models. In their effort, they use a mixing parameter, which controls the degree of optimization for both local models and the global model. The FedAvg ([McMahan et al., 2017](#)) can be considered a special case of this approach. They show that the learned model is in the convex hull of both local and global models, and at each iteration, depend on the local models’ optimization parameters, the global model is getting closer to the global model learned by FedAvg. Similarly, [Huang et al. \(2020\)](#) and [Dinh et al. \(2020\)](#) also propose to use the regularization between local and global model, to realize the personalized learning. [Shen et al. \(2020\)](#) propose a knowledge distillation way to achieve personalization, where they apply the regularization on the predictions between local model and global model.

Personalization via model interpolation: Parallel to our work, there are other studies to introduce different personalization approaches for federated learning by mixing the global and local models. The closest approach for personalization to our proposal is introduced by [Mansour et al. \(2020\)](#). In fact, they propose three different approaches for personalization with generalization guarantees, namely, client clustering, data interpolation, and model interpolation. Out of these three, the first two approaches need some meta-features from all clients that makes them not a feasible approach for federated learning, due to privacy concerns. The third schema, which is the most promising one in practice as well, has a close formulation to ours in the interpolation of the local and global models. However, in their theory, the generalization bound does not demonstrate the advantage of mixing models, but in our analysis, we show how the model mixing can impact the generalization

bound, by presenting its dependency on the mixture parameter, data diversity and optimal models on local and global distributions.

Beyond different techniques for personalization in federated learning, Kairouz et al. (2019) ask an essential question of “*when is a global FL-trained model better?*”, or as we can ask, when is personalization better? The answer to these questions mostly depends on the distribution of data across clients. As we theoretically prove and empirically verify in this paper, when the data is distributed IID, we cannot benefit from personalization, and it is similar to the local SGD scenario (Stich, 2018; Haddadpour et al., 2019a;b; Woodworth et al., 2020b). However, when the data is non-IID across clients, which is mostly the case in federated learning, personalization can help to balance between shared and local knowledge. Then, the question becomes, what degree of personalization is best for each client? While this was an open problem in Mohri et al. (2019) on how to appropriately mix the global and local model, we answer this question by adaptively tuning the degree of personalization for each client, as discussed in Section 3, so it can perfectly become agnostic to the local data distributions.

B ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present additional experimental results to demonstrate the efficacy of the proposed APFL algorithm. First, we describe different datasets we have used in this paper, and then, present additional results.

B.1 DATASETS

For the experiments we use 4 different data sources as follows:

MNIST and CIFAR10 For the MNIST and CIFAR10 datasets to be similar to the setting in federated learning, we need to manually distribute them in a non-IID way, hence the data distribution is pathologically heterogeneous. To this end, we follow the steps used by McMahan et al. (2017), where they partitioned the dataset based on labels and for each client draw samples from some limited number of classes. We use the same way to create 3 datasets for the MNIST, that are, MNIST non-IID with 2 classes per client, MNIST non-IID with 4 classes per client, and MNIST IID, where the data is distributed uniformly random across different clients. Also, we create a non-IID CIFAR10 dataset, where each client has access to only 2 classes of data.

EMNIST In addition to pathological heterogeneous data distributions, we applied our algorithm on a real-world heterogeneous dataset, which is an extension to MNIST dataset. The EMNIST dataset includes images of characters divided by authors, where each author has a different style, make their distributions different Caldas et al. (2018). We use only digit characters and 1000 authors’ data to train our models on.

Synthetic For generating the synthetic dataset, we follow the procedure used by Li et al. (2018), where they use two parameters, say $\text{synthetic}(\gamma, \beta)$, that control how much the local model and the local dataset of each client differ from that of other clients, respectively. Using these parameters, we want to control the diversity between data and model of different clients. The procedure is that for each client we generate a weight matrix $\mathbf{W}_i \in \mathbb{R}^{m \times c}$ and a bias $\mathbf{b} \in \mathbb{R}^c$, where the output for the i th client is $y_i = \arg \max \left(\sigma \left(\mathbf{W}_i^\top \mathbf{x}_i + \mathbf{b} \right) \right)$, where $\sigma(\cdot)$ is the softmax. In this setting, the input data $\mathbf{x}_i \in \mathbb{R}^m$ has m features and the output y can have c different values indicating number of classes. The model is generated based on a Gaussian distribution $\mathbf{W}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, 1)$ and $\mathbf{b}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, 1)$, where $\boldsymbol{\mu}_i \sim \mathcal{N}(0, \gamma)$. The input is drawn from a Gaussian distribution $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\nu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\nu}_i \sim \mathcal{N}(V_i, 1)$ and $V_i \sim \mathcal{N}(0, \beta)$. Also the variance $\boldsymbol{\Sigma}$ is a diagonal matrix with value of $\Sigma_{k,k} = k^{-1.2}$. Using this procedure, we generate three different datasets, namely $\text{synthetic}(0.0, 0.0)$, $\text{synthetic}(0.5, 0.5)$, and $\text{synthetic}(1.0, 1.0)$, where we move from an IID dataset to a highly non-IID data.

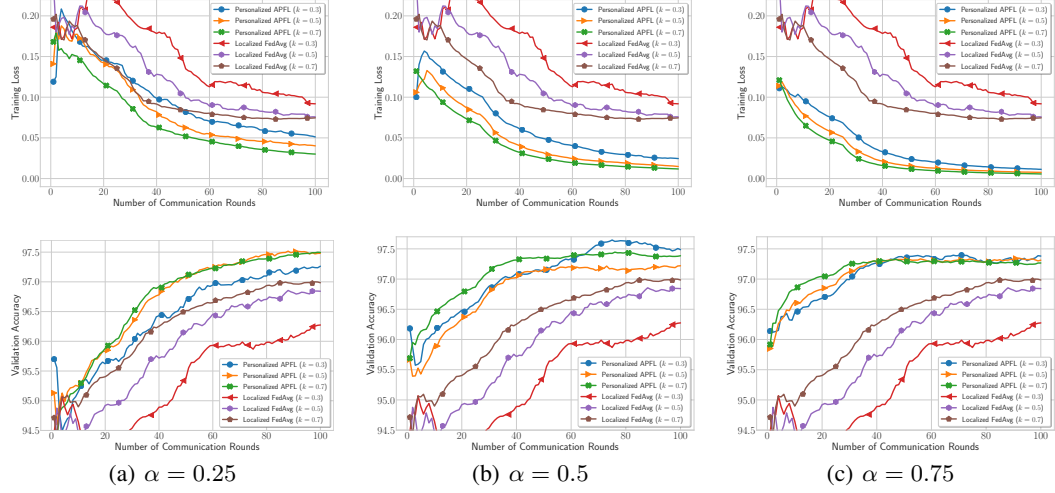


Figure 4: Evaluating the effect of sampling on APFL and FedAvg algorithm using the MNIST dataset that is non-IID with only 2 classes per client with logistic regression as the loss. The first row is training performance on the local model of FedAvg and personalized model of APFL with different sampling rates from $\{0.3, 0.5, 0.7\}$. The second row is the generalization performance of models on local validation data, aggregated over all clients. It can be inferred that despite the sampling ratio, APFL can superbly outperform FedAvg.

B.2 ADDITIONAL RESULTS

In this part, we present more experimental results that can further illustrate the effectiveness of APFL on other datasets and models.

Effect of sampling. To understand how the sampling of different clients will affect the performance of the APFL algorithm, we run the same experiment with different sampling rates for the MNIST dataset. The results of this experiment are depicted in Figure 4, where we run the experiment for different sampling rates of $K \in \{0.3, 0.5, 0.7\}$. Also, we run it with different values of $\alpha \in \{0.25, 0.5, 0.75\}$. The results are reported for the personalized model of APFL and localized FedAvg. As it can be inferred, decreasing the sampling ratio has a negative impact on both the training and generalization performance of FedAvg. However, we can see that despite the sampling ratio, APFL is outperforming local model of the FedAvg in both training and generalization. Also, from the results of Figure 2, we know that for this dataset that is highly non-IID, larger α values are preferred. Increasing α can diminish the negative impacts of sampling on personalized models both in training and generalization.

Natural heterogeneous data In addition to the CIFAR10 and MNIST datasets with pathological heterogeneous data distributions, we apply our algorithm on a natural heterogeneous dataset, EMNIST (Caldas et al., 2018). We use the data from 1000 clients, and for each round of communication we randomly select 10% of clients to participate in the training. We use an MLP model with 2 hidden layers, each with 200 neurons and ReLU as the activation function, using cross entropy as the loss function. For APFL, we use the adaptive α scheme with initial value of 0.5 for each client. We run both algorithms for 250 rounds of communication. In each round, each online client performs the local updates for 1 epoch on its data. Figure 5 shows the results of this experiment for personalized model of APFL and the localized model of the FedAvg. APFL with adaptive α can reach to the same training loss of the local FedAvg, while greatly outperforms the local FedAvg model in generalization on local validation data.

C DISCUSSIONS AND EXTENSIONS

Connection between learning guarantee and convergence. As Theorem 1 suggests, the generalization bound depends on the divergence of the local and global distributions. In the language

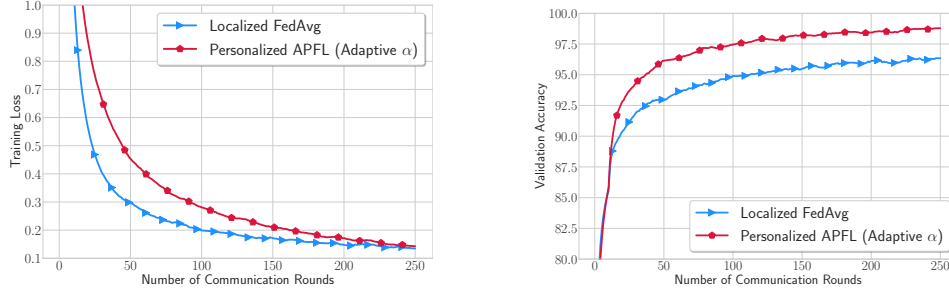


Figure 5: The results of applying FedAvg and APFL (with adaptive α) on an MLP model using EMNIST dataset, which is naturally heterogeneous. APFL achieves the same training loss of localized FedAvg, while outperforms it in validation accuracy.

of optimization, the counter-part of divergence of distribution is the gradient diversity; hence, the gradient diversity appears in our empirical loss convergence rate (Theorem 2). The other interesting discovery is in the generalization bound, we have the term $\lambda_{\mathcal{H}}$ and $\mathcal{L}_{\mathcal{D}_i}(h_i^*)$, which are intrinsic to the distributions and hypothesis class. Meanwhile, in the convergence result, we have the term $\|v_i^* - w^*\|^2$, which also only depends on the data distribution and hypothesis class we choose. In addition, $\|v_i^* - w^*\|^2$ also reveals the divergence between local and global optimal solutions.

Why APFL is “Adaptive”. Both information-theoretically (Theorem 1) and computationally (Theorem 2), we prove that when the local distribution drifts far away from the average distribution, the global model does not contribute too much to improve the local generalization and we have to tune the mixing parameter α to a larger value. Thus it is necessary to make α updated adaptively during empirical risk minimization. In Section 3, (6) shows that the update of α depends on the correlation of local gradient and deviation between local and global models. Experimental results show that our method can adaptively tune α , and can outperform the training scheme using fixed α .

Comparison with local ERM model A crucial question about personalization is *when it is preferable to employ a mixed model?*, and *how bad a local ERM model will be?* In the following corollary, we answer this by showing that the risk of local ERM model can be strictly worse than that of our personalized model.

Corollary 1. *Continuing with Theorem 1, there exist a distribution \mathcal{D}_i , constant C_1 and C_2 , such that with probability at least $1 - \delta$, the following upper bound for the difference between risks of personalized model h_{α_i} and local ERM model \hat{h}_i^* on \mathcal{D}_i , holds :*

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}) - \mathcal{L}_{\mathcal{D}_i}(\hat{h}_i^*) &\leq (2\alpha_i^2 - 1)\mathcal{L}_{\mathcal{D}_i}(h_i^*) + (2\alpha_i^2 C_1 - C_2) \sqrt{\frac{d + \log(1/\delta)}{m_i}} + 2\alpha_i^2 G \lambda_{\mathcal{H}}(\mathcal{S}_i) \\ &\quad + 2(1 - \alpha_i)^2 \left(\hat{\mathcal{L}}_{\bar{\mathcal{D}}}(\bar{h}^*) + B \|\bar{\mathcal{D}} - \mathcal{D}_i\|_1 + C_1 \sqrt{\frac{d + \log(1/\delta)}{m}} \right). \end{aligned}$$

By examining the above bound, the personalized model is preferable to local model if this value is less than 0. In this case, we require $(2\alpha_i^2 - 1)$ and $(2\alpha_i^2 C_1 - C_2)$ to be negative, which is satisfied by choosing $\alpha_i \leq \min\{\frac{\sqrt{2}}{2}, \sqrt{\frac{C_2}{2C_1}}\}$. Then, the term $\sqrt{\frac{d + \log(1/\delta)}{m_i}}$, should be sufficiently large, and the divergence term, as well as the global model generalization error has to be small. In this case, from the local model perspective, it can benefit from incorporate some global model. Using the similar technique, we can prove the supremacy of mixed model over global model as well.

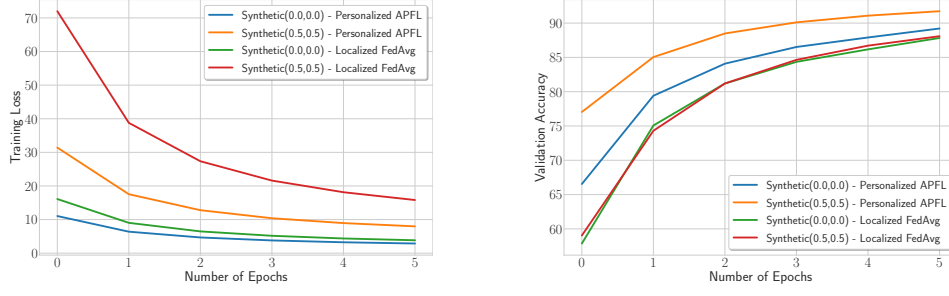


Figure 6: Comparing the effect of fine-tuning with the local model of FedAvg and with the personalized model of APFL on the synthetic datasets. The model is trained for 100 rounds of communication with 97 clients, and then 3 clients will join in fine-tuning the global model based on their own data. It can be seen that the model from APFL can better personalize the global model with respect to the FedAvg method both in training loss and validation accuracy. Increasing diversity makes it harder to personalize, however, APFL surpasses FedAvg again.

