Towards Flexible Device Participation in Federated Learning

Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, Carlee Joe-Wong
Carnegie Mellon University, Sun Yat-Sen University, Carnegie Mellon University

Abstract

Traditional federated learning algorithms impose strict requirements on the participation rates of devices, which limit the potential reach of federated learning. This paper extends the current learning paradigm to include devices that may become inactive, compute incomplete updates, and depart or arrive in the middle of training. We derive analytical results to illustrate how allowing more flexible device participation can affect the learning convergence when data is not independently and identically distributed (non-IID). We then propose a new federated aggregation scheme that converges even when devices may be inactive or return incomplete updates. We also study how the learning process can adapt to early departures or late arrivals, and analyze their impacts on the convergence.

1 Introduction

Federated learning is a cutting-edge learning framework that allows distributed devices to train a shared machine learning model cooperatively without sharing the raw data. In recent years, federated learning has exhibited remarkable performance in many applications such as next word suggestion, fault detection, and learning on private medical data (Li et al., 2020a). Generic federated learning involves a coordinator and a collection of devices. The training procedure consists of multiple rounds, each of which includes the following three steps: 1) Synchronization: the coordinator synchronizes the latest global model with all devices. 2) Local updates: each device trains a local model for a few local epochs, using samples from its local dataset. 3) Aggregation: the coordinator aggregates some, or all, of the local models to produce the next global model.

Considering that federated learning typically takes thousands of communication rounds to converge, it is difficult to ensure that all devices will be available during the entire training in practice. Moreover, there are typically multiple apps running simultaneously on user devices, competing for already highly constrained hardware resources. As such, it cannot be guaranteed that devices will complete their assigned training tasks in every training round as expected. A similar challenge also arises in cloud-based distributed learning due to the increasingly popular usage of preemptive cloud services, where the user process can be interrupted unexpectedly (McMahan et al., 2016).

While many methods have been proposed to mitigate the workload of individual devices, such as weight compression and federated dropout (Caldas et al., 2018; Konečný et al., 2016), they cannot completely remove the possibility that devices are unable to fulfill their training responsibilities, e.g., due to poor wireless
connectivity. Thus, in large scale federated learning, many resource-constrained devices have to be excluded from joining federated learning in the first place, which restricts the potential availability of training datasets, and weakens the applicability of federated learning. Furthermore, existing work does not specify how to react when confronting unexpected device behaviors, and also does not analyze the (negative) effects of such behaviors on the training progress.

In this paper, we relax these restrictions and allow devices to follow more flexible participation patterns. Specifically, the paper incorporates four situations that are not yet well discussed in the literature: 1) Incompleteness: devices might submit only partially completed work in a round. 2) Inactivity: furthermore, devices might not complete any updates, or respond to the coordinator at all. 3) Early departures: in the extreme case, existing devices might quit the training without finishing all training rounds. 4) Late arrivals: apart from existing devices, new devices might join after the training has already started.

The difference between inactivity and departure is that inactive devices will temporarily disconnect with the coordinator, but are expected to come back in the near future. In contrast, departing devices will inform the coordinator that they do not plan to rejoin the training. For example, if a user quits the app running federated learning, a message can be sent to the coordinator; the coordinator thus knows who is departing. In the meanwhile, although devices’ arriving and departing seem symmetric, they affect the model training differently, and thus require distinct treatments. The key difference is that arriving devices offer extra information about the data distribution, which can be utilized to accelerate the training, while departing devices reduce our available knowledge, thus degrading the applicability of the trained model.

Our approach to improve the flexibility of device participation comprises the following components that supplement the existing FedAvg algorithm and handle the challenges brought by flexible device participation.

- **Debiasing for partial model updates.** FedAvg aggregates device updates as a weighted sum, with weights that are proportional to the sizes of the local datasets. This choice of aggregation coefficients yields an unbiased gradient as in the centralized setting only when all data points from all devices are equally likely to join the learning (Li et al., 2020b). However, it in general fails to guarantee convergence to the globally optimal point in the presence of partial aggregation from incomplete and inactive devices. We show that by adapting the aggregation coefficients, the bias can be reduced and the convergence to a global optimum can still be established. Furthermore, our analysis shows the bias originates from the heterogeneity in device participation, as well as from the degree to which local datasets are not IID.

- **Fast-rebooting for device arrivals.** Arriving devices interrupt the training by forcing the model to re-orient to the new device’s data, thus slowing the convergence process. In this paper, we propose to rapidly reboot the training in the case of device arrivals by applying extra updates from the new devices. Intuitively, since an arriving device misses all previous epochs, the model training should emphasize more on its updates to compensate. We will rigorously prove this method indeed expedites learning convergence under certain conditions.

- **Redefining model applicability for device departures.** A model successfully trained by federated learning is expected to be applicable to the data from all participating devices. However, when a device withdraws itself from the learning, due to the lack of its future updates, we may no longer require the trained model to perform well on its data. It is then important to redefine the model’s applicability. Namely, one can either keep the departing device as a part of the global learning objective, or exclude it to focus only on the remaining devices. The decision depends on which definition yields smaller training loss. We will show the key to this determination lies in the remaining training time.

In Section 2 we review relevant literature. In Section 3 we give a convergence analysis that incorporates flexible device participation. Based on this analysis, we detail our contributions, as outlined above, in Section 4 and we experimentally verify our theoretical results in Section 5. Finally we conclude in Section 6.

2 Related Works

The celebrated federated learning algorithm named FedAvg runs the stochastic gradient descent (SGD) algorithm in parallel on each device in the system and periodically averages the updated parameters from a small set of end devices. However, its performance degrades when the local data is non-IID (Hsieh et al., 2019; Zhao et al., 2018). A few recent works provide theoretical results for the non-IID data case. For instance, Li et al. (2020b) analyze the convergence of FedAvg on non-IID data and establish an $O(1/T)$ convergence rate for strongly convex and smooth optimization problems, where $T$ is the number of rounds of local SGD updates. These works either simplify the heterogeneity of the devices, e.g., ignoring cases where some devices may partially finish some aggregation rounds or quit forever during the training (Li et al., 2020b), or consider alternative objective functions for the SGD.
algorithm to optimize [Li et al. 2018]. Alternatively, some recent papers propose to combine federated learning with the multi-task learning paradigm (Smith et al. 2017) (Corinzia and Buhmann 2019) where multiple models are trained simultaneously, but they also entail dedicated device participation throughout the training.

The FedAvg algorithm with non-IID data across devices has also been modified in specific edge computing scenarios to reduce the communication overhead (Liu et al. 2019) (Sattler et al. 2019) (Bonawitz et al. 2019) or maintain a good training convergence under a resource budget constraint (Wang et al. 2019). However, these works do not consider the possibility that the edge devices can be unavailable during the training process or join at different times, which are the main challenges of this work. An online learning framework (Chen et al. 2019) (Han et al. 2020) (Damaskinos et al. 2020) is a possible way to enable flexible device participation in the federated learning scenario. For instance, Chen et al. (2019) propose an asynchronous federated learning algorithm to handle unbalanced data that arrives in an online fashion onto different devices. Although the asynchronous aggregation in their proposed algorithm can be naturally applied to randomly inactive devices, the authors do not analyze how their algorithm’s convergence is affected by the device inactivity or incompleteness and the data heterogeneity.

In recent years, some attempts have been made to relax the strict training requirements on the participating devices. For example, Tu et al. (2020) study federated learning in a fog network topology with possible data sharing among devices, Yang et al. (2020) incorporate heterogeneity of devices into the design of the learning systems; Nishio and Yonenari (2019) propose a client selection policy that adapts to the change of devices’ hardware status. However, these works do not show how the variations in the devices could affect the convergence of training, nor do they incorporate the heterogeneity of user data into the algorithm design.