Proof of Corollary 1. Since in Theorem 1, we already obtained upper bound for $\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i})$ as following,

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}) &\leq 2\alpha_i^2 \left(\mathcal{L}_{\mathcal{D}_i}(h_i^*) + 2C_1 \sqrt{\frac{d + \log(1/\delta)}{m_i}} + G\lambda_{\mathcal{H}}(\mathcal{S}_i) \right) \\ &\quad + 2(1 - \alpha_i)^2 \left(\hat{\mathcal{L}}_{\bar{\mathcal{D}}}(\bar{h}^*) + B\|\bar{\mathcal{D}} - \mathcal{D}_i\|_1 + C_1 \sqrt{\frac{d + \log(1/\delta)}{m}} \right), \end{aligned}$$

to find the upper bound of $\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}) - \mathcal{L}_{\mathcal{D}_i}(\hat{h}_i^*)$, we just need the lower bound of $\mathcal{L}_{\mathcal{D}_i}(\hat{h}_i^*)$. The fundamental theorem of statistical learning (Shalev-Shwartz & Ben-David, 2014; Mohri et al., 2018) states a lower risk bound for agnostic PAC learning: for a hypothesis class with finite VC dimension d , then there exists a distribution \mathcal{D} , such that for any learning algorithm, which learns a hypothesis $h \in \mathcal{H}$ on m i.i.d. samples from \mathcal{D} , there exists a constant C , with the probability at least $1 - \delta$, we have:

$$\mathcal{L}_{\mathcal{D}}(h) - \min_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h') \geq C \sqrt{\frac{d + \log(1/\delta)}{m}}.$$

Since \hat{h}_i^* is learnt by ERM algorithm, the agnostic PAC learning lower risk bound also holds for it, so in worst case it might hold that under distribution \mathcal{D}_i , if \hat{h}_i^* is learnt by ERM algorithm using m_i samples, then there is a C_2 , such that with probability at least $1 - \delta$, we have:

$$\mathcal{L}_{\mathcal{D}_i}(\hat{h}_i^*) \geq \mathcal{L}_{\mathcal{D}_i}(h_i^*) + C_2 \sqrt{\frac{d + \log(1/\delta)}{m_i}}.$$

Thus we can bound $\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}) - \mathcal{L}_{\mathcal{D}_i}(\hat{h}_i^*)$ as Corollary 1 claims. \square

Personalization for new participant nodes. Suppose we already have a trained global model $\hat{\mathbf{w}}$, and now a new device k joins in the network, which is desired to personalize the global model to adapt its own domain. This can be done by performing a few local stochastic gradient descent updates from the given global model as an initial local model:

$$\mathbf{v}_k^{(t+1)} = \mathbf{v}_k^{(t)} - \eta_t \nabla_{\mathbf{v}} f_k(\alpha_k \mathbf{v}_k^{(t)} + (1 - \alpha_k) \hat{\mathbf{w}}; \xi_k^{(t)}) \quad (7)$$

to quickly learn a personalized model for the newly joined device. One thing worthy of investigation is the difference between APFL and meta-learning approaches, such as model-agnostic meta-learning (Finn et al., 2017). Our goal is to share the knowledge among the different users, in order to reduce the generalization error; while meta-learning cares more about how to build a meta-learner, to help training models faster and with fewer samples. In this scenario, similar to

FedAvg, when a new node joins the network, it gets the global model and takes a few stochastic steps based on its own data to update the global model. In Figure 6, we show the results of applying FedAvg and APFL on synthetic data with two different rates of diversity, `synthetic(0.0, 0.0)` and `synthetic(0.5, 0.5)`. In this experiment, we keep 3 nodes with their data off in the entire training for 100 rounds of communication between 97 nodes. In each round, each client updates its local and personalized models for one epoch. After the training is done, those 3 clients will join the network and get the latest global model and start training local and personalized models of their own. Figure 6 shows the training loss and validation accuracy of these 3 nodes during the 5 epochs of updates. The local model represents the model that will be trained in FedAvg, while the personalized model is the one resulting from APFL. Although the goal of APFL is to adaptively learn the personalized model during the training, it can be inferred that APFL can learn a better personalized model in a meta-learning scenario as well.

Agnostic global model. As pointed out by Mohri et al. (2019), the global model can be distributionally robust if we optimize the agnostic loss:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \max_{\mathbf{q} \in \Delta_n} F(\mathbf{w}) := \sum_i^n q_i f_i(\mathbf{w}), \quad (8)$$

where $\Delta_n = \{\mathbf{q} \in \mathbb{R}_+^n \mid \sum q_i = 1\}$ is the n -dimensional simplex. We call this scenario ‘‘Adaptive Personalized Agnostic Federated Learning’’. In this case, the analysis will be more challenging since the global empirical risk minimization is performed at a totally different domain, so the risk upper bound for h_{α_i} we derived does not hold anymore. Also, from a computational standpoint, since the resulted problem is a minimax optimization problem, the convergence analysis of agnostic APFL will be more involved, which we will leave as an interesting future work.

D PROOF OF GENERALIZATION BOUND

In this section we present the proof of generalization bound for APFL algorithm. Recall that we define the following hypotheses on i th local true and empirical distributions:

$$\begin{aligned}
\hat{h}_i^* &= \arg \min_{h \in \mathcal{H}} \hat{\mathcal{L}}_{\mathcal{D}_i}(h) && \text{(LOCAL EMPIRICAL RISK MINIMIZER)} \\
h_i^* &= \arg \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_i}(h) && \text{(LOCAL TRUE RISK MINIMIZER)} \\
\bar{h}^* &= \arg \min_{h \in \mathcal{H}} \mathcal{L}_{\bar{\mathcal{D}}}(h) && \text{(GLOBAL EMPIRICAL RISK MINIMIZER)} \\
\hat{h}_{loc,i}^* &= \arg \min_{h \in \mathcal{H}} \hat{\mathcal{L}}_{\mathcal{D}_i}(\alpha_i h + (1 - \alpha_i) \bar{h}^*) && \text{(MIXED EMPIRICAL RISK MINIMIZER)} \\
h_{loc,i}^* &= \arg \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_i}(\alpha_i h + (1 - \alpha_i) \bar{h}^*) && \text{(MIXED TRUE RISK MINIMIZER)}
\end{aligned}$$

where $\hat{\mathcal{L}}_{\mathcal{D}_i}(h)$ and $\mathcal{L}_{\mathcal{D}_i}(h)$ denote the empirical and true risks on \mathcal{D}_i , respectively.

From a high-level technical view, since we wish to bound the risk of the mixed model on local distribution \mathcal{D}_i , first we need to utilize the convex property of the risk function, and decompose it into two parts: $\mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*)$ and $\mathcal{L}_{\mathcal{D}_i}(\bar{h}^*)$. To bound $\mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*)$, a natural idea is to characterize it by the risk of optimal model $\mathcal{L}_{\mathcal{D}_i}(h_i^*)$, plus some excess risk. However, due to fact that $\hat{h}_{loc,i}^*$ is not the sole local empirical risk minimizer, rather it partially incorporates the global model, we need to characterize to what extent it drifts from the local empirical risk minimizer \hat{h}_i^* . This drift can be depicted by the hypothesis capacity, so that is our motivation to define $\lambda_{\mathcal{H}}(\mathcal{S})$ to quantify the empirical loss discrepancy over \mathcal{S} among pair of hypotheses in \mathcal{H} . We have to admit that there should be a tighter theory to bound this drift, depending how global model is incorporated, which we leave it as a future work.

The following simple result will be useful in the proof of generalization.

Lemma 1. *Let \mathcal{H} be a hypothesis class and \mathcal{D} and \mathcal{D}' denote two probability measures over space Ξ . Let $\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\ell(h(\mathbf{x}), y)]$ denote the risk of h over \mathcal{D} . If the loss function $\ell(\cdot)$ is bounded by B , then for every $h \in \mathcal{H}$:*

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_{\mathcal{D}'}(h) + B \|\mathcal{D} - \mathcal{D}'\|_1, \quad (9)$$

where $\|\mathcal{D} - \mathcal{D}'\|_1 = \int_{\Xi} |\mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} - \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}'}| d\mathbf{x} dy$.

Proof.

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(h) &\leq \mathcal{L}_{\mathcal{D}'}(h) + |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_{\mathcal{D}'}(h)| \\
&\leq \mathcal{L}_{\mathcal{D}}(h) + \int_{\Xi} |\ell(y, h(\mathbf{x}))| |\mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} - \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}'}| d\mathbf{x} dy \\
&= \mathcal{L}_{\mathcal{D}}(h) + B \|\mathcal{D} - \mathcal{D}'\|_1.
\end{aligned}$$

□

Proof of Theorem 1 We now turn to proving the generalization bound for the proposed APFL algorithm. Recall that for the classification task we consider squared hinge loss, and for the regression case we consider MSE loss. We will first prove that in both cases we can decompose the risk as follows:

$$\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}^*) \leq 2\alpha_i^2 \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) + 2(1 - \alpha_i)^2 \mathcal{L}_{\mathcal{D}_i}(\bar{h}^*(\mathbf{x})). \quad (10)$$

We start with the classification case first. Note that, hinge loss: $\max\{0, 1 - z\}$ is convex in z , so $\max\{0, 1 - y(\alpha_i h + (1 - \alpha_i) h')\} \leq \alpha_i \max\{0, 1 - y h\} + (1 - \alpha_i) \max\{0, 1 - y h'\}$, according to

Jensen's inequality. Hence, we have:

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}^*) &= \mathcal{L}_{\mathcal{D}_i}(\alpha_i \hat{h}_{loc,i}^* + (1 - \alpha_i) \bar{h}^*) \\
&= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left(\max\{0, 1 - y(\alpha_i \hat{h}_{loc,i}^*(\mathbf{x}) + (1 - \alpha_i) \bar{h}^*(\mathbf{x}))\} \right)^2 \\
&= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left(\alpha_i \max\{0, 1 - y \hat{h}_{loc,i}^*(\mathbf{x})\} + (1 - \alpha_i) \max\{0, 1 - y \bar{h}^*(\mathbf{x})\} \right)^2 \\
&\leq 2\alpha_i^2 \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left(\max\{0, 1 - y \hat{h}_{loc,i}^*(\mathbf{x})\} \right)^2 \\
&\quad + 2(1 - \alpha_i)^2 \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left(\max\{0, 1 - y \bar{h}^*(\mathbf{x})\} \right)^2 \\
&\leq 2\alpha_i^2 \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) + 2(1 - \alpha_i)^2 \mathcal{L}_{\mathcal{D}_i}(\bar{h}^*).
\end{aligned}$$

For regression case:

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}^*) &= \mathcal{L}_{\mathcal{D}_i}(\alpha_i \hat{h}_{loc,i}^* + (1 - \alpha_i) \bar{h}^*) \\
&= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left\| y - (\alpha_i \hat{h}_{loc,i}^*(\mathbf{x}) + (1 - \alpha_i) \bar{h}^*(\mathbf{x})) \right\|^2 \\
&= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left\| \alpha_i y - \alpha_i \hat{h}_{loc,i}^*(\mathbf{x}) + (1 - \alpha_i) y - (1 - \alpha_i) \bar{h}^*(\mathbf{x}) \right\|^2 \\
&\leq 2\alpha_i^2 \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left\| y - \hat{h}_{loc,i}^*(\mathbf{x}) \right\|^2 + 2(1 - \alpha_i)^2 \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_i} \left\| y - \bar{h}^*(\mathbf{x}) \right\|^2 \\
&\leq 2\alpha_i^2 \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) + 2(1 - \alpha_i)^2 \mathcal{L}_{\mathcal{D}_i}(\bar{h}^*)
\end{aligned}$$

Thus we can conclude:

$$\mathcal{L}_{\mathcal{D}_i}(h_{\alpha_i}^*) \leq \underbrace{2\alpha_i^2 \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*)}_{T_1} + \underbrace{2(1 - \alpha_i)^2 \mathcal{L}_{\mathcal{D}_i}(\bar{h}^*)}_{T_2}. \quad (11)$$

We proceed to bound the terms T_1 and T_2 in RHS of above inequality. We first bound T_1 as follows. The first step is to utilize uniform VC dimension error bound over \mathcal{H} [Mohri et al. \(2018\)](#); [Shalev-Shwartz & Ben-David \(2014\)](#):

$$\forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{D}_i}(h) - \hat{\mathcal{L}}_{\mathcal{D}_i}(h)| \leq C \sqrt{\frac{d + \log(1/\delta)}{m_i}},$$

where C is constant factor. So we can bound T_1 as:

$$\begin{aligned}
T_1 &= \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) = \mathcal{L}_{\mathcal{D}_i}(h_i^*) + \mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) - \mathcal{L}_{\mathcal{D}_i}(h_i^*) \\
&= \mathcal{L}_{\mathcal{D}_i}(h_i^*) \\
&\quad + \underbrace{\mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) - \hat{\mathcal{L}}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*)}_{\leq C \sqrt{\frac{d + \log(1/\delta)}{m_i}}} + \hat{\mathcal{L}}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) - \hat{\mathcal{L}}_{\mathcal{D}_i}(h_i^*) + \underbrace{\hat{\mathcal{L}}_{\mathcal{D}_i}(h_i^*) - \mathcal{L}_{\mathcal{D}_i}(h_i^*)}_{\leq C \sqrt{\frac{d + \log(1/\delta)}{m_i}}} \\
&\leq \mathcal{L}_{\mathcal{D}_i}(h_i^*) + 2C \sqrt{\frac{d + \log(1/\delta)}{m_i}} + \hat{\mathcal{L}}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) - \hat{\mathcal{L}}_{\mathcal{D}_i}(h_i^*).
\end{aligned}$$

Note that

$$\hat{\mathcal{L}}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*) - \hat{\mathcal{L}}_{\mathcal{D}_i}(h_i^*) \leq G \frac{1}{|\mathcal{S}_i|} \sum_{(\mathbf{x}, y) \in \mathcal{S}_i} |\hat{h}_{loc,i}^*(\mathbf{x}) - \hat{h}_i^*(\mathbf{x})| \leq G \lambda_{\mathcal{H}}(\mathcal{S}_i),$$

As a result we can bound T_1 by:

$$T_1 \leq \mathcal{L}_{\mathcal{D}_i}(h_i^*) + 2C \sqrt{\frac{d + \log(1/\delta)}{m_i}} + G \lambda_{\mathcal{H}}(\mathcal{S}_i).$$

We now turn to bounding T_2 . Plugging Lemma 1 in (11) and using uniform generalization risk bound will immediately give:

$$T_2 \leq \hat{\mathcal{L}}_{\bar{\mathcal{D}}}(\bar{h}^*) + B^2 \|\mathcal{D} - \bar{\mathcal{D}}\|_1 + C \sqrt{\frac{d + \log(1/\delta)}{m}}.$$

Plugging T_1 and T_2 back into (11) concludes the proof. \square

Remark 2. One thing worth mentioning is that, we assume the customary boundedness of loss functions. Actually it can be satisfied if the data and the parameters of hypothesis are bounded. For example, considering the scenario where we are learning a linear model \mathbf{w} with the constraint $\|\mathbf{w}\| \leq 1$, and also the data tuples (\mathbf{x}, y) are drawn from some bounded domain, then the loss is obviously bounded by some finite real value.

Remark 3. As $\mathcal{L}_{\mathcal{D}_i}(\hat{h}_{loc,i}^*)$ is the risk of the empirical risk minimizer on \mathcal{D}_i after incorporating a model learned on a different domain (i.e., global distribution), one might argue that generalization techniques established in multi-domain learning theory (Ben-David et al., 2010; Mansour et al., 2009; Zhang et al., 2020) can be utilized to serve our purpose. However, we note that the techniques developed in Ben-David et al. (2010); Mansour et al. (2009); Zhang et al. (2020) are only applicable to a settings where we aim at directly learning a model in some combination of source and target domain, while in our setting, we partially incorporate the model learned from source domain and then perform ERM on joint model over target domain. Moreover, their results only apply to very simple loss functions, e.g., absolute loss or MSE loss, while we consider squared hinge loss in the classification case. Analogous to multiple domain theory, we derive the multi domain learning bound based on the divergence of source and target domains but measured in absolute distance, $\|\cdot\|_1$. As Mansour et al. (2009) points out, divergence measured by absolute loss can be large, and as a result we leave the development of a more general multiple domain learning theory that can deal with most popular loss functions like hinge loss, cross entropy loss and optimal transport, with tighter divergence measure on distributions as an open question.

E PROOF OF CONVERGENCE RATE IN CONVEX SETTING

In this section, we present the proof of convergence rates. For ease of mathematical derivations, we first consider the case *without sampling clients* at each communication step and then generalize the proof to the setting where K devices are sampled uniformly at random by the server as employed in the proposed algorithm.

Technical challenges. The analysis of convergence rates in our setting is more involved compared to analysis of local SGD with periodic averaging by Stich (2018); Woodworth et al. (2020a). The key difficulty arises from the fact that unlike local SGD where local solutions are evolved by employing mini-batch SGD, in our setting we also partially incorporate the global model to compute stochastic gradients over local data. In addition, our goal is to find the convergence rate of the mixed model, rather than merely the local model or global model. To better illustrate this, let us first clarify the notations of models that will be used in analysis. Let us consider the simple case for now where we set $K = n$ (all device participate averaging). We define three virtual sequences: $\{\mathbf{w}^{(t)}\}_{t=1}^T$, $\{\bar{\mathbf{v}}^{(t)}\}_{t=1}^T$ and $\{\hat{\mathbf{v}}^{(t)}\}_{t=1}^T$ where $\mathbf{w}^{(t)} = \frac{1}{n} \sum_{j=1}^n \mathbf{w}_j^{(t)}$, $\bar{\mathbf{v}}^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}_i^{(t)}$, $\hat{\mathbf{v}}^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}^{(t)}$. Since the personalized model incorporates $1 - \alpha_i$ percentage of global model, then the key challenge in the convergence analysis is to find out how much the global model benefits/hurts the local convergence. To this end, we analyze how much the dynamics of personalized model $\hat{\mathbf{v}}_i^{(t)}$ and global model $\mathbf{w}^{(t)}$ differ from each other at each iteration. To be more specific, we study the distance between gradients $\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\|^2$. Surprisingly, we relate this distance to gradient diversity, personalized model convergence, global model convergence and local-global optimality gap:

$$\mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \leq 6\zeta_i + 2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + 6L^2 \Delta_i.$$

$\mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right]$ and $\mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right]$ will converge very fast under smooth strongly convex objective, and ζ_i and Δ_i will serve as residual error that indicates the heterogeneity among local functions.

Algorithm 2: Local Descent APFL (without sampling)

input: Mixture weights $\alpha_1, \dots, \alpha_n$, Synchronization gap τ , Local models $\mathbf{v}_i^{(0)}$ for $i \in [n]$ and local version of global model $\mathbf{w}_i^{(0)}$ for $i \in [n]$.

for $t = 0, \dots, T$ **do**

if t *not divides* τ **then**

$\mathbf{w}_i^{(t)} = \mathbf{w}_i^{(t-1)} - \eta_t \nabla f_i(\mathbf{w}_i^{(t-1)}; \xi_i^t)$

$\mathbf{v}_i^{(t)} = \mathbf{v}_i^{(t-1)} - \eta_t \nabla_{\mathbf{v}} f_i(\mathbf{v}_i^{(t-1)}; \xi_i^t)$

$\bar{\mathbf{v}}_i^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}_i^{(t)}$

else

each client sends $\mathbf{w}_j^{(t)}$ to the server

$\mathbf{w}^{(t)} = \frac{1}{n} \sum_{j=1}^n \mathbf{w}_j^{(t)}$

server broadcast $\mathbf{w}^{(t)}$ to all clients

end

end

for $i = 1, \dots, n$ **do**

output: Personalized model: $\hat{\mathbf{v}}_i = \frac{1}{S_T} \sum_{t=1}^T p_t (\alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \mathbf{w}_j^{(t)})$;

Global model: $\hat{\mathbf{w}} = \frac{1}{n S_T} \sum_{t=1}^T p_t \sum_{j=1}^n \mathbf{w}_j^{(t)}$.

end

E.1 PROOF WITHOUT SAMPLING

Before giving the proof of convergence analysis of the Algorithm 1 in the main paper, we first discuss a warm-up case: local descent APFL without client sampling. As Algorithm 2 shows, all clients will participate in the averaging stage every τ iterations. The convergence of global and local models in Algorithm 2 are given in the following theorems. We start by stating the convergence of global model.

Theorem 4 (Global model convergence of Local Descent APFL without Sampling). *If each client's objective function is μ -strongly convex and L -smooth, and satisfies Assumption 1, using Algorithm 2, choosing the mixing weight $\alpha_i \geq \max\{1 - \frac{1}{4\sqrt{6}\kappa}, 1 - \frac{1}{4\sqrt{6}\kappa\sqrt{\mu}}\}$, learning rate $\eta_t = \frac{16}{\mu(t+a)}$, where $a = \max\{128\kappa, \tau\}$, and using average scheme $\hat{\mathbf{w}} = \frac{1}{n S_T} \sum_{t=1}^T p_t \sum_{j=1}^n \mathbf{w}_j^{(t)}$, where $p_t = (t+a)^2$, $S_T = \sum_{t=1}^T p_t$, then the following convergence holds:*

$$\mathbb{E}[F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \leq O\left(\frac{\mu}{T^3}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + \tau \frac{\zeta}{n})}{\mu T^2}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + \tau \frac{\zeta}{n}) \ln T}{\mu T^3}\right) + O\left(\frac{\sigma^2}{nT}\right),$$

where $\mathbf{w}_* = \arg \min_{\mathbf{w}} F(\mathbf{w})$ is the optimal global solution.

Proof. Proof is deferred to Appendix E.1.2. □

The following theorem obtains the convergence of personalized model in Algorithm 2.

Theorem 5 (Personalized model convergence of Local Descent APFL without Sampling). *If each client's objective function is μ -strongly convex and L -smooth, and satisfies Assumption 1, using Algorithm 2, choosing the mixing weight $\alpha_i \geq \max\{1 - \frac{1}{4\sqrt{6}\kappa}, 1 - \frac{1}{4\sqrt{6}\kappa\sqrt{\mu}}\}$, learning rate $\eta_t = \frac{16}{\mu(t+a)}$, where $a = \max\{128\kappa, \tau\}$, and using average scheme $\hat{\mathbf{v}}_i = \frac{1}{S_T} \sum_{t=1}^T p_t (\alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \mathbf{w}_j^{(t)})$, where $p_t = (t+a)^2$, $S_T = \sum_{t=1}^T p_t$, and f_i^* is the local minimum of the i th client, then the following convergence holds for all $i \in [n]$:*

$$\begin{aligned} \mathbb{E}[f_i(\hat{\mathbf{v}}_i)] - f_i^* &\leq O\left(\frac{\mu}{T^3}\right) + \alpha_i^2 O\left(\frac{\sigma^2}{\mu T}\right) + (1 - \alpha_i)^2 O\left(\frac{\zeta_i}{\mu} + \kappa L \Delta_i\right) \\ &+ (1 - \alpha_i)^2 \left(O\left(\frac{\kappa L \ln T}{T^3}\right) + O\left(\frac{\kappa^2 \sigma^2}{\mu n T}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + \tau (\zeta_i + \frac{\zeta}{n}))}{\mu T^2}\right) + O\left(\frac{\kappa^4 \tau (\sigma^2 + 2\tau \frac{\zeta}{n})}{\mu T^2}\right) \right). \end{aligned}$$

Proof. Proof is deferred to Appendix E.1.3. \square

E.1.1 PROOF OF USEFUL LEMMAS

Before giving the proof of Theorem 4 and 5, we first prove few useful lemmas. Recall that we define virtual sequences $\{\mathbf{w}^{(t)}\}_{t=1}^T, \{\bar{\mathbf{v}}_i^{(t)}\}_{t=1}^T, \{\hat{\mathbf{v}}_i^{(t)}\}_{t=1}^T$ where $\mathbf{w}^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i^{(t)}, \bar{\mathbf{v}}_i^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}_i^{(t)}, \hat{\mathbf{v}}_i^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}^{(t)}$.

We start with the following lemma that bounds the difference between the gradients of local objective and global objective at local and global models.

Lemma 2. *For Algorithm 2, at each iteration, the gap between local gradient and global gradient is bounded by*

$$\mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \leq 2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6\zeta_i + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + 6L^2 \Delta_i.$$

Proof. From the smoothness assumption and by applying the Jensen's inequality we have:

$$\begin{aligned} & \mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \\ & \leq 2\mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla f_i(\mathbf{v}_i^*)\|^2 \right] + 2\mathbb{E} \left[\|\nabla f_i(\mathbf{v}_i^*) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \\ & \leq 2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6\mathbb{E} \left[\|\nabla f_i(\mathbf{v}_i^*) - \nabla f_i(\mathbf{w}^*)\|^2 \right] \\ & \quad + 6\mathbb{E} \left[\|\nabla f_i(\mathbf{w}^*) - \nabla F(\mathbf{w}^*)\|^2 \right] + 6\mathbb{E} \left[\|\nabla F(\mathbf{w}^*) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \\ & \leq 2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6L^2 \mathbb{E} \left[\|\mathbf{v}_i^* - \mathbf{w}^*\|^2 \right] + 6\zeta_i + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] \\ & \leq 2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6L^2 \Delta_i + 6\zeta_i + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right]. \end{aligned}$$

\square

Lemma 3 (Local model deviation without sampling). *For Algorithm 2, at each iteration, the deviation between each local version of the global model $\mathbf{w}_i^{(t)}$ and the global model $\mathbf{w}^{(t)}$ is bounded by:*

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] & \leq 3\tau\sigma^2\eta_{t-1}^2 + 3\left(\zeta_i + \frac{\zeta}{n}\right)\tau^2\eta_{t-1}^2, \\ \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] & \leq 3\tau\sigma^2\eta_{t-1}^2 + 6\tau^2\frac{\zeta}{n}\eta_{t-1}^2, \end{aligned}$$

where $\frac{\zeta}{n} = \frac{1}{n} \sum_{i=1}^n \zeta_i$.

Proof. According to Lemma 8 in Woodworth et al. (2020a):

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] & \leq \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\|\mathbf{w}_j^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\ & \leq 3 \left(\sigma^2 + \zeta_i\tau + \frac{\zeta}{n}\tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} (1 - \mu\eta_q) \\ \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] & \leq \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E} \left[\|\mathbf{w}_j^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\ & \leq 3 \left(\sigma^2 + 2\tau\frac{\zeta}{n} \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} (1 - \mu\eta_q). \end{aligned}$$

Plugging in $\eta_q = \frac{16}{\mu(a+q)}$ yields:

$$\begin{aligned}
\mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] &\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} \frac{a+q-16}{a+q} \\
&\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} \frac{a+q-16}{a+q} \\
&\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} \frac{a+q-2}{a+q} \\
&\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \frac{(a+p-1)(a+p)}{(a+t-2)(a+t-1)} \\
&\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \frac{\eta_{t-1}^2}{\eta_p^2} \\
&\leq 3 \tau \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{n} \tau \right) \eta_{t-1}^2.
\end{aligned}$$

Similarly,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \leq 3 \tau \sigma^2 \eta_{t-1}^2 + 6 \tau^2 \frac{\zeta}{n} \eta_{t-1}^2.$$

□

Lemma 4. (Convergence of global model) Let $\mathbf{w}^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i^{(t)}$. Under the setting of Theorem 5, we have:

$$\begin{aligned}
\mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{a^3}{(T+a)^3} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] \\
&+ \left(T + 16 \left(\frac{1}{a+1} + \ln(T+a) \right) \right) \frac{1536a^2\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) L^2}{(a-1)^2 \mu^4 (T+a)^3} + \frac{128\sigma^2 T(T+2a)}{n\mu^2 (T+a)^3}.
\end{aligned}$$

Proof. By the updating rule we have:

$$\mathbf{w}^{(t+1)} - \mathbf{w}^* = \mathbf{w}^{(t)} - \mathbf{w}^* - \eta_t \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}; \xi_j^t).$$

Then, taking square of norm and expectation on both sides, as well as applying strong convexity and smoothness assumptions yields:

$$\begin{aligned}
&\mathbb{E} \left[\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 \right] \\
&\leq \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] - 2\eta_t \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \right] \\
&+ \eta_t^2 \frac{\sigma^2}{n} + \eta_t^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\leq \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] - 2\eta_t \mathbb{E} \left[\left\langle \nabla F(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \right] + \eta_t^2 \frac{\sigma^2}{n} + \eta_t^2 \underbrace{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right]}_{T_1} \\
&\underbrace{- 2\eta_t \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla F(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \right]}_{T_2} \\
&\leq (1 - \mu\eta_t) \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] - 2\eta_t (\mathbb{E}[F(\mathbf{w}^{(t)})] - F(\mathbf{w}^*)) + \eta_t^2 \frac{\sigma^2}{n} + T_1 + T_2,
\end{aligned} \tag{12}$$

where at the last step we used the strongly convex property.