In (Rizk et al. 2020) and (Wang et al. 2020), the authors reveal that incomplete devices can block the convergence, but they consider neither other dynamic participation patterns such as inactivity, arrivals and departures, nor probabilistic models for uncertain device participation. To relieve the impact of incomplete devices, these works propose similar strategies as our paper by reweighting the contribution of local models. However, they focus mostly on removing the additional bias term originating from heterogeneous device updates, without looking into how this bias is related to the participation frequency of devices and the divergence among them. They also do not compare the proposed methods with alternative extensions of FedAvg. In this work, we model the device participation as random variables and incorporate them into the convergence analysis, and we compare the convergence rates for three reasonable aggregation schemes.

3 Convergence Analysis

In this section, we establish a convergence bound for federated learning with flexible device participation patterns. Our analysis generalizes the standard FedAvg to incorporate arbitrary aggregation coefficients. In the aggregation step, all devices are counted even if they cannot finish all local epochs. The analysis considers a non-IID data distribution and heterogeneous devices, i.e., some devices can be more stable than the others. We first derive the convergence bound with incomplete and inactive devices in Sections 3.1 to 3.2 and then discuss arrivals and departures in Section 3.3.

3.1 Algorithm Description

Suppose there are $N$ devices, where each device $k$ defines a local objective function $F_k(w)$. Here $w$ represents the parameters of the machine learning model to be optimized, and $F_k(w)$ may be defined as the average empirical loss over all data points at device $k$, as in typical federated learning frameworks (McMahan et al. 2016). The global objective is to minimize $F(w) = \sum_{k=1}^{N} p_k F_k(w)$, where $p_k = \frac{n_k}{n}$, $n_k$ is the number of data points device $k$ owns, and $n = \sum_{k=1}^{N} n_k$.

Let $w^*$ be the minimizer of $F$, and denote by $F_k^*$ the minimum value of $F_k$. We quantify the degree to which data at each device $k$ is distributed differently than that at other devices as $\Gamma_k = F_k(w^*) - F_k^*$, to capture that data distributions at different devices are non-IID, and let $\Gamma = \sum_{k=1}^{N} p_k \Gamma_k$ as in (Li et al. 2020).

We consider discrete time steps $t = 0, 1, \ldots$. Model weights are synchronized when $t$ is a multiple of $E$, i.e., each round consists of $E$ time steps. Assume there are at most $T$ rounds. For each round (say the $r$th round), the following three steps are executed:

- **Synchronization:** the coordinator broadcasts the latest global weight $w^*_E$ to all devices. Each device updates its local weight so that: $w^*_E = w_k^*$
- **Local updates:** each device runs stochastic gradient descent (SGD) on $F_k$ for $i = 0, \ldots, s_k - 1$,

\[
    w_{r,E+i+1}^k = w_{r,E+i}^k - \eta \hat{g}_{r,E+i}^k
\]

Here $\eta$ is a staircase learning rate that decays with $\tau$, $0 \leq s_k \leq E$ represents the number of local updates this device completes in this round, $\hat{g}_{r,E+i}^k = \nabla F_k(w_{r,E+i}^k, \xi_k)$ is the stochastic gradient at device $k$.

While some papers define local epochs and local updates separately, we use them interchangeably in this paper. Both refer to the times (1) is conducted in a global round.
k, and $\xi_k^t$ is a mini-batch sampled from device $k$’s local dataset. We also define $g^t_k = \nabla F_k(u^t_k)$ as the full batch gradient at device $k$, hence $g^t_k = \mathbb{E}[\xi^t_k | g^t_k]$.

**Aggregation**: the coordinator aggregates the gradients and generates the next global weight as

$$w^{(t+1)}_\tau = w^t_\tau + \sum_{i=1}^{N} p^t_i \left( w^{(t)}_\tau + s^t_i - w^t_\tau \right)$$

The variance of the stochastic gradients is bounded:

$$\mathbb{E} \left[ \left\| g^t_k - g^t_k \right\|^2 \right] \leq \sigma^2_k, \quad \forall k, t.$$  

**Assumption 3.4.** The expected squared norm of the stochastic gradients at each local device is uniformly bounded:

$$\mathbb{E} \left[ \left\| g^t_k \right\|^2 \right] \leq G^2$$

**Assumption 3.5.** There exists an upper bound $\theta > 0$ for the aggregation coefficient: $p^t_i/p^t_k \leq \theta, \forall k$. Assume the following expectations exist and do not vary with time: $\mathbb{E}[\xi_k^t], \mathbb{E}[p^t_k s^t_k], \mathbb{E}[(p^t_k)^2 s^t_k], \mathbb{E}[(\sum_{k=1}^{N} p^t_k - 2) + (\sum_{k=1}^{N} p^t_k s^t_k)]$ for all rounds $\tau$ and devices $k$, and assume $\mathbb{E}[\sum_{k=1}^{N} p^t_k s^t_k] \neq 0$. Intuitively, this last assumption ensures that some updates are aggregated in each round, otherwise this round can be simply omitted.

Generally, $p^t_k$’s are functions of $s^t_k$, and these expectations can be estimated from device histories.

**Theorem 3.1.** By choosing the learning rate $\eta_{\tau} = \frac{16E}{n\mathbb{E}[\sum_{k=1}^{N} p^t_k s^t_k]}/\tau + \gamma$, we can obtain

$$\mathbb{E}[\|w^{(t)}_\tau - w^*\|^2] \leq M_{\tau} D + V$$

Here we define $\gamma = \max \left\{ \frac{32E(1+\theta)L}{n\mathbb{E}[\sum_{k=1}^{N} p^t_k s^t_k]} \right\}$, $M_{\tau} = \sum_{t=0}^{\tau-1} \mathbb{E}[z_t]$, $D = \frac{64E\sum_{k=1}^{N} \mathbb{E}[p^t_k s^t_k] \Gamma_{k}}{n\mathbb{E}[\sum_{k=1}^{N} p^t_k s^t_k]}$, $V = \max \left\{ \gamma^2 \mathbb{E}[w^0 - w^*]^2 \left( \frac{16E}{n\mathbb{E}[\sum_{k=1}^{N} p^t_k s^t_k]} \right)^2 \mathbb{E}[z_{\tau}] \right\}$,

$$B_{\tau} = (2+\theta) L \left( \sum_{k=1}^{N} p^t_k s^t_k \right) + \left( 2 + \frac{n}{2(1+\theta)L} \right) (E - 1) G^2 \left( \sum_{k=1}^{N} p^t_k s^t_k \right)^2 + \left( \sum_{k=1}^{N} p^t_k s^t_k \right)^2 + \left( \sum_{k=1}^{N} (p^t_k)^2 s^t_k \right)^2 \theta_k^2.$$

Theorem 3.1 shows that the convergence rate is affected by the aggregation coefficients $p^t_k$’s as they determine $M_{\tau}, D$, and $V$. From $\mathbb{E}[w^{(t)}_\tau - w^*]$ eventually converges to a globally optimal solution only if $M_{\tau}$ increases sub-linearly with $\tau$. In the original full-participation FedAvg, $p^{t \tau}_k s^t_k \equiv p^t_k$, so $z_{\tau} \equiv 0$ and $M_{\tau} \equiv 0$ as per the definitions. Thus, full-participation FedAvg converges according to $\mathbb{E}[w^{(t)}_\tau - w^*]$.

3.2 General Convergence Bound

The analysis relies on the following five assumptions. The first four are standard (Li et al. 2020b). The last assumption ensures bounded aggregation coefficients and is satisfied by all schemes discussed in Section 3. In Section 4, we experimentally show that our proposed learning algorithm performs well even when some assumptions (like strong convexity) are violated.