Now we are going to bound T_1 . By the Jensen's inequality and smoothness, we have:

$$\begin{aligned} T_1 &\leq 2\eta_t^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] + 2\eta_t^2 \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] \\ &\leq 2\eta_t^2 L^2 \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + 4\eta_t^2 L \left(\mathbb{E} \left[F(\mathbf{w}^{(t)}) \right] - F(\mathbf{w}^*) \right) \end{aligned} \quad (13)$$

Then, we bound T_2 as:

$$\begin{aligned} T_2 &\leq \eta_t \left(\frac{2}{\mu} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] + \frac{\mu}{2} \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}^* \right\|^2 \right] \right) \\ &\leq \frac{2\eta_t L^2}{\mu} \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + \frac{\mu \eta_t}{2} \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}^* \right\|^2 \right]. \end{aligned} \quad (14)$$

Now, by plugging back T_1 and T_2 from (13) and (14) in (12), we have:

$$\begin{aligned} &\mathbb{E} \left[\left\| \mathbf{w}^{(t+1)} - \mathbf{w}^* \right\|^2 \right] \\ &\leq \left(1 - \frac{\mu \eta_t}{2} \right) \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}^* \right\|^2 \right] \underbrace{- (2\eta_t - 4\eta_t^2 L) \left(\mathbb{E} \left[F(\mathbf{w}^{(t)}) \right] - F(\mathbf{w}^*) \right)}_{\leq -\eta_t} + \eta_t^2 \frac{\sigma^2}{n} \\ &\quad + \left(\frac{2\eta_t L^2}{\mu} + 2\eta_t^2 L^2 \right) \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] \\ &\leq \left(1 - \frac{\mu \eta_t}{2} \right) \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}^* \right\|^2 \right] + \eta_t^2 \frac{\sigma^2}{n} + \left(\frac{2\eta_t L^2}{\mu} + 2\eta_t^2 L^2 \right) \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right]. \end{aligned} \quad (15)$$

Now, by using Lemma 3 we have:

$$\begin{aligned} &\mathbb{E} \left[\left\| \mathbf{w}^{(t+1)} - \mathbf{w}^* \right\|^2 \right] \\ &\leq \left(1 - \frac{\mu \eta_t}{2} \right) \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}^* \right\|^2 \right] + \left(\frac{2\eta_t L^2}{\mu} + 2\eta_t^2 L^2 \right) 3\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) \eta_{t-1}^2 + \eta_t^2 \frac{\sigma^2}{n}. \end{aligned}$$

Note that $(1 - \frac{\mu \eta_t}{2}) \frac{p_t}{\eta_t} = \frac{\mu(t+a)^2(t-8+a)}{16} \leq \frac{\mu(t-1+a)^3}{16} = \frac{p_{t-1}}{\eta_{t-1}}$, so we multiply $\frac{p_t}{\eta_t}$ on both sides and do the telescoping sum:

$$\begin{aligned} &\frac{p_T}{\eta_T} \mathbb{E} \left[\left\| \mathbf{w}^{(T+1)} - \mathbf{w}^* \right\|^2 \right] \\ &\leq \frac{p_0}{\eta_0} \mathbb{E} \left[\left\| \mathbf{w}^{(1)} - \mathbf{w}^* \right\|^2 \right] + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) p_t \eta_{t-1}^2 + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{n} \\ &\leq \frac{p_0}{\eta_0} \mathbb{E} \left[\left\| \mathbf{w}^{(1)} - \mathbf{w}^* \right\|^2 \right] + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) \frac{256a^2}{\mu^2(a-1)^2} + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{n}. \end{aligned} \quad (16)$$

Then, by re-arranging the terms will conclude the proof:

$$\begin{aligned} &\mathbb{E} \left[\left\| \mathbf{w}^{(T+1)} - \mathbf{w}^* \right\|^2 \right] \\ &\leq \frac{a^3}{(T+a)^3} \mathbb{E} \left[\left\| \mathbf{w}^{(1)} - \mathbf{w}^* \right\|^2 \right] \\ &\quad + \left(T + 16 \left(\frac{1}{a+1} + \ln(T+a) \right) \right) \frac{1536a^2\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) L^2}{(a-1)^2 \mu^4 (T+a)^3} + \frac{128\sigma^2 T(T+2a)}{n\mu^2 (T+a)^3}, \end{aligned}$$

where we use the inequality $\sum_{t=1}^T \frac{1}{t+a} \leq \frac{1}{a+1} + \int_1^T \frac{1}{t+a} < \frac{1}{a+1} + \ln(T+a)$. \square

E.1.2 PROOF OF THEOREM 4

Proof. According to (15) and (16) in the proof of Lemma 4 we have:

$$\begin{aligned} \frac{p_T}{\eta_T} \mathbb{E} [\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2] &\leq \frac{p_0}{\eta_0} \mathbb{E} [\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2] - \sum_{t=1}^T p_t \left(\mathbb{E} [F(\mathbf{w}^{(t)})] - F(\mathbf{w}^*) \right) \\ &\quad + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) \frac{256a^2}{\mu^2(a-1)^2} + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{n}, \end{aligned}$$

re-arranging term and dividing both sides by $S_T = \sum_{t=1}^T p_t > T^3$ yields:

$$\begin{aligned} \frac{1}{S_T} \sum_{t=1}^T p_t \left(\mathbb{E} [F(\mathbf{w}^{(t)})] - F(\mathbf{w}^*) \right) &\leq \frac{p_0}{S_T \eta_0} \mathbb{E} [\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2] \\ &\quad + \frac{1}{S_T} \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\tau \frac{\zeta}{n} \right) \frac{256a^2}{\mu^2(a-1)^2} + \frac{1}{S_T} \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{n} \\ &\leq O\left(\frac{\mu}{T^3}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{n})}{\mu T^2}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{n}) \ln T}{\mu T^3}\right) + O\left(\frac{\sigma^2}{nT}\right). \end{aligned}$$

Recall that $\hat{\mathbf{w}} = \frac{1}{nS_T} \sum_{t=1}^T \sum_{j=1}^n \mathbf{w}_j^{(t)}$ and convexity of F , we can conclude that:

$$\mathbb{E} [F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \leq O\left(\frac{\mu}{T^3}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{n})}{\mu T^2}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{n}) \ln T}{\mu T^3}\right) + O\left(\frac{\sigma^2}{nT}\right).$$

□

E.1.3 PROOF OF THEOREM 5

Proof. Recall that we defined virtual sequences $\{\mathbf{w}^{(t)}\}_{t=1}^T$ where $\mathbf{w}^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i^{(t)}$ and $\hat{\mathbf{v}}_i^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}^{(t)}$, then by the updating rule we have:

$$\begin{aligned} &\mathbb{E} [\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2] \\ &= \mathbb{E} \left[\left\| \hat{\mathbf{v}}_i^{(t)} - \alpha_i^2 \eta_t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i) \eta_t \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \mathbf{v}_i^* \right\|^2 \right] \\ &\quad + \mathbb{E} \left[\left\| \alpha_i^2 \eta_t (\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}; \xi_i^t)) + (1 - \alpha_i) \eta_t \frac{1}{n} \sum_{j \in U_t} (\nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_j(\mathbf{w}_j^{(t)}; \xi_j^t)) \right\|^2 \right] \\ &\leq \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2] - 2\mathbb{E} \left[\left\langle \alpha_i^2 \eta_t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \eta_t \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right] \\ &\quad + \eta_t^2 \mathbb{E} \left[\left\| \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] + \alpha_i^2 \eta_t^2 \sigma^2 + (1 - \alpha_i)^2 \eta_t^2 \frac{\sigma^2}{n} \\ &= \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2] - \underbrace{2(\alpha_i^2 + 1 - \alpha_i) \eta_t \mathbb{E} \left[\left\langle \nabla f_i(\bar{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right]}_{T_1} \\ &\quad - \underbrace{2\eta_t (1 - \alpha_i) \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right]}_{T_2} \\ &\quad + \underbrace{\eta_t^2 \mathbb{E} \left[\left\| \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right]}_{T_3} + \alpha_i^2 \eta_t^2 \sigma^2 + (1 - \alpha_i)^2 \eta_t^2 \frac{\sigma^2}{n}. \end{aligned} \tag{17}$$

Now, we bound the term T_1 as follows:

$$\begin{aligned}
T_1 &= -2\eta_t(\alpha_i^2 + 1 - \alpha_i)\mathbb{E}\left[\left\langle \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle\right] \\
&\quad - 2\eta_t(\alpha_i^2 + 1 - \alpha_i)\mathbb{E}\left[\left\langle \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle\right] \\
&\leq -2\eta_t(\alpha_i^2 + 1 - \alpha_i)\left(\mathbb{E}\left[f_i(\hat{\mathbf{v}}_i^{(t)})\right] - f_i(\mathbf{v}_i^*) + \frac{\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]\right) \\
&\quad + (\alpha_i^2 + 1 - \alpha_i)\eta_t\left(\frac{8L^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \bar{\mathbf{v}}_i^{(t)}\|^2\right] + \frac{\mu(1 - 8(\alpha_i - \alpha_i^2))}{8}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]\right) \\
&\leq -2\eta_t(\alpha_i^2 + 1 - \alpha_i)\left(\mathbb{E}\left[f_i(\hat{\mathbf{v}}_i^{(t)})\right] - f_i(\mathbf{v}_i^*) + \frac{\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]\right) \\
&\quad + \eta_t\left(\frac{8L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))}\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2\right] + \frac{\mu(1 - 8(\alpha_i - \alpha_i^2))}{8}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]\right) \\
&\leq -2\eta_t(\alpha_i^2 + 1 - \alpha_i)\left(\mathbb{E}\left[f_i(\hat{\mathbf{v}}_i^{(t)})\right] - f_i(\mathbf{v}_i^*)\right) - \frac{7\mu\eta_t}{8}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right] \\
&\quad + \frac{8\eta_t L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))}\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2\right], \tag{18}
\end{aligned}$$

where we use the fact $(\alpha_i^2 + 1 - \alpha_i) \leq 1$. Note that, because we set $\alpha_i \geq \max\{1 - \frac{1}{4\sqrt{6}\kappa}, 1 - \frac{1}{4\sqrt{6}\kappa\sqrt{\mu}}\}$, and hence $1 - 8(\alpha_i - \alpha_i^2) \geq 0$, so in the second inequality we can use the arithmetic-geometry inequality.

Next, we turn to bounding the term T_2 in (17):

$$\begin{aligned}
T_2 &= -2\eta_t(1 - \alpha_i)\mathbb{E}\left[\left\langle \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle\right] \\
&\leq \eta_t(1 - \alpha_i)\left(\frac{2(1 - \alpha_i)}{\mu}\mathbb{E}\left[\left\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)})\right\|^2\right] + \frac{\mu}{2(1 - \alpha_i)}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]\right) \\
&\leq \frac{6(1 - \alpha_i)^2\eta_t}{\mu}\left(\mathbb{E}\left[\left\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)})\right\|^2\right] + \mathbb{E}\left[\left\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\right\|^2\right]\right. \\
&\quad \left. + \mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(t)}) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)})\right\|^2\right]\right) + \frac{\eta_t\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right] \\
&\leq \frac{6(1 - \alpha_i)^2\eta_t}{\mu}\left(L^2\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2\right] + \mathbb{E}\left[\left\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\right\|^2\right]\right. \\
&\quad \left. + \mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{(t)}) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)})\right\|^2\right]\right) + \frac{\eta_t\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2\right]. \tag{19}
\end{aligned}$$

And finally, we bound the term T_3 in (17) as follows:

$$\begin{aligned}
T_3 &= \mathbb{E}\left[\left\|\alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i)\frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)})\right\|^2\right] \\
&\leq 2(\alpha_i^2 + 1 - \alpha_i)^2\mathbb{E}\left[\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)})\|^2\right] + 2\mathbb{E}\left[\left\|(1 - \alpha_i)\left(\frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)})\right)\right\|^2\right] \\
&\leq 2\left(2(\alpha_i^2 + 1 - \alpha_i)^2\mathbb{E}\left[\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i^*\|^2\right] + 2(\alpha_i^2 + 1 - \alpha_i)^2\mathbb{E}\left[\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2\right]\right) \\
&\quad + 2(1 - \alpha_i)^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)})\right\|^2\right] \\
&\leq 8L(\alpha_i^2 + 1 - \alpha_i)\left(\mathbb{E}\left[f_i(\hat{\mathbf{v}}_i^{(t)})\right] - f_i^*\right) + 4(1 - \alpha_i)^2L^2\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2\right] \\
&\quad + 6(1 - \alpha_i)^2\left(L^2\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2\right] + \mathbb{E}\left[\left\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\right\|^2\right]\right. \\
&\quad \left. + \frac{1}{n}\sum_{j=1}^n L^2\mathbb{E}\left[\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2\right]\right). \tag{20}
\end{aligned}$$

Now, using Lemma 3, $(1 - \alpha_i)^2 \leq 1$ and plugging back T_1 , T_2 , and T_3 from (18), (19), and (20) into (17), yields:

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - 2(\eta_t - 4\eta_t^2 L)(\alpha_i^2 + 1 - \alpha_i) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \quad + \left(\frac{8\eta_t L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2 \right] \\
& \quad + \left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] + \alpha_i^2 \eta_t^2 \sigma^2 + (1 - \alpha_i)^2 \eta_t^2 \frac{\sigma^2}{n}, \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - 2(\eta_t - 4\eta_t^2 L)(\alpha_i^2 + 1 - \alpha_i) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \quad + \left(\frac{8\eta_t L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \left(\sigma^2 + (\zeta_i + \frac{\zeta}{n})\tau \right) \eta_{t-1}^2 \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \left(\sigma^2 + 2\frac{\zeta}{n}\tau \right) \eta_{t-1}^2 \\
& \quad + \underbrace{\left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right]}_{T_4} + \alpha_i^2 \eta_t^2 \sigma^2 + (1 - \alpha_i)^2 \eta_t^2 \frac{\sigma^2}{n}, \tag{21}
\end{aligned}$$

where using Lemma 2 we can bound T_4 as:

$$\begin{aligned}
T_4 & \leq \frac{6\eta_t}{\mu} (1 - \alpha_i)^2 \left(2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6\zeta_i + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + 6L^2 \Delta_i \right) \\
& \quad + 6\eta_t^2 (1 - \alpha_i)^2 \left(2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 6\zeta_i + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + 6L^2 \Delta_i \right). \tag{22}
\end{aligned}$$

Note that we choose $\alpha_i \geq \max\{1 - \frac{1}{4\sqrt{6}\kappa}, 1 - \frac{1}{4\sqrt{6}\kappa\sqrt{\mu}}\}$, hence $\frac{12L^2(1-\alpha_i)^2}{\mu} \leq \frac{\mu}{8}$ and $12L^2(1 - \alpha_i)^2 \leq \frac{\mu}{8}$, thereby we have:

$$T_4 \leq \frac{\mu\eta_t}{4} \|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 + 36\eta_t \left(\frac{1}{\mu} + \eta_t \right) (1 - \alpha_i)^2 \left(\zeta_i + L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + L^2 \Delta_i \right).$$

Now, using Lemma 4 we have:

$$\begin{aligned}
T_4 & \leq \frac{\mu\eta_t}{4} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2 \right] + 36\eta_t \left(\frac{1}{\mu} + \eta_t \right) (1 - \alpha_i)^2 \\
& \quad \left(\zeta_i + L^2 \left(\frac{a^3}{(t+a-1)^3} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] \right. \right. \\
& \quad \left. \left. + \left(t + 16 \left(\frac{1}{a+1} + \ln(t+a) \right) \right) \frac{1536\tau(\sigma^2 + 2\tau\frac{\zeta}{n})L^2}{\mu^4(t+a-1)^3} + \frac{128\sigma^2 t(t+2a)}{n\mu^2(t+a-1)^3} \right) + L^2 \Delta_i \right). \tag{23}
\end{aligned}$$

By plugging back T_4 from (23) in (21) and using the fact $-(\eta_t - 4\eta_t^2 L) \leq -\frac{1}{2}\eta_t$, and $(\alpha_i^2 + 1 - \alpha_i) \geq \frac{3}{4}$, we have:

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{\mu\eta_t}{8}\right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - \frac{3\eta_t}{4} \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) + \alpha_i^2 \eta_t^2 \sigma^2 + (1 - \alpha_i)^2 \eta_t^2 \frac{\sigma^2}{n} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \left(\sigma^2 + \left(\zeta_i + \frac{\zeta}{n}\right)\tau \right) \eta_{t-1}^2 \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \left(\sigma^2 + 2\frac{\zeta}{n}\tau \right) \eta_{t-1}^2 \\
& \quad + 36\eta_t \left(\frac{1}{\mu} + \eta_t \right) (1 - \alpha_i)^2 \left(\zeta_i + L^2 \left(\frac{a^3 \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right]}{(t - 1 + a)^3} \right. \right. \\
& \quad \left. \left. + \left(t + 16 \left(\frac{1}{a + 1} + \ln(t + a) \right) \right) \frac{1536\tau (\sigma^2 + 2\tau \frac{\zeta}{n}) L^2}{\mu^4 (t + a - 1)^3} + \frac{128\sigma^2 t(t + 2a)}{n\mu^2 (t - 1 + a)^3} + \Delta_i \right) \right).
\end{aligned}$$

Note that $(1 - \frac{\mu\eta_t}{8}) \frac{p_t}{\eta_t} \leq \frac{p_{t-1}}{\eta_{t-1}}$ where $p_t = (t + a)^2$, so, we multiply $\frac{p_t}{\eta_t}$ on both sides, and re-arrange the terms:

$$\begin{aligned}
& \frac{3p_t}{4} \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \leq \frac{p_{t-1}}{\eta_{t-1}} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - \frac{p_t}{\eta_t} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] + p_t \eta_t \left(\alpha_i^2 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
& \quad + \left(\frac{8L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} + \frac{6(1 - \alpha_i)^2 L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t L^2 \right) 3\tau \left(\sigma^2 + \left(\zeta_i + \frac{\zeta}{n}\right)\tau \right) p_t \eta_{t-1}^2 \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\frac{\zeta}{n}\tau \right) p_t \eta_{t-1}^2 + 36p_t \left(\frac{1}{\mu} + \eta_t \right) (1 - \alpha_i)^2 (\zeta_i + L^2 \Delta_i) \\
& \quad + 36p_t \left(\frac{1}{\mu} + \eta_t \right) (1 - \alpha_i)^2 \\
& \quad L^2 \left(\frac{a^3}{(t - 1 + a)^3} + (t + 16\Theta(\ln t)) \frac{1536\tau (\sigma^2 + 2\tau \frac{\zeta}{n}) L^2}{\mu^4 (t + a - 1)^3} + \frac{128\sigma^2 t(t + 2a)}{n\mu^2 (t - 1 + a)^3} \right).
\end{aligned}$$

By applying the telescoping sum and dividing both sides by $S_T = \sum_{t=1}^T p_t \geq T^3$ we have:

$$\begin{aligned}
& f_i(\hat{\mathbf{v}}_i) - f_i(\mathbf{v}_i^*) \\
& \leq \frac{1}{S_T} \sum_{t=1}^T p_t (f_i(\hat{\mathbf{v}}_i^{(t)}) - f_i(\mathbf{v}_i^*)) \\
& \leq \frac{4p_0 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(1)} - \mathbf{v}_i^*\|^2]}{3\eta_0 S_T} + \frac{1}{S_T} \frac{4}{3} \sum_{t=1}^T p_t \eta_t \left(\alpha_i^2 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
& + \frac{1}{S_T} \frac{4}{3} \sum_{t=1}^T \left(\frac{8L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} + \frac{6(1 - \alpha_i)^2 L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t L^2 \right) 3\tau \left(\sigma^2 + (\zeta_i + \frac{\zeta}{n})\tau \right) p_t \eta_{t-1}^2 \\
& + \frac{1}{S_T} \frac{4}{3} \sum_{t=1}^T \left(\frac{6(1 - \alpha_i)^2 L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t L^2 \right) 3\tau \left(\sigma^2 + 2\frac{\zeta}{n}\tau \right) p_t \eta_{t-1}^2 \\
& + 48(1 - \alpha_i)^2 \\
& \frac{L^2}{S_T} \sum_{t=1}^T p_t \left(\frac{1}{\mu} + \eta_t \right) \left(\frac{a^3}{(t-1+a)^3} + (t+16\Theta(\ln t)) \frac{1536\tau(\sigma^2 + 2\tau\frac{\zeta}{n})L^2}{\mu^4(t+a-1)^3} + \frac{128\sigma^2 t(t+2a)}{n\mu^2(t-1+a)^3} \right) \\
& + 48(1 - \alpha_i)^2 (\zeta_i + L^2 \Delta_i) \frac{1}{S_T} \sum_{t=1}^T p_t \left(\frac{1}{\mu} + \eta_t \right) \\
& \leq \frac{4p_0 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(1)} - \mathbf{v}_i^*\|^2]}{3\eta_0 S_T} + \frac{32T(T+a)}{3\mu S_T} \left(\alpha_i^2 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
& + \frac{4(1 - \alpha_i)^2}{3} \left(\frac{8L^2 T}{\mu(1 - 8(\alpha_i - \alpha_i^2))S_T} + \frac{6L^2 T}{\mu S_T} + \frac{10L^2 \Theta(\ln T)}{\mu S_T} \right) 3\tau \left(\sigma^2 + (\zeta_i + \frac{\zeta}{n})\tau \right) \frac{256a^2}{\mu^2(a-1)^2} \\
& + \frac{4}{3} \left(\frac{6(1 - \alpha_i)^2 L^2 T}{\mu S_T} + \frac{6(1 - \alpha_i)^2 L^2 \Theta(\ln T)}{\mu S_T} \right) 3\tau \left(\sigma^2 + 2\frac{\zeta}{n}\tau \right) \frac{256a^2}{\mu^2(a-1)^2} \\
& + 48(1 - \alpha_i)^2 L^2 \frac{a^2}{(a-1)^2 S_T} \\
& \left(\frac{a^3 \Theta(\ln T)}{\mu} + \left(\frac{T}{a} + \Theta(\ln T) \right) \frac{1536L^2 \tau (\sigma^2 + 2\tau\frac{\zeta}{n})}{\mu^5} + \frac{64(2a+1)\sigma^2 T(T+a)}{na\mu^3} \right) \\
& + 48(1 - \alpha_i)^2 L^2 \frac{a^2}{(a-1)^2 S_T} \left(\frac{16a^3 \pi^2}{6\mu} + (\Theta(\ln T)) \frac{1536L^2 \tau (\sigma^2 + 2\tau\frac{\zeta}{n})}{\mu^5} + \frac{2048(2a+1)\sigma^2 T}{na\mu^3} \right) \\
& + 48(1 - \alpha_i)^2 (\zeta_i + L^2 \Delta_i) \frac{1}{S_T} \left(\frac{S_T}{\mu} + \frac{8T(T+2a)}{\mu} \right) \\
& = O\left(\frac{\mu}{T^3}\right) + \alpha_i^2 O\left(\frac{\sigma^2}{\mu T}\right) + (1 - \alpha_i)^2 O\left(\frac{\zeta_i}{\mu} + \kappa L \Delta_i\right) \\
& + (1 - \alpha_i)^2 \left(O\left(\frac{\kappa L \ln T}{T^3}\right) + O\left(\frac{\kappa^2 \sigma^2}{\mu n T}\right) + O\left(\frac{\kappa^2 \tau^2 (\zeta_i + \frac{\zeta}{n}) + \kappa^2 \tau \sigma^2}{\mu T^2}\right) + O\left(\frac{\kappa^4 \tau (\sigma^2 + 2\tau\frac{\zeta}{n})}{\mu T^2}\right) \right).
\end{aligned}$$

where we use the convergence of $\sum_{t=1}^{\infty} \frac{\ln t}{t^2} \rightarrow O(1)$, and $\sum_{t=1}^{\infty} \frac{1}{t^2} \rightarrow \frac{\pi^2}{6}$.

□

E.2 PROOF OF CONVERGENCE OF APFL WITH SAMPLING

In this section we will provide the formal proof of the Theorem 2. Before proceed to the proof, we would like to give the convergence of global model here first. The following theorem establishes the convergence of global model in APFL.

Theorem 6 (Global model convergence of Local Descent APFL). *If each client's objective function is μ -strongly convex and L -smooth, and satisfies Assumption 1, using Algorithm 1, by choosing the learning rate $\eta_t = \frac{16}{\mu(t+a)}$, where $a = \max\{128\kappa, \tau\}$, $\kappa = \frac{L}{\mu}$, and using average scheme $\hat{\mathbf{w}} = \frac{1}{KS_T} \sum_{t=1}^T p_t \sum_{j \in U_t} \mathbf{w}_j^{(t)}$, where $p_t = (t+a)^2$, $S_T = \sum_{t=1}^T p_t$, and letting F^* to denote the*

minimum of the F , then the following convergence holds:

$$\mathbb{E}[F(\hat{\mathbf{w}})] - F^* \leq O\left(\frac{\mu}{T^3}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K})}{\mu T^2}\right) + O\left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K}) \ln T}{\mu T^3}\right) + O\left(\frac{\sigma^2}{KT}\right), \quad (24)$$

where τ is the number of local updates (i.e., synchronization gap).

Proof. The proof is provided in Appendix E.2.2. \square

Remark 4. It is noticeable that the obtained rate matches the convergence rate of the FedAvg, and if we choose $\tau = \sqrt{T/K}$, we recover the rate $O(\sqrt{1/KT})$, which is the convergence rate of well-known local SGD with periodic averaging (Woodworth et al., 2020a).

Now we switch to the proof of the Theorem 2. The proof pipeline is similar to what we did in Appendix E.1.3, non-sampling setting. The only difference is that we use sampling method here, hence, we will introduce the variance depending on sampling size K . Now we first begin with the proof of some technique lemmas.

E.2.1 PROOF OF USEFUL LEMMAS

Lemma 5. For Algorithm 1, at each iteration, the gap between local gradient and global gradient is bounded by

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) \right\|^2 \right] \\ & \leq 2L^2 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2] + 6 \left(2\zeta_i + 2\frac{\zeta}{K} \right) + 6L^2 \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2] + 6L^2 \Delta_i. \end{aligned}$$

Proof. From the smoothness assumption and by applying the Jensen's inequality we have:

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) \right\|^2 \right] \\ & \leq 2\mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla f_i(\mathbf{v}_i^*)\|^2] + 2\mathbb{E} \left[\left\| \nabla f_i(\mathbf{v}_i^*) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) \right\|^2 \right] \\ & \leq 2L^2 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2] + 6\mathbb{E} [\|\nabla f_i(\mathbf{v}_i^*) - \nabla f_i(\mathbf{w}^*)\|^2] \\ & \quad + 6\mathbb{E} \left[\left\| \nabla f_i(\mathbf{w}^*) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^*) \right\|^2 \right] + 6\mathbb{E} \left[\left\| \nabla \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^*) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) \right\|^2 \right] \\ & \leq 2L^2 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2] + 6L^2 \mathbb{E} [\|\mathbf{v}_i^* - \mathbf{w}^*\|^2] + 6 \left(2\zeta_i + 2\frac{1}{K} \sum_{j \in U_t} \zeta_j \right) + 6L^2 \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2] \\ & \leq 2L^2 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}^*\|^2] + 6L^2 \Delta_i + 6 \left(2\zeta_i + 2\frac{\zeta}{K} \right) + 6L^2 \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2]. \end{aligned}$$

\square

Lemma 6 (Local model deviation with sampling). For Algorithm 1, at each iteration, the deviation between each local version of the global model $\mathbf{w}_i^{(t)}$ and the global model $\mathbf{w}^{(t)}$ is bounded by:

$$\begin{aligned} \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2] & \leq 3\tau\sigma^2\eta_{t-1}^2 + 3(\zeta_i + \frac{\zeta}{K})\tau^2\eta_{t-1}^2, \\ \frac{1}{K} \sum_{i \in U_t} \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2] & \leq 3\tau\sigma^2\eta_{t-1}^2 + 6\tau^2\frac{\zeta}{K}\eta_{t-1}^2. \end{aligned}$$

where $\frac{\zeta}{K} = \frac{1}{K} \sum_{i=1}^n \zeta_i$.