**Assumption 3.1.** $F_1, \ldots, F_N$ are all $L$-smooth, so that $F$ is also $L$-smooth.

**Assumption 3.2.** $F_1, \ldots, F_N$ are all $\mu$-strongly convex, so that $F$ is also $\mu$-strongly convex.

**Assumption 3.3.** The variance of the stochastic gradients is bounded: $\mathbb{E}[\xi^t_k g^t_k - g^t_k]^2 \leq \sigma^2_k, \forall k, t$.

**Assumption 3.4.** The expected squared norm of the stochastic gradients at each local device is uniformly bounded: $\mathbb{E}[\|g^t_k\|^2] \leq G^2$ for all $k$ and $t$.

**Assumption 3.5.** There exists an upper bound $\theta > 0$ for the aggregation coefficient: $p^t_i/p^t_k \leq \theta, \forall k$.

While we only have results for $s^t_k$ whose distributions are static with time, Theorem 3.1 can be easily extended to time-varying distributed $s^t_k$ by replacing the corresponding expectations of $p^t_k s^t_k$ and $(p^t_k)^2 s^t_k$ with their minimum or maximum expectations over $\tau$.

3.3 Shifts in the Global Objective

Recall the global objective is $F(w) = \sum_{k \in C} F_k(w)$, i.e., an average of local objectives for participating devices $C$. A well trained model $w^*$ is expected to perform well on all data points generated by devices in $C$. In the presence of departing and arriving devices,
Consider the objective shift the convergence to the new objective can be bounded by and in the departure case rate back to \(E\) with \(E\). Here \(\tilde{\gamma}\) Combining Theorems 3.1 and 3.2, we can obtain the w of the training is then equivalent to starting over from device l. We formally define objective shift as the process of changing the global objective, and the applicability of the trained model, by adding or removing devices from \(C\).

The following theorem bounds the offset between the global optima due to the objective shift. As we can intuitively expect, the difference reduces when the data becomes more IID (\(\Gamma_l \rightarrow 0\)), and when the departing/arriving device owns fewer data points (\(n_l \rightarrow 0\):

**Theorem 3.2.** Suppose a device \(l\) arrives/departs, and let \(n\) be the total number of data points originally. Consider the objective shift \(F \rightarrow \tilde{F}\), \(w^* \rightarrow \tilde{w}^*\). Let \(\Gamma_k = F_k(\tilde{w}^*) - F_k^*\) quantify the degree of non-IID with respect to the new objective. Then in the arrival case

\[
\|w^* - \tilde{w}^*\| \leq \frac{2\sqrt{2L}}{\mu} \frac{n_l}{n + n_l} \sqrt{\Gamma_l}
\]

and in the departure case

\[
\|w^* - \tilde{w}^*\| \leq \frac{2\sqrt{2L}}{\mu} \frac{n_l}{n} \sqrt{\Gamma_l}
\]

Objective shift is mandatory when a new device (say device \(l\)) arrives: Unless \(F_l \equiv F\) (which is highly unlikely), incorporating updates from \(l\) will always move \(F(w)\) away from \(F^*\). The best strategy without objective shift is then not to aggregate updates from \(l\), and thus not to admit \(l\) into the learning process in the first place. In contrast, objective shift is optional when devices depart: we can keep the original objective \(F\) even if we will no longer receive updates from a departing device, if doing so yields smaller training loss.

Suppose an objective shift occurs at \(\tau_0\). The remainder of the training is then equivalent to starting over from \(w_{\tau_0}^\beta\) but converging towards the new objective \(\tilde{w}^*\). Combining Theorems 3.1 and 3.2 we can obtain the following convergence bound after the objective shifts:

**Corollary 3.2.1.** Assume the objective shifts at \(\tau_0\) with \(E\|w_{\tau_0}^\beta - w^*\|^2 \leq \Delta_{\tau_0}\). By increasing the learning rate back to \(\eta_\tau = \frac{16E}{\mu^2(\sum_{k=1}^{K} p_k + 1) (\tau - \tau_0) \beta + 1}\) for \(\tau > \tau_0\), the convergence to the new objective can be bounded by

\[
E\|w_{\tau}^\beta - \tilde{w}^*\|^2 \leq \frac{M_r D + \tilde{V}}{\tau - \tau_0} E + \tilde{\gamma}
\]

Here \(M_r, D, \tilde{V}, \tilde{\gamma}\) are defined analogously to \(M, D, V, \gamma\) but they respectively include/exclude the arriving/departing device. The first term in \(\tilde{V}\) equals \(\tilde{\gamma}^2 (\sqrt{\Delta_{\tau_0}} + \|w^* - \tilde{w}^*\|) = O\left(\frac{\sqrt{\Delta_{\tau_0}}}{\tau_0 + \tilde{\gamma}} + \Gamma_l\right)\).

The increase of the learning rate after the objective shift is necessary. Intuitively, if the shift happens at a large time \(\tau_0\) when \(w_{\tau_0}^\beta\) is close to the old optimal \(w^*\) and \(\eta_\tau\) is close to zero, the learning rate used in Theorem 3.1 will be too small to steer the model to the new optimum, since \(\|w_{\tau_0}^\beta - \tilde{w}^*\| \approx \|w^* - \tilde{w}^*\|\).

Comparing (3) and (6), an objective shift yields an one-time increase in the loss, which forces us to take actions when confronting departures and arrivals. In the case of device departure, it is possible that retaining the old objective can result in a smaller training loss compared to doing a shift. In this situation, the trained model is still applicable to data of the departing device. In the arrival case, though objective shift is mandatory, we can still accelerate the training by a “fast-reboot”, applying extra gradient updates from the arriving device.

We will discuss in Section 4.2 the fast-reboot method for the arrival case, and in Section 4.3 the decision of model applicability for the departure case.

### 4 Main Results

Based on the convergence analysis in Section 3, in this section, we present corollaries that can guide operators in reacting to flexible device participation.

#### 4.1 Debiasing on Incomplete Aggregation

According to Theorem 3.1, the convergence bound is controlled by the expectation of \(p^\beta_k\) and its functions. Below we discuss three plausible schemes of choosing \(p^\beta_k\) and compare their convergence rates in Table 1:

- **Scheme A:** Only aggregate parameters from devices that complete all \(E\) local epochs, with aggregation coefficient \(p^\beta_k = \frac{Np_k}{K\tau}\), where \(K\tau\) is the number of complete devices, \(q^k_e \in \{0, 1\}\) denotes if client \(k\) is complete. If \(K\tau = 0\), this round is discarded.

- **Scheme B:** Allow clients to upload incomplete work (with \(s^k_e < E\) updates), with fixed aggregation coefficient \(p^\beta_k = \frac{E}{s^k_e} p^\beta_k\), or \(p^\beta_k = 0\) if \(s^k_e = 0\).

Schemes A and B are natural extensions of FedAvg. Scheme C assigns a greater aggregation coefficient to devices that complete fewer local epochs. Though this idea seems counter-intuitive, as fewer local updates might lead to less optimal parameters (cf. Table 1), it turns out to be the only scheme that guarantees convergence when device participation is heterogeneous.

**Corollary 4.0.1.** Let \(K_e\) be the number of devices that run all \(E\) epochs, \(I_e\) indicate the appearance of any inactive devices in round \(\tau\), and write \(\sigma_K = \sum_k (p^k \sigma_k)^2\).
Table 1 gives the convergence rates of Schemes A, B, C when device updates may be incomplete and inactive.

Table 1: Convergence rates with incomplete and inactive devices. The bound for Scheme A assumes there is at least one complete device \((K_τ \neq 0)\), and those for Schemes B, C assume \(s^k_τ\) is not trivially zero \((\mathbb{E}[s^k_τ] \neq 0)\). While the three schemes have similar performance in the homogeneous setting, Schemes A and B fail to converge to the global optimum even assuming all devices are active. Scheme C works if inactive devices do not occur in every round \((\sum_t I_t < O(τ))\).