Proof. According to Lemma 8 in [Woodworth et al. \(2020a\)](#):

$$\begin{aligned}
\mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] &\leq \frac{1}{K} \sum_{j \in U_t} \mathbb{E} \left[\|\mathbf{w}_j^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\
&\leq 3 \left(\sigma^2 + \zeta_i \tau + \frac{\zeta}{K} \tau \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} (1 - \mu \eta_q) \\
\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] &\leq \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E} \left[\|\mathbf{w}_j^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\
&\leq 3 \left(\sigma^2 + 2\tau \frac{\zeta}{K} \right) \sum_{p=t_c}^{t-1} \eta_p^2 \prod_{q=p+1}^{t-1} (1 - \mu \eta_q).
\end{aligned}$$

Then the rest of the proof follows Lemma 3. \square

Lemma 7. (Convergence of Global Model) Let $\mathbf{w}^{(t)} = \frac{1}{K} \sum_{j \in U_t} \mathbf{w}_j^{(t)}$. In Theorem 2's setting, using Algorithm 1 by choosing learning rate as $\eta_t = \frac{16}{\mu(t+a)}$, we have:

$$\begin{aligned}
\mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{a^3}{(T+a)^3} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] \\
&+ \left(T + 16 \left(\frac{1}{a+1} + \ln(T+a) \right) \right) \frac{1536a^2\tau(\sigma^2 + 2\tau\frac{\zeta}{K})L^2}{(a-1)^2\mu^4(T+a)^3} + \frac{128\sigma^2T(T+2a)}{K\mu^2(T+a)^3}.
\end{aligned}$$

Proof. First, we note that from the updating rule we have

$$\mathbf{w}^{(t+1)} - \mathbf{w}^* = \mathbf{w}^{(t)} - \mathbf{w}^* - \eta_t \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}; \xi_j^t). \quad (25)$$

Now, making both sides squared and according to the strong convexity we have:

$$\begin{aligned}
&\mathbb{E} \left[\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 \right] \\
&\leq \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] - 2\eta_t \mathbb{E} \left[\left\langle \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \right] \\
&\quad + \eta_t^2 \mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] + \eta_t^2 \frac{\sigma^2}{K} \\
&\leq (1 - \mu\eta_t) \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] - (2\eta_t - 2L\eta_t^2) \mathbb{E} \left[F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*) \right] + \eta_t^2 \frac{\sigma^2}{K} \\
&\quad + \eta_t^2 \frac{1}{K} \sum_{j \in U_t} L^2 \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] - 2\eta_t \mathbb{E} \left[\left\langle \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_j(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \right] \\
&\leq (1 - \mu\eta_t) \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] \underbrace{-(2\eta_t - 4L\eta_t^2)}_{\leq -\eta_t} \mathbb{E} \left[F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*) \right] + \eta_t^2 \frac{\sigma^2}{K} \\
&\quad + 2\eta_t^2 L^2 \frac{1}{K} \sum_{j \in U_t} \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + \frac{2\eta_t L^2}{\mu} \frac{1}{K} \sum_{j \in U_t} \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + \frac{\mu\eta_t}{2} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right]. \quad (26)
\end{aligned}$$

Then, merging the term, multiplying both sides with $\frac{p_t}{\eta_t}$, and do the telescoping sum yields:

$$\begin{aligned} \frac{p_T}{\eta_T} \mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{p_0}{\eta_0} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] - \mathbb{E}[F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*)] \\ &\quad + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) p_t \frac{1}{K} \sum_{j \in U_t} \mathbb{E} \left[\|\mathbf{w}_j^{(t)} - \mathbf{w}^{(t)}\|^2 \right] + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{K}. \end{aligned} \quad (27)$$

Plugging Lemma 6 into (27) yields:

$$\begin{aligned} \frac{p_T}{\eta_T} \mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{p_0}{\eta_0} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] - \mathbb{E}[F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*)] \\ &\quad + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3p_t \eta_{t-1}^2 \tau \left(\sigma^2 + 2\tau \frac{\zeta}{K} \right) + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{K}. \end{aligned} \quad (28)$$

Then, by re-arranging the terms will conclude the proof as

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{a^3}{(T+a)^3} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] \\ &\quad + \left(T + 16 \left(\frac{1}{a+1} + \ln(T+a) \right) \right) \frac{1536a^2 L^2 \tau \left(\sigma^2 + 2\tau \frac{\zeta}{K} \right)}{(a-1)^2 \mu^4 (T+a)^3} \\ &\quad + \frac{128\sigma^2 T(T+2a)}{K\mu^2 (T+a)^3}. \end{aligned}$$

□

E.2.2 PROOF OF THEOREM 6

Proof. According to (29) we have:

$$\begin{aligned} \frac{p_T}{\eta_T} \mathbb{E} \left[\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \right] &\leq \frac{p_0}{\eta_0} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] - \mathbb{E}[F(\mathbf{w}^{(t)}) - F(\mathbf{w}^*)] \\ &\quad + \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3p_t \eta_{t-1}^2 \tau \left(\sigma^2 + 2\tau \frac{\zeta}{K} \right) + \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{K}. \end{aligned} \quad (29)$$

By re-arranging the terms and dividing both sides by $S_T = \sum_{t=1}^T p_t > T^3$ yields:

$$\begin{aligned} \frac{1}{S_T} \sum_{t=1}^T p_t \left(\mathbb{E} [F(\mathbf{w}^{(t)})] - F(\mathbf{w}^*) \right) &\leq \frac{p_0}{S_T \eta_0} \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right] + \frac{1}{S_T} \sum_{t=1}^T \left(\frac{2L^2}{\mu} + 2\eta_t L^2 \right) 3p_t \eta_{t-1}^2 \tau \left(\sigma^2 + 2\tau \frac{\zeta}{K} \right) + \frac{1}{S_T} \sum_{t=1}^T p_t \eta_t \frac{\sigma^2}{K} \\ &\leq O \left(\frac{\mu \mathbb{E} \left[\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 \right]}{T^3} \right) + O \left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K})}{\mu T^2} \right) + O \left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K}) \ln T}{\mu T^3} \right) + O \left(\frac{\sigma^2}{KT} \right). \end{aligned}$$

Recalling that $\hat{\mathbf{w}} = \frac{1}{nS_T} \sum_{t=1}^T p_t \sum_{j=1}^n \mathbf{w}_j^{(t)}$, from the convexity of $F(\cdot)$, we can conclude that

$$\mathbb{E} [F(\hat{\mathbf{w}})] - F(\mathbf{w}^*) \leq O \left(\frac{\mu}{T^3} \right) + O \left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K})}{\mu T^2} \right) + O \left(\frac{\kappa^2 \tau (\sigma^2 + 2\tau \frac{\zeta}{K}) \ln T}{\mu T^3} \right) + O \left(\frac{\sigma^2}{KT} \right).$$

□

E.2.3 PROOF OF THEOREM 2

Now we provide the formal proof of Theorem 2. The main difference from without-sampling setting is that only a subset of local models get updated each period due to partial participation of devices, i.e., K out of all n devices that are sampled uniformly at random. To generalize the proof, we will use an indicator function to model this stochastic update, and show that while the stochastic gradient is unbiased, the variance is changed.

Proof. Recall that we defined virtual sequences of $\{\mathbf{w}^{(t)}\}_{t=1}^T$ where $\mathbf{w}^{(t)} = \frac{1}{K} \sum_{j \in U_t} \mathbf{w}_j^{(t)}$ and $\hat{\mathbf{v}}_i^{(t)} = \alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \mathbf{w}^{(t)}$. We also define an indicator variable to denote whether i th client was selected at iteration t :

$$\mathbb{I}_i^t = \begin{cases} 1 & \text{if } i \in U_t \\ 0 & \text{else} \end{cases}$$

obviously, $\mathbb{E}[\mathbb{I}_i^t] = \frac{K}{n}$. We start by writing the updating rule:

$$\hat{\mathbf{v}}_i^{(t+1)} = \hat{\mathbf{v}}_i^{(t)} - \alpha_i^2 \mathbb{I}_i^t \eta_t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}; \xi_i^t) - (1 - \alpha_i) \eta_t \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}; \xi_j^t).$$

Now, subtracting \mathbf{v}_i^* on both sides, taking the square of norm and expectation, yields:

$$\begin{aligned} & \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\ &= \mathbb{E} \left[\left\| \hat{\mathbf{v}}_i^{(t)} - \alpha_i^2 \mathbb{I}_i^t \eta_t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i) \eta_t \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \mathbf{v}_i^* \right\|^2 \right] \\ &+ \mathbb{E} \left[\left\| \alpha_i^2 \mathbb{I}_i^t \eta_t \left(\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}; \xi_i^t) \right) + (1 - \alpha_i) \eta_t \left(\frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}; \xi_j^t) \right) \right\|^2 \right] \\ &= \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - 2 \left\langle \frac{K}{n} \alpha_i^2 \eta_t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \eta_t \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \\ &+ \eta_t^2 \mathbb{E} \left[\left\| \alpha_i^2 \mathbb{I}_i^t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] + \alpha_i^2 \eta_t^2 \frac{2K^2 \sigma^2}{n^2} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K}. \\ &= \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - \underbrace{2\eta_t \left\langle \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \nabla f_i(\bar{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle}_{T_1} \\ &\quad - \underbrace{2\eta_t (1 - \alpha_i) \mathbb{E} \left[\left\langle \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right]}_{T_2} \\ &+ \underbrace{\eta_t^2 \mathbb{E} \left[\left\| \alpha_i^2 \mathbb{I}_i^t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right]}_{T_3} + \alpha_i^2 \eta_t^2 \frac{2K^2 \sigma^2}{n^2} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K}. \end{aligned}$$

Now we switch to bound T_1 :

$$\begin{aligned}
T_1 &= -2\eta_t \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \mathbb{E} \left[\left\langle \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right] \\
&\quad - 2\eta_t \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \mathbb{E} \left[\left\langle \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^* \right\rangle \right] \\
&\leq -2\eta_t \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) + \frac{\mu}{2} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] \right) \\
&\quad + \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \eta_t \left(\frac{8L^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \bar{\mathbf{v}}_i^{(t)}\|^2 \right] \right. \\
&\quad \quad \left. + \frac{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))}{8} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] \right) \\
&\leq -2\eta_t \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) + \frac{\mu}{2} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] \right) \\
&\quad + \eta_t \left(\frac{8L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \frac{K}{n} \alpha_i^2))} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] + \frac{\mu(1 - 8(\alpha_i - \frac{K}{n} \alpha_i^2))}{8} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] \right) \\
&\leq -2\eta_t \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) - \frac{7\mu\eta_t}{8} \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] \\
&\quad + \frac{8\eta_t L^2(1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2))} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right], \tag{30}
\end{aligned}$$

For T_2 , we use the same approach as we did in (19); To deal with T_3 , we also employ the similar technique in (20):

$$\begin{aligned}
T_3 &= \mathbb{E} \left[\left\| \alpha_i^2 \mathbb{I}_i^t \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\leq 2 \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right)^2 \mathbb{E} \left[\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)})\|^2 \right] + 2\mathbb{E} \left[\left\| (1 - \alpha_i) \left(\frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) \right) \right\|^2 \right] \\
&\leq 2 \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right)^2 \mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla f_i^*\|^2 \right] \\
&\quad + 2 \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right)^2 \mathbb{E} \left[\|\nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2 \right] \\
&\quad + 2(1 - \alpha_i)^2 \mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}_j^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \\
&\leq 8L \left(\frac{K}{n} \alpha_i^2 + 1 - \alpha_i \right) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i^* \right) + 4(1 - \alpha_i)^2 L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\
&\quad + 6(1 - \alpha_i)^2 \left(L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] + \mathbb{E} \left[\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla F(\mathbf{w}^{(t)})\|^2 \right] \right. \\
&\quad \quad \left. + \frac{1}{K} \sum_{j \in U_t} L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2 \right] \right). \tag{31}
\end{aligned}$$

Then plugging T_1, T_2, T_3 back, we obtain the similar formulation as the without sampling case in (17). Thus:

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - 2(\eta_t - 4\eta_t^2 L) \left(\alpha_i^2 \frac{K}{n} + 1 - \alpha_i \right) \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \quad + \alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2 \right] \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) \frac{1}{K} \sum_{j \in U_t} \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2 \right] \\
& \quad + \left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right]. \tag{32}
\end{aligned}$$

we then examine the lower bound of $\alpha_i^2 \frac{K}{n} + 1 - \alpha_i$. Notice that: $\alpha_i^2 \frac{K}{n} + 1 - \alpha_i = \frac{K}{n} ((\alpha_i - \frac{n}{2K})^2 + \frac{n}{K} - \frac{n^2}{4K^2})$.

Case 1: $\frac{n}{2K} \geq 1$ The lower bound is attained when $\alpha_i = 1$: $\alpha_i^2 \frac{K}{n} + 1 - \alpha_i \geq \frac{K}{n}$.

Case 2: $\frac{n}{2K} < 1$ The lower bound is attained when $\alpha_i = \frac{n}{2K}$: $\alpha_i^2 \frac{K}{n} + 1 - \alpha_i \geq 1 - \frac{n}{4K} > \frac{1}{2}$.

So $\alpha_i^2 \frac{K}{n} + 1 - \alpha_i \geq b := \min\{\frac{K}{n}, \frac{1}{2}\}$ always holds.