<table>
<thead>
<tr>
<th>Homogeneous</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (O\left(\frac{E[s^k_τ]+1}{τ}\right))</td>
<td>(\frac{D}{E})</td>
</tr>
<tr>
<td>B (O\left(\frac{s^k_τ+1}{τ}\right))</td>
<td>(\frac{D}{E})</td>
</tr>
<tr>
<td>C (O\left(\frac{1}{τ(\mathbb{E}[s^k_τ]+1)}\right))</td>
<td>(O\left(\sum_{k=0}^{N} l_i D + \sum_{k} (\frac{p^k}{τ})^2 \mathbb{E}\left[\frac{1}{τ}\right]+1\right))</td>
</tr>
</tbody>
</table>

The reason for enlarging the aggregation coefficients in Scheme C can be understood by observing from \(2\) that increasing \(p^k_τ\) is equivalent to increasing the learning rate of device \(k\). Thus, by assigning devices that run fewer epochs a greater aggregation coefficient, these devices effectively run further in each local step, compensating for the additional epochs other devices completed. As shown in Figure 1, Scheme C ensures an unbiased gradient after aggregation, while Schemes A and B will favor devices that run more epochs. Ideally, allowing devices to adapt learning rates by themselves would effectively lead to the same result. However, when a device is running local updates, it may not yet know or be able to estimate the number of local epochs it will complete. In contrast, centralized intervention can make accurate adjustments a posteriori.

Figure 1: Snapshot of one aggregation round. The bottom two devices completed all \(E = 5\) local epochs, while the top two completed only 3 and 4 epochs. Scheme C enlarges the incomplete gradients by respectively \(5/3\) and \(5/4\) and produces unbiased aggregation results. Aggregations with Schemes A and B are biased towards devices that run more epochs.

Table 1 also reveals how the following system and statistical factors affect the convergence asymptotically:

- The non-IID metric \(Γ\) is the major obstacle of convergence in the homogeneous case. In the heterogeneous setting, the \(D\) term (which grows with \(Γ\)) dominates the training loss. It controls the maximum non-diminishing loss \(D/E\) of Scheme A and B, and decelerates the training of Scheme C in the presence of inactive devices.
- Devices’ activeness \(s^k_τ\) and \(K_τ\) contribute inversely to the training loss: The more devices participate, the faster the loss decays. When inactivity occurs frequently, Scheme C cannot converge either. E.g., if a device never responds to the coordinator (so \(I_t = 1\)), its training loss can never converge to zero.
- The variance \(σ_N, σ_k\) in the stochastic gradient descent algorithm slows down the training as expected.

### 4.2 Fast-rebooting on Arrivals

Intuitively, when a device \(l\) arrives, \(\tilde{w}^*\) will be “dragged” towards its local optimum \(w^*_l\). The gradients from device \(l\) may thus encode more information about the new optimum \(\tilde{w}^*\) compared to those from the other devices. Thus, by adding an extra update \(-δ\nabla F_l(w^*)\), \(δ^l > 0\) to the gradient aggregation, it is likely that \(w\) can move closer to \(\tilde{w}^*\), allowing the training to fast-reboot from the point of arrivals. However, as shown in Figure 2, this intuition may not hold: it is also possible that \(-δ\nabla F_l(w^\tilde{w})\) ends up driving \(w^\tilde{w}\) away from \(\tilde{w}^*\). In fact, the success of this method is determined by the distance \(b = \|w^\tilde{w} - w^*\|\). When \(b\) is small, it is highly likely the extra update can rapidly reboot the training. We formalize this statement in Corollary 4.0.2.

Figure 2: Left: when the distance to the old optimum \(b = \|w^\tilde{w} - w^*\|\) is small, applying an extra update to \(w^\tilde{w}\) following the direction \(-\nabla F_l(w^\tilde{w})\) moves it closer to \(\tilde{w}^*\) \((d_2 < d_1)\). Right: for a large \(b\), the extra update may on the contrary enlarge this distance \((d_2 > d_1)\).

**Corollary 4.0.2.** Assume \(\nabla F(w)\) is continuous, and \(0 < \|\nabla F(w)\|_2, \|\nabla^2 F(w)\|_2 \leq W\) for any \(w\) (The latter is the induced \(l_2\) norm for matrices). Let \(w' = w - \delta\nabla F_l(w)\), then there exists a \(δ^l > 0\) such that \(\|w' - \tilde{w}^*\| < \|w - \tilde{w}^*\|\) if \(w\) satisfies

\[
\|w - w^*\| < \left(\frac{\tilde{F}(w^*) - \tilde{F}(\tilde{w}^*)}{(2\sqrt{2E} + 1) \sqrt{Γ_l}}\right) \frac{1}{b^l W} \tag{7}
\]
defines a sphere around the original global optimum \( w^* \) within which the extra update helps fast-reboot. The radius of the sphere depends on the divergence between the new (arriving) and old data points. Generally, the longer the training has elapsed, the closer the global model is to \( w^* \). Thus, the extra updating works best for devices that arrive late in the training.

When applied in practice, the extra updating can be conducted on-the-fly, by augmenting the aggregation coefficient of the arriving device so that \( p'_e = p' + \delta' \). Furthermore, the distance \( b \) can be estimated by the gradient norm with respect to the original objective.

As the name suggests, fast-reboot only accelerates the training for a certain duration after the device arrives. In fact, if there are no future interrupts, models with or without fast-rebooting eventually converge to the same global optimum. Nevertheless, fast-reboot is still beneficial if there is insufficient training time remaining (e.g., a device arrives near the end of the training).

### 4.3 Redefining Applicability on Departures

As is discussed in Section 3.3, when a device leaves, we need to redefine the applicability of the trained model. Namely, one can decide to either exclude this departing device and shift the objective, or keep including it and stick to the old objective. The decision depends on the time at which the device leaves. When including the device as a part of the global objective, from (6), since \( M_r = \tau - \tau_0 \) from then on, the training loss will always exceed a structural bias \( D/E \). In contrast, if the device is excluded and the model is trained with a shifted global objective, there will be an immediate increase in the convergence bound as in Theorem 3.2. But afterwards, the bound will decrease and eventually the parameters will converge to the new global optimum.

Assume a device leaves at \( \tau_0 < T \) and there are no subsequent arrivals/departures. Let \( f_0(\tau) \) be the convergence bound if we include the device, and \( f_1(\tau) \) be the bound if it is excluded. We can obtain

\[
f_0(\tau) = \frac{(\tau-\tau_0)D+V}{E+\gamma}, \quad f_1(\tau) = \frac{\tilde{V}}{(\tau-\tau_0)E+\gamma}.
\]

Here \( M_r, \tilde{V}, \gamma \) are defined analogously to \( M_r, V, \gamma \) but they exclude the departing device. A device is excluded if by doing so, a smaller training loss can be obtained at the deadline \( T \), which is summarized in the following corollary:

**Corollary 4.0.3.** Excluding a device that departs at \( \tau_0 \) leads to smaller training loss if

\[
\min_{\tau \geq \tau_0} f_0(\tau) \geq f_1(T) \tag{8}
\]

Further assume \( \tilde{\gamma} = \gamma \), and \( \tilde{V} \) is dominated by its first term so that \( \tilde{V} = \frac{\tilde{V}}{\tau_0 M_r + \gamma} + \Gamma_1 \). Then becomes

\[
T - \tau_0 \geq O\left(\sqrt{\Gamma_1 \tau_0}\right) \tag{9}
\]

From (9), when the remaining training time \( T - \tau_0 \) is at least \( O(\sqrt{T \tau_0}) \), applying the trained model to the departing device becomes less promising. It is thus better to exclude it and shift the objective. As we can expect, the bound grows with \( \Gamma_1 \), since the non-IID contribution from the departing device increases the initial \( \tilde{V} \). As \( \tau_0 \) increases, the learning rate without shift gets smaller, mitigating the increase of the training loss from departing devices.

### 5 Experiments

In this section, we experimentally evaluate Section 4’s results. Due to the limitations on hardware resources, the training process is performed in computer simulations. To ensure the simulation is consistent with the real learning environment, we use real-world traces to represent the participation patterns of simulated devices. We present our experiment setup in Section 5.1 and verify our theory results in Sections 5.2 - 5.4.

#### 5.1 Experiment Setup

We create various data traces to represent the heterogeneous participation patterns of local devices. We set up a simple federated learning experiment with five Raspberry PIs as workers, and a desktop server as the coordinator. Each PI has a training process that runs the original FedAvg algorithm, and a competitor process doing CPU-intensive work simultaneously. We manually tune the workload of the competitor process so that it takes up 0%, 30%, 50%, 70%, 90% of the PI’s CPU resources, simulating different device configurations in federated learning. Under the five settings, for each round, we record the percentage of required epochs the PI ends up submitting before a preset, fixed deadline. Due to the default load-balancing behavior of the operating system’s CPU scheduler, these traces do not contain zero epochs (i.e. inactive cases). To generate inactive device participation patterns, we create another set of three traces with respectively low, medium and high bandwidth. Devices can thus be inactive due to weak transmission. Table 2 shows the mean and standard deviation of the percentage of epochs completed for each trace. In the following experiments, each simulated device is randomly assigned a trace. For each aggregation round \( \tau \), it randomly samples from its trace to obtain the number of local epochs \( s_\tau \).

Three datasets are used in this paper: MNIST (LeCun et al., 1998), EMNIST (Cohen et al., 2017) and SYNTHETIC(\( \alpha, \beta \)) (Li et al., 2018). We build a two-layer MLP model and a two-convolution-layer CNN model respectively for MNIST and EMNIST, both models are defined by McMahan et al. (2016). For SYNTHETIC(\( \alpha, \beta \)), we use an ordinary logistic re-
We first examine the effects of the device heterogeneity on the convergence model. All models use the vanilla SGD as local optimizers, with batch sizes of 10 for MNIST and EMNIST, and 20 for SYNTHETIC. When generating non-IID data, we sort the MNIST and EMNIST data by labels so that each device is assigned data from one label chosen uniformly at random. For SYNTHETIC($\alpha, \beta$), we vary the parameters $\alpha, \beta$ from 0 to 1. The larger $\alpha, \beta$ are, the less IID the dataset becomes. We use the staircase learning rate $\eta_r = \eta_0/\tau$ as adopted in our convergence analysis. The initial $\eta_0$ is 2e-3 for MNIST, 5e-4 for EMNIST, and 1 for SYNTHETIC($\alpha, \beta$). Unless otherwise noted, the number of samples at each device follows the Type-I Pareto distribution with the Pareto index of 0.5.

### 5.2 Comparison of Aggregation Schemes

We first examine the effects of the device heterogeneity and the non-IID data distributions on the convergence for each aggregation scheme. We conduct eight sets of experiments where we incrementally increase the number of participation traces to reflect the increasing heterogeneity in device participation. For SYNTHETIC, we use $\alpha = \beta = 0$ for the IID case, and $\alpha = \beta = 1$ for the non-IID case. We train on 100 devices for MNIST, 62 devices for EMNIST (by merge), and 50 devices for SYNTHETIC($\alpha, \beta$). Table 2 records the differences in the test accuracies between different aggregation schemes after 200 global epochs. The typical convergence process is depicted in Figure 3.

![Figure 3: Test accuracy for non-IID EMNIST. Plots from left to right correspond to $|\mathcal{T}| = 1, 3, 5, 8$. (increasing device heterogeneity)](image)

As we can see, Scheme C yields the best test accuracy on average. Compared to Schemes A and B, it achieves higher accuracy when devices get more heterogeneous and less IID. This is consistent with our loss bounds in Table 1 since Schemes A and B fail to converge to the global optimum in the heterogeneous case with non-IID data. On the other hand, Scheme A performs extremely badly with large $|\mathcal{T}|$. This is because the last few traces contain very few complete rounds, significantly increasing $E[1/K_r]$. Noteworthily, Scheme C is no different from, or even worse than Scheme B in more homogeneous settings, this is consistent with Table 1 since $\frac{1}{E[^{|\mathcal{T}|]}]} \leq E[\frac{1}{K_r}]$. When the traces contain inactive devices ($|\mathcal{T}| \geq 6$), Scheme C becomes less stable due to the variance introduced by $I_t$ in Corollary 4.0.1.

Table 2: The means and standard deviations for the percentage of required local epochs actually submitted to the coordinator during the federated training. The first five traces do not contain inactive cases.

<table>
<thead>
<tr>
<th>Name</th>
<th>$T_0$</th>
<th>$T_{30}$</th>
<th>$T_{50}$</th>
<th>$T_{70}$</th>
<th>$T_{90}$</th>
<th>$T_{10}$</th>
<th>$T_{15}$</th>
<th>$T_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>100</td>
<td>75.3</td>
<td>67.2</td>
<td>57.2</td>
<td>56.3</td>
<td>82.5</td>
<td>74.1</td>
<td>51.2</td>
</tr>
<tr>
<td>Stdev</td>
<td>0</td>
<td>14.8</td>
<td>11.3</td>
<td>11.7</td>
<td>14.8</td>
<td>23.3</td>
<td>22.3</td>
<td>18.3</td>
</tr>
</tbody>
</table>

![Table 3: The % improvement in the test accuracies of Scheme C w.r.t. Scheme B (right numbers). Table 3 records the % improvement in the test accuracies of Scheme C w.r.t. Schemes A(left numbers) and Scheme B (right numbers). $|\mathcal{T}| = j$ represents using the first $j$ traces in Table 2.](image)

(a) MNIST Data

| $|\mathcal{T}|$ | 1 | 2 | 3 | 4 |
|---------------|---|---|---|---|
| IID           | -0.6 | 0.3 | 1.9 | 0.1 | 5.6 | 0.1 | 8.3 | 0.7 |
| NIIID         | 0.2 | -0.3 | 9.5 | 1.8 | 19.3 | 1.6 | 33.8 | 3.3 |

(b) EMNIST Data

| $|\mathcal{T}|$ | 1 | 2 | 3 | 4 |
|---------------|---|---|---|---|
| IID           | 0.7 | -0.6 | 0.9 | 0.1 | 4.2 | 0.7 | 4.8 | 1.0 |
| NIIID         | -0.1 | -0.7 | 17.0 | -2.0 | 34.2 | 1.8 | 37.9 | 4.8 |

(c) SYNTHETIC Data

| $|\mathcal{T}|$ | 1 | 2 | 3 | 4 |
|---------------|---|---|---|---|
| IID           | 6.9 | 1.1 | 6.6 | 1.2 | 4.0 | 1.5 | 7.6 | 1.2 |
| NIIID         | 30.2 | 2.5 | 22.5 | 3.0 | 25.3 | 2.2 | 18.6 | 1.8 |

### 5.3 Effectiveness of Fast-Reboot

We now investigate the effectiveness of the fast-reboot method described in Section 4.2. The experiments involve $N - 1$ existing devices, and the arriving device joins at $T_0$. As is discussed in Section 4.2, the method makes no difference when data distribution is IID. We thus only consider non-IID cases. We set $N = 10$ for MNIST and EMNIST (balanced) and $N = 30$ for
SYNTHETIC(1, 1). To avoid the interference brought by inactive devices, for this experiment we only use the first five traces in Table 2, and we adopt Scheme C as the aggregation method. All devices are given the same number of samples for fair comparison.