Now we plug it and Lemma 6 back to (32):

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - b\eta_t \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) + \alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau\eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau\eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + \left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right]. \tag{33}
\end{aligned}$$

Plugging Lemma 5 yields:

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - b\eta_t \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) + \alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau\eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau\eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + \left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \left[2L^2 \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] + 6 \left(2\zeta_i + 2\frac{\zeta}{K} \right) + 6L^2 \mathbb{E} \left[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right] + 6L^2 \Delta_i \right].
\end{aligned}$$

Then following the same procedure in Appendix E.1.3, together with the application of Lemma 7 we can conclude that:

$$\begin{aligned}
& f_i(\hat{\mathbf{v}}_i) - f_i(\mathbf{v}_i^*) \\
& \leq \frac{1}{S_T} \sum_{t=1}^T p_t (f_i(\hat{\mathbf{v}}_i^{(t)}) - f_i(\mathbf{v}_i^*)) \\
& \leq \frac{p_0 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(1)} - \mathbf{v}_i^*\|^2]}{b\eta_0 S_T} + \frac{1}{bS_T} \sum_{t=1}^T p_t \eta_t \left(\alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \right) \\
& \quad + \frac{1}{bS_T} \sum_{t=1}^T (1 - \alpha_i)^2 L^2 \left(\frac{8}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6}{\mu} + 10\eta_t \right) 3\tau p_t \eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \frac{1}{bS_T} \sum_{t=1}^T (1 - \alpha_i)^2 L^2 \left(\frac{6}{\mu} + 10\eta_t \right) 3\tau p_t \eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + 36(1 - \alpha_i)^2 \frac{L^2}{bS_T} \sum_{t=1}^T p_t \left(\frac{1}{\mu} + \eta_t \right) \left(\frac{a^3}{(t-1+a)^3} \mathbb{E} [\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2] \right. \\
& \quad \quad \left. + \left(t + 16 \left(\frac{1}{a+1} + \ln(t+a) \right) \right) \frac{1536a^2\tau(\sigma^2 + 2\tau\frac{\zeta}{K})L^2}{(a-1)^2\mu^4(t-1+a)^3} + \frac{128\sigma^2 t(t+2a)}{K\mu^2(t-1+a)^3} \right) \\
& \quad + 36(1 - \alpha_i)^2 \left(2\zeta_i + 2\frac{\zeta}{K} + L^2\Delta_i \right) \frac{1}{bS_T} \sum_{t=1}^T p_t \left(\frac{1}{\mu} + \eta_t \right). \\
& = O\left(\frac{\mu}{bT^3}\right) + \alpha_i^2 O\left(\frac{\sigma^2}{\mu bT}\right) + (1 - \alpha_i)^2 O\left(\frac{2\zeta_i + 2\frac{\zeta}{K}}{\mu b} + \frac{\kappa L\Delta_i}{b}\right) \\
& \quad + (1 - \alpha_i)^2 \left(O\left(\frac{\kappa L \ln T}{bT^3}\right) + O\left(\frac{\kappa^2 \sigma^2}{\mu bKT}\right) + O\left(\frac{\kappa^2 \tau^2 (\zeta_i + \frac{\zeta}{K}) + \kappa^2 \tau \sigma^2}{\mu bT^2}\right) \right. \\
& \quad \quad \left. + O\left(\frac{\kappa^4 \tau (\sigma^2 + 2\tau\frac{\zeta}{K})}{\mu bT^2}\right) \right).
\end{aligned}$$

□

F CONVERGENCE RATE WITHOUT ASSUMPTION ON α_i

In this section, we provide the convergence results of Algorithm 1 without assumption on α_i . The following Theorem establish the convergence rate:

Theorem 7 (Personalized model convergence of Local Descent APFL without assumption on α_i). *If each client's objective function is μ -strongly-convex and L -smooth, and its gradient is bounded by G , using Algorithm 1, learning rate: $\eta_t = \frac{8}{\mu(t+a)}$, where $a = \max\{64\kappa, \tau\}$, and using average scheme $\hat{\mathbf{v}}_i = \frac{1}{S_T} \sum_{t=1}^T p_t (\alpha_i \mathbf{v}_i^{(t)} + (1 - \alpha_i) \frac{1}{K} \sum_{j \in U_t} \mathbf{w}_j^{(t)})$, where $p_t = (t+a)^2$, $S_T = \sum_{t=1}^T p_t$, and f_i^* is the local minimum of the i th client, then the following convergence holds for all $i \in [n]$:*

$$\begin{aligned}
\mathbb{E}[f_i(\hat{\mathbf{v}}_i)] - f_i^* & \leq O\left(\frac{\mu}{bT^3}\right) + \alpha_i^2 O\left(\frac{\sigma^2}{\mu bT}\right) + (1 - \alpha_i)^2 O\left(\frac{G^2}{\mu b}\right) \\
& \quad + (1 - \alpha_i)^2 \left(O\left(\frac{\kappa L \ln T}{bT^3}\right) + O\left(\frac{\kappa^2 \sigma^2}{\mu bKT}\right) + O\left(\frac{\kappa^2 \tau^2 (\zeta_i + \frac{\zeta}{K}) + \kappa^2 \tau \sigma^2}{\mu bT^2}\right) \right. \\
& \quad \quad \left. + O\left(\frac{\kappa^4 \tau (\sigma^2 + 2\tau\frac{\zeta}{K})}{\mu bT^2}\right) \right), \tag{34}
\end{aligned}$$

where $b = \min\{\frac{K}{n}, \frac{1}{2}\}$

Remark 5. Here we remove the assumption $\alpha_i \geq \max\{1 - \frac{1}{4\sqrt{6}\kappa}, 1 - \frac{1}{4\sqrt{6}\kappa\sqrt{\mu}}\}$. The key difference is that we can only show the residual error with dependency on G , instead of more accurate quantities ζ_i and Δ_i . Apparently, when the diversity among data shards is small, ζ_i and Δ_i terms become small which leads to a tighter convergence rate. Also notice that, to realize the bounded gradient assumption, we need to require the parameters come from a bounded domain \mathcal{W} . Thus, we need to do projection during parameter update, which is inexpensive.

Proof. According to (33):

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{3\mu\eta_t}{8} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - b\eta_t \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \quad + \alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + \left(\frac{6\eta_t}{\mu} + 6\eta_t^2 \right) (1 - \alpha_i)^2 \mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right].
\end{aligned}$$

Here, we directly use the bound $\mathbb{E} \left[\left\| \frac{1}{K} \sum_{j \in U_t} \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \leq 2G^2$. Then we have:

$$\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{v}_i^*\|^2 \right] \\
& \leq \left(1 - \frac{\mu\eta_t}{4} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_i^{(t)} - \mathbf{v}_i^*\|^2 \right] - b\eta_t \left(\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - f_i(\mathbf{v}_i^*) \right) \\
& \quad + \alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \\
& \quad + \left(\frac{8\eta_t L^2 (1 - \alpha_i)^2}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 10(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \left(\frac{6(1 - \alpha_i)^2 \eta_t L^2}{\mu} + 6(1 - \alpha_i)^2 \eta_t^2 L^2 \right) 3\tau \eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + \left(\frac{12\eta_t}{\mu} + 12\eta_t^2 \right) (1 - \alpha_i)^2 G^2.
\end{aligned}$$

Then following the same procedure in Appendix E.1.3, we can conclude that:

$$\begin{aligned}
& f_i(\hat{\mathbf{v}}_i) - f_i(\mathbf{v}_i^*) \\
& \leq \frac{1}{S_T} \sum_{t=1}^T p_t (f_i(\hat{\mathbf{v}}_i^{(t)}) - f_i(\mathbf{v}_i^*)) \\
& \leq \frac{p_0 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(1)} - \mathbf{v}_i^*\|^2]}{b\eta_0 S_T} + \frac{1}{bS_T} \sum_{t=1}^T p_t \eta_t \left(\alpha_i^2 \eta_t^2 \frac{2K\sigma^2}{n} + (1 - \alpha_i)^2 \eta_t^2 \frac{2\sigma^2}{K} \right) \\
& \quad + \frac{1}{bS_T} \sum_{t=1}^T (1 - \alpha_i)^2 L^2 \left(\frac{8}{\mu(1 - 8(\alpha_i - \alpha_i^2 \frac{K}{n}))} + \frac{6}{\mu} + 10\eta_t \right) 3\tau p_t \eta_{t-1}^2 \left(\sigma^2 + (\zeta_i + \frac{\zeta}{K})\tau \right) \\
& \quad + \frac{1}{bS_T} \sum_{t=1}^T (1 - \alpha_i)^2 L^2 \left(\frac{6}{\mu} + 10\eta_t \right) 3\tau p_t \eta_{t-1}^2 \left(\sigma^2 + 2\frac{\zeta}{K}\tau \right) \\
& \quad + 12(1 - \alpha_i)^2 G^2 \frac{1}{bS_T} \sum_{t=1}^T p_t \left(\frac{1}{\mu} + \eta_t \right). \\
& = O\left(\frac{\mu}{bT^3}\right) + \alpha_i^2 O\left(\frac{\sigma^2}{\mu b T}\right) + (1 - \alpha_i)^2 O\left(\frac{G^2}{\mu b}\right) \\
& \quad + (1 - \alpha_i)^2 \left(O\left(\frac{\kappa L \ln T}{bT^3}\right) + O\left(\frac{\kappa^2 \sigma^2}{\mu b K T}\right) + O\left(\frac{\kappa^2 \tau^2 (\zeta_i + \frac{\zeta}{K}) + \kappa^2 \tau \sigma^2}{\mu b T^2}\right) + O\left(\frac{\kappa^4 \tau (\sigma^2 + 2\tau \frac{\zeta}{K})}{\mu b T^2}\right) \right).
\end{aligned}$$

□

G PROOF OF CONVERGENCE RATE IN NONCONVEX SETTING

In this section we will provide the proof of convergence results on nonconvex functions. Let us first present the convergence rate of the global model of APFL, on nonconvex function:

Theorem 8 (Global model convergence of Local Descent APFL). *If each client's objective function is L -smooth and satisfies Assumptions -I, using Algorithm 1, by choosing $K = n$ and learning rate $\eta = \frac{1}{2\sqrt{5}L\sqrt{T}}$, then the following convergence holds:*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] \leq O\left(\frac{L}{\sqrt{T}}\right) + O\left(\frac{\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n}}{T}\right) + O\left(\frac{\sigma^2}{\sqrt{T}}\right).$$

Proof. The proof is provided in Appendix G.2. □

As usual, let us introduce several useful lemmas before the formal proof of Theorem 3 and 8.

G.1 PROOF OF TECHNICAL LEMMAS

Lemma 8. *Under Theorem 3's assumptions, the following statement holds true:*

$$\begin{aligned}
\mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t+1)})] & \leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
& \quad + 2\alpha_i^4 (1 - \alpha_i)^2 \eta L^2 \mathbb{E} \left[\left\| \mathbf{w}_i^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}_j^{(t)} \right\|^2 \right] \\
& \quad + 4\eta (1 - \alpha_i^2)^2 \zeta_i + 8\eta L^2 (1 - \alpha_i^2)^2 \Gamma + 8(\alpha_i - \alpha_i^2)^2 \eta \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right].
\end{aligned}$$

Proof. According to the updating rule and smoothness of f_i , we have:

$$f_i(\hat{\mathbf{v}}_i^{(t+1)}) \leq f_i(\hat{\mathbf{v}}_i^{(t)}) + \left\langle \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t+1)} - \hat{\mathbf{v}}_i^{(t)} \right\rangle + \frac{L}{2} \left\| \hat{\mathbf{v}}_i^{(t+1)} - \hat{\mathbf{v}}_i^{(t)} \right\|^2.$$

Taking expectation on both sides yields:

$$\begin{aligned}
\mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t+1)})] &\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] + \mathbb{E} [\langle \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \hat{\mathbf{v}}_i^{(t+1)} - \hat{\mathbf{v}}_i^{(t)} \rangle] + \frac{L}{2} \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t+1)} - \hat{\mathbf{v}}_i^{(t)}\|^2] \\
&\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \eta \mathbb{E} \left[\left\langle \nabla f_i(\hat{\mathbf{v}}_i^{(t)}), \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\rangle \right] \\
&\quad + \frac{\eta^2 L}{2} \mathbb{E} \left[\left\| \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right).
\end{aligned}$$

Using the identity: $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2} \|\mathbf{a}\|^2 + \frac{1}{2} \|\mathbf{b}\|^2 - \frac{1}{2} \|\mathbf{a} - \mathbf{b}\|^2$ we have:

$$\begin{aligned}
\mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t+1)})] &\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] \\
&\quad - \underbrace{\left(\frac{\eta}{2} - \frac{\eta^2 L}{2} \right)}_{\geq 0} \mathbb{E} \left[\left\| \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) + (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\quad + \frac{\eta}{2} \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \alpha_i^2 \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i) \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\quad + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\quad + \eta \mathbb{E} \left[\left\| \alpha_i^2 \left(\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) \right) - (1 - \alpha_i) \nabla F(\mathbf{w}^{(t)}) + (1 - \alpha_i^2) \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \\
&\quad + (1 - \alpha_i)^2 \eta \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) - \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\quad + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2] + 2\eta \mathbb{E} \left[\left\| \alpha_i^2 \left(\nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - \nabla f_i(\bar{\mathbf{v}}_i^{(t)}) \right) \right\|^2 \right] \\
&\quad + 2\eta \mathbb{E} \left[\left\| (1 - \alpha_i^2) \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i^2) \nabla F(\hat{\mathbf{v}}_i^{(t)}) \right. \right. \\
&\quad \quad \left. \left. + (1 - \alpha_i^2) \nabla F(\hat{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i) \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right]
\end{aligned}$$

Using the smoothness of f and F , together with applying Jensen's inequality on the last term yields:

$$\begin{aligned}
\mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t+1)})] &\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\quad + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2] + 2\alpha_i^4 \eta L^2 \mathbb{E} [\|\hat{\mathbf{v}}_i^{(t)} - \bar{\mathbf{v}}_i^{(t)}\|^2] \\
&\quad + 4\eta \mathbb{E} [\|(1 - \alpha_i^2) \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i^2) \nabla F(\hat{\mathbf{v}}_i^{(t)})\|^2] \\
&\quad + 4\eta \mathbb{E} [\|(1 - \alpha_i^2) \nabla F(\hat{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i) \nabla F(\mathbf{w}^{(t)})\|^2] \\
&\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\quad + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2] + 2\alpha_i^4 (1 - \alpha_i)^2 \eta L^2 \mathbb{E} [\|\mathbf{w}_i^{(t)} - \mathbf{w}^{(t)}\|^2] \\
&\quad + 4\eta (1 - \alpha_i^2)^2 \zeta_i + 8\eta \mathbb{E} [\|(1 - \alpha_i^2) \nabla F(\hat{\mathbf{v}}_i^{(t)}) - (1 - \alpha_i^2) \nabla F(\mathbf{w}^{(t)})\|^2] \\
&\quad + 8\eta \mathbb{E} [\|(\alpha_i - \alpha_i^2) \nabla F(\mathbf{w}^{(t)})\|^2] \\
&\leq \mathbb{E} [f_i(\hat{\mathbf{v}}_i^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla f_i(\hat{\mathbf{v}}_i^{(t)})\|^2] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\
&\quad + 2\alpha_i^4 (1 - \alpha_i)^2 \eta L^2 \mathbb{E} [\|\mathbf{w}_i^{(t)} - \mathbf{w}^{(t)}\|^2] + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_j^{(t)}\|^2] \\
&\quad + 4\eta (1 - \alpha_i^2)^2 \zeta_i + 8\eta L^2 (1 - \alpha_i^2)^2 \Gamma + 8(\alpha_i - \alpha_i^2)^2 \eta \mathbb{E} [\|\nabla F(\mathbf{w}^{(t)})\|^2]
\end{aligned}$$