5.4 Model Applicability upon Departures

Next we study the situation when multiple devices arrive in a row. Figure 5 shows the training process for MNIST data. Every time a device arrives, we increase the learning rate as per Corollary 3.2.1. Initially, seven devices are in the training. After 100 global epochs, the remaining three devices arrive at 50 epoch intervals, without waiting for the model to fully converge. From Figure 5, the fast-reboot trick accelerates the convergence for every device arrival.

Table 5: The number of global epochs after the departing time $\tau_0$ until the test losses coincide for including and excluding options. The rows correspond to three choices of parameters $(\alpha, \beta)$ in SYNTHETIC$(\alpha, \beta)$.

<table>
<thead>
<tr>
<th>$\tau_0$</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1, 1)$</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>10</td>
<td>26</td>
<td>40</td>
</tr>
<tr>
<td>$(1, 0.5)$</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>14</td>
<td>13</td>
<td>7</td>
<td>12</td>
<td>36</td>
<td>34</td>
</tr>
<tr>
<td>$(1, 1.1)$</td>
<td>10</td>
<td>9</td>
<td>27</td>
<td>18</td>
<td>34</td>
<td>17</td>
<td>28</td>
<td>62</td>
<td>77</td>
</tr>
</tbody>
</table>

6 Conclusion and Future Work

This paper extends the federated learning paradigm to incorporate more flexible device participation. The analysis shows that incomplete local device updates can be utilized by scaling the corresponding aggregation coefficients, and a mild degree of device inactivity will not impact the convergence. Further investigation reveals how the convergence relates to heterogeneity in both the data and the device participation. The paper also proposes techniques to fast-reboot the training after new devices arrive, and provides an analytical criterion on when to exclude a departing device. In the future work, we will analyze groups of arrivals or departures, and investigate the possibility for users to dynamically update their datasets during the training.
Acknowledgements

This research was partially supported by the CMU CyLab IoT Initiative, and NSF CNS-1909306, CNS-1751075.

References


Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 2019.


A Proof of Theorems and Corollaries

A.1 Proof of Theorem 3.1

A.1.1 Equivalent View

For ease of the analysis, we introduce for each client $k$ and each global round $\tau$ a sequence of virtual variables $\alpha^k_{\tau E}, \alpha^k_{\tau E+1}, \ldots, \alpha^k_{(\tau+1)E-1}$. Here each $\alpha^k_t \in \{0, 1\}$ and $\sum_{i=0}^{E} \alpha^k_{\tau E+i} = s^k_{\tau}$. Since $s^k_{\tau}$ is a random variable, $\alpha^k_t$’s are also random variables, and the distributions of $\alpha^k_t$’s determine the distribution of $s^k_{\tau}$. For example, if $\alpha^k_{\tau E+1} \sim \text{Bernoulli}(p)$, then $s^k_{\tau} \sim \text{Bin}(E, p)$. In general, we do not make any assumption on the distributions and correlations of $\alpha^k_t$’s. Our results are thus valid for any realization of $s^k_{\tau}$.

With the definition of $\alpha^k_t$’s, we can rewrite (1)-(2) as:

$$w^k_{\tau E+i+1} = w^k_{\tau E+i} - \eta_t g^k_{\tau E+i} \alpha^k_{\tau E+i}$$  \hspace{1cm} (10)

$$w^G_{(\tau+1)E} = w^G_{\tau E} - \sum_{k=1}^{N} p^k_{\tau} \sum_{i=0}^{E} \eta_t g^k_{\tau E+i} \alpha^k_{\tau E+i}$$  \hspace{1cm} (11)

Note that $w^G_{\tau E}$ is visible only when $t$ is a multiple of $E$. To generalize it to arbitrary $t$, we define $\bar{w}_t$ such that $\bar{w}_0 = w^G_0$, and

$$\bar{w}_{\tau E+i+1} = \bar{w}_{\tau E+i} - \eta_t \sum_{k=1}^{N} p^k_{\tau} \alpha^k_{\tau E+i}$$  \hspace{1cm} (12)

Note that $\bar{w}_{\tau E+i} = \sum_{k=1}^{N} p^k_{\tau} w^k_{\tau E+i}$ only if $\sum_{k=1}^{N} p^k_{\tau} = 1$, which generally does not hold.

Lemma A.1. For any $\tau$, $\bar{w}_{\tau E} = w^G_{\tau E}$.

Proof. We will prove by induction. By definition, $\bar{w}_0 = w^G_0$. Suppose $\bar{w}_{\tau E} = w^G_{\tau E}$, then

$$\bar{w}_{(\tau+1)E} = \bar{w}_{(\tau+1)E-1} - \eta_t \sum_{k=1}^{N} p^k_{\tau} g^k_{(\tau+1)E-1} \alpha^k_{(\tau+1)E-1}$$

$$= \cdots = \bar{w}_{\tau E} - \sum_{i=0}^{E-1} \eta_t \sum_{k=1}^{N} p^k_{\tau} g^k_{\tau E+i} \alpha^k_{\tau E+i}$$  \hspace{1cm} (13)

$$= w^G_{\tau E} - \sum_{k=1}^{N} p^k_{\tau} \sum_{i=0}^{E-1} \eta_t g^k_{\tau E+i} \alpha^k_{\tau E+i} = w^G_{(\tau+1)E}$$

Thus, in the following analysis we will just use $\bar{w}_t$ to denote the global weight.

A.1.2 Key Lemmas

We first present a couple of important lemmas:

Lemma A.2.

$$\mathbb{E}_\xi \| \sum_{k=1}^{N} p^k \tilde{g}^k - g^k \|_2^2 \leq \sum_{k=1}^{N} (p^k)^2 \sigma^2_k$$  \hspace{1cm} (14)

Proof.

$$\| \sum_{k=1}^{N} p^k (g^k - \tilde{g}^k) \|_2^2 = \sum_{k=1}^{N} (p^k)^2 \| g^k - \tilde{g}^k \|_2^2 + \sum_{j \neq k} p^k p^j \langle g^k - \tilde{g}^k, g^j - \tilde{g}^j \rangle$$  \hspace{1cm} (15)
Thus,\[
\mathbb{E}_\xi \| \sum_{k=1}^{N} p_k^k (g^k_i - \bar{g}^k_i) \|^2 = \sum_{k=1}^{N} \mathbb{E}_\xi \| p_k^k (g^k_i - \bar{g}^k_i) \|^2 \leq \sum_{k=1}^{N} (p^k) \sigma^2_k
\] (17)

Lemma A.3. For \( i = 0, \ldots, E - 1 \) and all \( \tau, k \)
\[
\mathbb{E}_\xi \| \sum_{k=1}^{N} p_k^k (\bar{w}_{\tau E+i} - w^k_{\tau E+i}) \|^2 \leq (E - 1)G^2\eta^2 \left( \sum_{k=1}^{N} p_k^k s_k^2 + \left( \sum_{k=1}^{N} p_k^k - 2 \right) + \sum_{k=1}^{N} \frac{(p^k)^2}{p_k} s_k^2 \right)
\] (18)

Proof. Note that \( w^k_{\tau E} = \bar{w}_{\tau E} \) for all \( k \).
\[
\| \bar{w}_{\tau E+i} - \bar{w}_{\tau E+i} \|^2 = \| (\bar{w}_{\tau E+i} + \bar{w}_{\tau E}) - (w^k_{\tau E+i} + \bar{w}_{\tau E}) \|^2
\]
\[
= \| \bar{w}_{\tau E+i} - \bar{w}_{\tau E} \|^2 - 2\langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, w^k_{\tau E+i} - \bar{w}_{\tau E} \rangle + \| w^k_{\tau E+i} - \bar{w}_{\tau E} \|^2
\] (19)

From [10][12],
\[
\sum_{k=1}^{N} p_k^k w^k_{\tau E+i} = \sum_{k=1}^{N} p_k^k w^k_{\tau E+i-1} - \eta\tau \sum_{k=1}^{N} p_k^k g^k_{\tau E+i-1} \alpha^k_{\tau E+i-1}
\]
\[
= \sum_{k=1}^{N} p_k^k w^k_{\tau E+i-1} + \bar{w}_{\tau E+i-1} - \bar{w}_{\tau E+i-1}
\]
\[
= \ldots = \sum_{k=1}^{N} p_k^k w^k_{\tau E} + \bar{w}_{\tau E+i} - \bar{w}_{\tau E}
\] (20)

Thus,
\[
-2 \sum_{k=1}^{N} p_k^k \langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, w^k_{\tau E+i} - \bar{w}_{\tau E} \rangle
\]
\[
= -2 \langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, \sum_{k=1}^{N} p_k^k w^k_{\tau E} + \bar{w}_{\tau E+i} - \bar{w}_{\tau E} - \sum_{k=1}^{N} p_k^k \bar{w}_{\tau E} \rangle
\]
\[
= -2 \| \bar{w}_{\tau E+i} - \bar{w}_{\tau E} \|^2
\] (21)

\[
\sum_{k=1}^{N} p_k^k \| \bar{w}_{\tau E+i} - w^k_{\tau E+i} \|^2 = \left( \sum_{k=1}^{N} p_k^k - 2 \right) \| \bar{w}_{\tau E+i} - \bar{w}_{\tau E} \|^2 + \sum_{k=1}^{N} p_k^k \| w^k_{\tau E+i} - \bar{w}_{\tau E} \|^2
\] (22)

\[
\| \bar{w}_{\tau E+i} - \bar{w}_{\tau E} \|^2 = \| \sum_{j=0}^{i-1} \eta\tau \sum_{k=1}^{N} p_k^k g^k_{\tau E+j} \alpha^k_{\tau E+j} \|^2
\]
\[
= \sum_{j=0}^{i-1} \eta^2 \| \sum_{k=1}^{N} p_k^k \left( \sum_{j=0}^{i-1} g^k_{\tau E+j} \alpha^k_{\tau E+j} \right) \|^2
\]
\[
\leq \eta^2 \sum_{k=1}^{N} \frac{(p^k)^2}{p_k} \| \sum_{j=0}^{i-1} g^k_{\tau E+j} \alpha^k_{\tau E+j} \|^2
\] (23)
Here
\[ \| \sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 = \sum_{j=0}^{i-1} \| g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 + 2 \sum_{p<q} (g_{\tau E+p}^k \alpha_{\tau E+p}^k, g_{\tau E+q}^k \alpha_{\tau E+q}^k) \]
\[ \leq \sum_{j=0}^{i-1} \| g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 + 2 \sum_{p<q} \| g_{\tau E+p}^k \alpha_{\tau E+p}^k \| \| g_{\tau E+q}^k \alpha_{\tau E+q}^k \| \]
\[ \leq \sum_{j=0}^{i-1} \| g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 + \sum_{p<q} (\| g_{\tau E+p}^k \alpha_{\tau E+p}^k \|^2 + \| g_{\tau E+q}^k \alpha_{\tau E+q}^k \|^2) \]
\[ = i \sum_{j=0}^{i-1} \| g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 \]

So
\[ \mathbb{E}_\xi \| \sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 \leq iG^2 \sum_{j=0}^{i-1} \alpha_{\tau E+j}^k \leq (E-1)G^2 s_\tau^k \]
(25)

Plug (25) into (23) we have
\[ \mathbb{E}_\xi \| \bar{w}_{E+i} - \bar{w}_{E} \|^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^{N} \frac{(p_k^k)^2}{p^k} s_k^k \]
(26)

Similarly
\[ \mathbb{E}_\xi \sum_{k=1}^{N} p_k^k \| w_{\tau E+i}^k - \bar{w}_{\tau E} \|^2 = \mathbb{E}_\xi \sum_{k=1}^{N} p_k^k \| \eta_{\tau} \sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k \|^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^{N} p_k^k s_k^k \]
(27)

Plug (26) (27) into (22) we have
\[ \mathbb{E}_\xi \left( \sum_{k=1}^{N} p_k^k \| w_{\tau E+i}^k - \bar{w}_{\tau E+i}^k \|^2 \right) \leq (E-1)G^2 \eta_\tau^2 \left( \sum_{k=1}^{N} p_k^k s_k^k + \left( \sum_{k=1}^{N} p_k^k - 2 \right) \frac{\sum_{k=1}^{N} (p_k^k)^2}{p^k} s_k^k \right) \]
(28)

A.1.3 Bounding $\| \bar{w}_{\tau E+i+1} - w^* \|^2$

\[ \| \bar{w}_{\tau E+i+1} - w^* \|^2 = \| \bar{w}_{\tau E+i} - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k - w^* - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k + \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k \bar{g}_{\tau E+i}^k \|^2 \]
\[ = \| \bar{w}_{\tau E+i} - w^* - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k \|^2 + \eta_k^2 \| \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k (g_{\tau E+i}^k - \bar{g}_{\tau E+i}^k) \|^2 \]
\[ + 2\eta_k \langle \bar{w}_{\tau E+i} - w^* - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k, \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k (g_{\tau E+i}^k - \bar{g}_{\tau E+i}^k) \rangle \]
\[ \leq \| \bar{w}_{\tau E+i} - w^* - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k \|^2 + \eta_k^2 \| \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k (g_{\tau E+i}^k - \bar{g}_{\tau E+i}^k) \|^2 \]
\[ + 2\eta_k \langle \bar{w}_{\tau E+i} - w^* - \eta_k \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k g_{\tau E+i}^k, \sum_{k=1}^{N} p_k^k \alpha_{\tau E+i}^k (g_{\tau E+i}^k - \bar{g}_{\tau E+i}^k) \rangle \]
(29)
Since $E_\xi [g^k_{rE+i}] = g^k_{rE+i}$, we have $E_\xi [A_2] = 0$. We then bound $A_1$. 

$$A_1 = \|\bar{w}_{rE+i} - w^* - \eta_r \sum_{k=1}^N p^k_r \alpha^k_{rE+i} \bar{g}^k_{rE+i}\|^2$$

$$= \|\bar{w}_{rE+i} - w^*\|^2 - 2\eta_r \langle \bar{w}_{rE+i} - w^*, \sum_{k=1}^N p^k_r \alpha^k_{rE+i} \bar{g}^k_{rE+i} \rangle + \eta^2_r \sum_{k=1}^N p^k_r \alpha^k_{rE+i} \|\bar{g}^k_{rE+i}\|^2$$

(30)

Since $F_k$ is $L$-smooth,

$$\|\alpha^k_{rE+i} \bar{g}^k_{rE+i}\|^2 \leq 2L(F_k(w^k_{rE+i}) - F^*_k)\alpha^k_{rE+i}$$

(31)

By the convexity of $l_2$ norm

$$B_2 = \eta^2_r \sum_{k=1}^N (p^k_r)^2 \|\alpha^k_{rE+i} \bar{g}^k_{rE+i}\|^2 = \eta^2_r \sum_{k=1}^N p^k_r \|\alpha^k_{rE+i} \bar{g}^k_{rE+i}\|^2 \leq 2L\eta^2_r \sum_{k=1}^N p^k_r (F_k(w^k_{rE+i}) - F^*_k)\alpha^k_{rE+i}$$

(32)