□

Lemma 9. Under Theorem 3's assumptions, the following statement holds true:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla F(\mathbf{w}^{(t)})\|^2] \leq \frac{2}{\eta T} \mathbb{E} [F(\mathbf{w}^{(1)})] + L^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}_j^{(t)} - \mathbf{w}^{(t)}\|^2] + \frac{\eta L \sigma^2}{n}.$$

Proof. According to the updating rule and smoothness of f_i , we have:

$$F(\mathbf{w}^{(t+1)}) \leq F(\mathbf{w}^{(t)}) + \langle \nabla F(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle + \frac{L}{2} \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2.$$

Taking expectation on both sides yields:

$$\begin{aligned}
\mathbb{E} [F(\mathbf{w}^{(t+1)})] &\leq \mathbb{E} [F(\mathbf{w}^{(t)})] + \mathbb{E} [\langle \nabla F(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle] + \frac{L}{2} \mathbb{E} [\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2] \\
&\leq \mathbb{E} [F(\mathbf{w}^{(t)})] - \eta \mathbb{E} \left[\left\langle \nabla F(\mathbf{w}^{(t)}), \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\rangle \right] \\
&\quad + \frac{\eta^2 L}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] + \frac{\eta^2 L \sigma^2}{2n}.
\end{aligned}$$

Using the identity $\langle \mathbf{a}, \mathbf{b} \rangle = -\frac{1}{2} \|\mathbf{a} - \mathbf{b}\|^2 + \frac{1}{2} \|\mathbf{a}\|^2 + \frac{1}{2} \|\mathbf{b}\|^2$, we have:

$$\begin{aligned}
\mathbb{E} [F(\mathbf{w}^{(t+1)})] &\leq \mathbb{E} [F(\mathbf{w}^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}^{(t)})\|^2] - \left(\frac{\eta}{2} - \frac{\eta^2 L}{2}\right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{w}_j^{(t)}) \right\|^2 \right] \\
&\quad + \frac{\eta^2 L \sigma^2}{2n} + \frac{\eta L^2}{2} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\mathbf{w}_j^{(t)} - \mathbf{w}^{(t)}\|^2] \\
&\leq \mathbb{E} [F(\mathbf{w}^{(t)})] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}^{(t)})\|^2] + \frac{\eta^2 L \sigma^2}{2n} + \frac{\eta L^2}{2} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\mathbf{w}_j^{(t)} - \mathbf{w}^{(t)}\|^2].
\end{aligned}$$

Re-arranging terms and doing the telescoping sum from $t = 1$ to T :

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla F(\mathbf{w}^{(t)})\|^2] \leq \frac{2}{\eta T} \mathbb{E} [F(\mathbf{w}^{(1)})] + L^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\|\mathbf{w}_j^{(t)} - \mathbf{w}^{(t)}\|^2] + \frac{\eta L \sigma^2}{n}.$$

□

Lemma 10. *Under Theorem 3's assumptions, the following statement holds true:*

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2] &\leq 10\tau^2 \eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n} \right), \\
\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2] &\leq 200L^2 \tau^4 \eta^4 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n} \right) + 20\tau^2 \eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \zeta_i \right).
\end{aligned}$$

Proof. For the first statement, we define $\gamma_t = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\mathbf{w}^{(t)} - \mathbf{w}_i^{(t)}\|^2]$, and let t_c be the latest synchronization stage. Then we have:

$$\begin{aligned}
\gamma_t &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \mathbf{w}^{t_c} - \sum_{j=t_c}^t \frac{\eta}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \left(\mathbf{w}^{t_c} - \sum_{j=t_c}^t \eta \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right) \right\|^2 \right] \\
&= \tau \sum_{j=t_c}^t \frac{\eta^2}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \frac{1}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right\|^2 \right] \\
&= \tau \sum_{j=t_c}^t \frac{\eta^2}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \frac{1}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \nabla f_k(\mathbf{w}_k^{(j)}) + \nabla f_k(\mathbf{w}_k^{(j)}) - \nabla f_k(\mathbf{w}^{(j)}) \right. \right. \\
&\quad \left. \left. + \nabla f_k(\mathbf{w}^{(j)}) - \nabla f_i(\mathbf{w}^{(j)}) + \nabla f_i(\mathbf{w}^{(j)}) - \nabla f_i(\mathbf{w}_i^{(j)}) + \nabla f_i(\mathbf{w}_i^{(j)}) - \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right\|^2 \right] \\
&\leq \tau \sum_{j=t_c}^{t+\tau} 5\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2 \gamma^j + \frac{\zeta}{n} \right).
\end{aligned}$$

Summing over t from t_c to $t_c + \tau$ yields:

$$\begin{aligned}
\sum_{t=t_c}^{t_c+\tau} \gamma_t &\leq \sum_{t=t_c}^{t_c+\tau} \sum_{j=t_c}^{t_c+\tau} 5\tau \eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + 2L^2 \gamma^j + \frac{\zeta}{n} \right) \\
&\leq 10L^2 \tau^2 \eta^2 \sum_{j=r\tau}^{(r+1)\tau} \gamma^j + 5\tau^3 \eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 5\tau^3 \eta^2 \frac{\zeta}{n}. \tag{35}
\end{aligned}$$

Since $\eta \leq \frac{1}{2\sqrt{5}\tau L}$, we have $10L^2\tau^2\eta^2 \leq \frac{1}{2}$, hence by re-arranging the terms we have:

$$\sum_{t=t_c}^{t_c+\tau} \gamma_t \leq 10\tau^3\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 10\tau^3\eta^2 \frac{\zeta}{n}.$$

Summing over all synchronization stages t_c , and dividing both sides by T can conclude the proof of the first statement:

$$\frac{1}{T} \sum_{t=1}^T \gamma_t \leq 10\tau^2\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 10\tau^2\eta^2 \frac{\zeta}{n}. \quad (36)$$

To prove the second statement, let $\delta_t^i = \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}_i^{(t)} \right\|^2 \right]$. Notice that:

$$\begin{aligned} \delta_t^i &= \mathbb{E} \left[\left\| \mathbf{w}^{t_c} - \sum_{j=t_c}^t \frac{\eta}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \left(\mathbf{w}^{t_c} - \sum_{j=t_c}^t \eta \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right) \right\|^2 \right] \\ &= \tau \sum_{j=t_c}^t \eta^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right\|^2 \right] \\ &= \tau \sum_{j=t_c}^t \eta^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{k=1}^n \nabla f_k(\mathbf{w}_k^{(j)}; \xi_k^j) - \nabla f_k(\mathbf{w}_k^{(j)}) + \nabla f_k(\mathbf{w}_k^{(j)}) - \nabla f_k(\mathbf{w}^{(j)}) \right. \right. \\ &\quad \left. \left. + \nabla f_k(\mathbf{w}^{(j)}) - \nabla f_i(\mathbf{w}^{(j)}) + \nabla f_i(\mathbf{w}^{(j)}) - \nabla f_i(\mathbf{w}_i^{(j)}) + \nabla f_i(\mathbf{w}_i^{(j)}) - \nabla f_i(\mathbf{w}_i^{(j)}; \xi_i^j) \right\|^2 \right] \\ &\leq \tau \sum_{j=t_c}^{t+\tau} 5\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + L^2\gamma_j + L^2\delta_j^i + \zeta_i \right). \end{aligned}$$

Summing over t from t_c to $t_c + \tau$ yields:

$$\begin{aligned} \sum_{t=t_c}^{t_c+\tau} \gamma_t &\leq \sum_{t=t_c}^{t_c+\tau} \sum_{j=t_c}^{t_c+\tau} 5\tau\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + L^2\gamma_j + L^2\delta_j^i + \zeta_i \right) \\ &\leq 5L^2\tau^2\eta^2 \sum_{j=t_c}^{t_c+\tau} \gamma_j + 5L^2\tau^2\eta^2 \sum_{j=t_c}^{t_c+\tau} \delta_j^i + 5\tau^3\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 5\tau^3\eta^2 \zeta_i. \end{aligned}$$

Since $\eta \leq \frac{1}{2\sqrt{5}\tau L}$, we have $5L^2\tau^2\eta^2 \leq \frac{1}{4}$, hence by re-arranging the terms we have:

$$\sum_{t=t_c}^{t_c+\tau} \delta_t^i \leq 20L^2\tau^2\eta^2 \sum_{j=t_c}^{t_c+\tau} \gamma_j + 20\tau^3\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 20\tau^3\eta^2 \zeta_i.$$

Summing over all synchronization stages t_c , and dividing both sides by T can conclude the proof of the first statement:

$$\frac{1}{T} \sum_{t=1}^T \delta_t^i \leq 20L^2\tau^2\eta^2 \frac{1}{T} \sum_{t=1}^T \gamma_t + 20\tau^2\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} \right) + 20\tau^2\eta^2 \zeta_i.$$

Using the result from (36) to bound $2\frac{1}{T} \sum_{t=1}^T \gamma_t$ we have:

$$\frac{1}{T} \sum_{t=1}^T \delta_t^i \leq 200L^2\tau^4\eta^4 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n} \right) + 20\tau^2\eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \zeta_i \right).$$

□

G.2 PROOF OF THEOREM 8

Proof. According to Lemma 9:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] \leq \frac{2}{\eta T} \mathbb{E} \left[F(\mathbf{w}^{(1)}) \right] + L^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + \frac{\eta L \sigma^2}{n}.$$

Then plugging in Lemma 10 will conclude the proof. \square

G.3 PROOF OF THEOREM 3

Proof. According to Lemma 8:

$$\begin{aligned} \mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t+1)}) \right] &\leq \mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(t)}) \right] - \frac{\eta}{2} \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] + \frac{\eta^2 L}{2} \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\ &\quad + 2\alpha_i^4 (1 - \alpha_i)^2 \eta L^2 \mathbb{E} \left[\left\| \mathbf{w}_i^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + (1 - \alpha_i)^2 \eta \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}^{(t)} - \mathbf{w}_j^{(t)} \right\|^2 \right] \\ &\quad + 4\eta (1 - \alpha_i^2)^2 \zeta_i + 8\eta (1 - \alpha_i^2)^2 \Gamma + 8(\alpha_i - \alpha_i^2)^2 \eta \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right]. \end{aligned}$$

Re-arranging the terms, summing from $t = 1$ to T , and dividing both sides with T yields:

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \\ &\leq \frac{2\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(1)}) \right]}{\eta T} + \eta L \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) \\ &\quad + 4\alpha_i^4 (1 - \alpha_i)^2 L^2 \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathbf{w}_i^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + 2(1 - \alpha_i)^2 L^2 \frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] \\ &\quad + 16(1 - \alpha_i)^2 L^2 \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(\mathbf{w}^{(t)}) \right\|^2 \right] + 8(1 - \alpha_i^2)^2 \zeta_i + 16(1 - \alpha_i^2)^2 \Gamma, \end{aligned}$$

Then, plug in Lemma 9 and 10 :

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \\ &\leq \frac{2\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(1)}) \right]}{\eta T} + \eta L \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) + 8(1 - \alpha_i^2)^2 \zeta_i + 16(1 - \alpha_i^2)^2 \Gamma \\ &\quad + 4\alpha_i^4 (1 - \alpha_i)^2 L^2 \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathbf{w}_i^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + 2(1 - \alpha_i)^2 L^2 \frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] \\ &\quad + 16(1 - \alpha_i)^2 L^2 \left(\frac{2}{\eta T} \mathbb{E} \left[F(\mathbf{w}^{(1)}) \right] + L^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{w}_j^{(t)} - \mathbf{w}^{(t)} \right\|^2 \right] + \frac{\eta L \sigma^2}{n} \right) \\ &\leq \frac{2\mathbb{E} \left[f_i(\hat{\mathbf{v}}_i^{(1)}) \right]}{\eta T} + \eta L \left(\alpha_i^4 \sigma^2 + (1 - \alpha_i)^2 \frac{\sigma^2}{n} \right) + 8(1 - \alpha_i^2)^2 \zeta_i + 16(1 - \alpha_i^2)^2 \Gamma \\ &\quad + 4\alpha_i^4 (1 - \alpha_i)^2 L^2 \left[200L^2 \tau^4 \eta^4 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n} \right) + 20\tau^2 \eta^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \zeta_i \right) \right] \\ &\quad + 180\tau^2 \eta^2 (1 - \alpha_i)^2 L^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n} \right) + 16(1 - \alpha_i)^2 L^2 \left(\frac{2}{\eta T} \mathbb{E} \left[F(\mathbf{w}^{(1)}) \right] + \frac{\eta L \sigma^2}{n} \right). \end{aligned}$$

Plugging in $\eta = \frac{1}{2\sqrt{5}\sqrt{T}L}$ concludes the proof:

$$\begin{aligned}
& \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla f_i(\hat{\mathbf{v}}_i^{(t)}) \right\|^2 \right] \\
& \leq O\left(\frac{L}{\sqrt{T}}\right) + \alpha_i^4 O\left(\frac{\sigma^2}{\sqrt{T}}\right) + (1 - \alpha_i)^2 O\left(\frac{\sigma^2}{n\sqrt{T}}\right) + (1 - \alpha_i)^2 \left(\frac{L}{\sqrt{T}} + \frac{\sigma^2}{n\sqrt{T}}\right) + (1 - \alpha_i^2)^2 (\zeta_i + \Gamma) \\
& + \alpha_i^4 (1 - \alpha_i)^2 O\left(\frac{\tau^4 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n}\right)}{T^2} + \frac{\tau^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \zeta_i\right)}{T}\right) + (1 - \alpha_i)^2 O\left(\frac{\tau^2 \left(\sigma^2 + \frac{\sigma^2}{n} + \frac{\zeta}{n}\right)}{T}\right).
\end{aligned}$$

□