Here

$$-2\langle \bar{w}_{rE+i} - w^k_{rE+i}, \alpha^k_{rE+i} \bar{g}^k_{rE+i} \rangle \leq 2\|\bar{w}_{rE+i} - w^k_{rE+i}, \alpha^k_{rE+i} \bar{g}^k_{rE+i} \|$$

$$\leq 2\alpha^k_{rE+i} \|\bar{w}_{rE+i} - w^k_{rE+i} \| \|\bar{g}^k_{rE+i}\| \leq \left( \frac{1}{\eta_r} \|\bar{w}_{rE+i} - w^k_{rE+i}\|^2 + \eta_r \|\bar{g}^k_{rE+i}\|^2 \right)\alpha^k_{rE+i}$$

(34)

Since $F_k$ is $\mu$-strongly convex

$$\langle w^k_{rE+i} - w^*, \alpha^k_{rE+i} \bar{g}^k_{rE+i} \rangle \geq \left( (F_k(w^k_{rE+i}) - F_k(w^*)) + \mu \|w^k_{rE+i} - w^*\|^2 \right)\alpha^k_{rE+i}$$

(35)

Plug (34), (36) to (33)

$$B_1 \leq \sum_{k=1}^N p^k_r \alpha^k_{rE+i} \left( \|\bar{w}_{rE+i} - w^k_{rE+i}\|^2 + \eta^2_r \|\bar{g}^k_{rE+i}\|^2 - 2\eta_r ((F_k(w^k_{rE+i}) - F_k(w^*)) + \frac{\mu}{2} \|w^k_{rE+i} - w^*\|^2) \right)$$

(36)
Plug (32) (36) to (30)

\[
A_1 \leq \|\bar{w}_{\tau E+i} - w^*\|^2 + 2L\theta \eta_r^2 \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k^*) \\
+ \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k \left( \|w_{\tau E+i} - w_{\tau E+i}^k\|^2 + \eta_r^2 \|\bar{y}_{\tau E+i}^k\|^2 \right) - 2\eta_r \left((F_k(w_r^{k*}) - F_k(w^*)) + \frac{H}{2} \|w_{\tau E+i}^k - w^*\|^2 \right) \\
\leq \|\bar{w}_{\tau E+i} - w^*\|^2 - \mu \eta_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k \|w_{\tau E+i}^k - w_{\tau E+i}^k\|^2 + \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k \|\bar{w}_{\tau E+i} - w_{\tau E+i}^k\|^2 \\
+ 2(1 + \theta) L \eta_r^2 \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k^*) - 2\eta_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k(w^*))
\]

(37)

\[
\|w_{\tau E+i}^k - w^*\|^2 = \|w_{\tau E+i}^k - \bar{w}_{\tau E+i} + \bar{w}_{\tau E+i} - w^*\|^2 \\
= \|w_{\tau E+i}^k - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 + 2(w_{\tau E+i}^k - \bar{w}_{\tau E+i}, \bar{w}_{\tau E+i} - w^*) \\
\geq \|w_{\tau E+i}^k - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 - 2\|w_{\tau E+i}^k - \bar{w}_{\tau E+i}\| \|\bar{w}_{\tau E+i} - w^*\| \\
\geq \|w_{\tau E+i}^k - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 - (2\|w_{\tau E+i}^k - \bar{w}_{\tau E+i}\|^2 + \frac{1}{2} \|\bar{w}_{\tau E+i} - w^*\|^2)
\]

(38)

Thus,

\[
A_1 \leq (1 - \frac{1}{2} \mu \eta_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k) \|\bar{w}_{\tau E+i} - w^*\|^2 + (1 + \mu \eta_r) \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k \|w_{\tau E+i}^k - w_{\tau E+i}^k\|^2 + C
\]

(39)

Let \( \gamma_r = 2\eta_r (1 - (1 + \theta) L \eta_r) \). Assume \( \eta_r \leq \frac{1}{2(1 + \theta) L} \), hence \( \eta_r \leq \gamma_r \leq 2\eta_r \).

\[
C = -2\eta_r (1 - (1 + \theta) L \eta_r) \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k^*) + 2\eta_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w^*) - F_k) \\
= -\gamma_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k^*) + F_k(w^*) - F_k^* + 2\eta_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w^*) - F_k^*) \\
= -\gamma_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k^*) + (2\eta_r - \gamma_r) \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w^*) - F_k^*) \]

(40)

\[
\leq -\gamma_r \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k (F_k(w_r^k) - F_k(w^*) + 2(1 + \theta) L \eta_r^2 \sum_{k=1}^N p_r^k \alpha_{\tau E+i}^k F_k
\]

(41)
Next we bound $D$

$$
\sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F_k(w^*)) \geq \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\nabla F_k(\bar{w}_{\tau E+i}), w^k_{\tau E+i} - \bar{w}_{\tau E+i}) + \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$

$$
\geq \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\nabla F_k(\bar{w}_{\tau E+i}), w^k_{\tau E+i} - \bar{w}_{\tau E+i}) + \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$

$$
\geq \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\nabla F_k(\bar{w}_{\tau E+i})) \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\| + \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$

$$
\geq \frac{1}{2} \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\eta_r \|\nabla F_k(\bar{w}_{\tau E+i})\|^2 + \frac{1}{\eta_r} \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2) + \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$

$$
\sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\eta_r L(F_k(\bar{w}_{\tau E+i}) - F_k^*) + \frac{1}{2\eta_r} \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2) + \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$

Thus,

$$
C \leq \gamma_r \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(\eta_r L (F_k(\bar{w}_{\tau E+i}) - F_k^*) + \frac{1}{2\eta_r} \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2) - \gamma_r \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) + 2(1 + \theta) \eta_r^2 \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \Gamma_k
$$

$$
\leq \gamma_r (\eta_r L - 1) \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) + \frac{\gamma_r}{2\eta_r} \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2
$$

$$
+ 2(1 + \theta) \eta_r^2 \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \Gamma_k
$$

Plug (43) to (39) we have

$$
A_1 \leq \|\bar{w}_{\tau E+i} - w^*\|^2 + \mu \eta_r \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \|w^k_{\tau E+i} - w^*\|^2 + 2 \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2
$$

$$
+ 2(1 + \theta) \eta_r^2 \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i} \Gamma_k + \gamma_r (\eta_r L - 1) \sum_{k=1}^{N} p^k \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
$$
Plug [43] to [29],

\[
\|\bar{w}_{\tau E+i+1} - w^*\|^2 \leq \left(1 - \frac{1}{2} \mu_\tau \sum_{k=1}^{N} p^k \alpha^k \right) \|\bar{w}_{\tau E+i} - w^*\|^2
\]

\[
+ \eta_\tau^2 \| \sum_{k=1}^{N} p^k \alpha^k \| \bar{g}_{\tau E+i} - g_{\tau E+i} \|^2 + \left(2 + \mu_\tau \right) \sum_{k=1}^{N} p^k \alpha^k \| \bar{w}_{\tau E+i} - w^*_{\tau E+i} \|^2
\]

\[
\leq 2 + (2 + \theta) \sum_{k=1}^{N} p^k \alpha^k \| \bar{g}_{\tau E+i} - g_{\tau E+i} \|^2 + \gamma_\tau \| \bar{w}_{\tau E+i} - w^*_{\tau E+i} \|^2
\]

(44)

Thus,

\[
\|\bar{w}_{\tau E+i+1} - w^*\|^2 \leq \left(1 - \frac{1}{2} \mu_\tau \sum_{k=1}^{N} p^k \alpha^k \right) \|\bar{w}_{\tau E+i} - w^*\|^2 + \eta_\tau^2 B_{\tau E+i}
\]

\[
+ 2 \eta_\tau \sum_{k=1}^{N} p^k \alpha^k (F_k(w^*) - F_k(\bar{w}_{\tau E+i}))
\]

(46)

Apply the lemmas we have

\[
\mathbb{E}_\xi [C_{\tau E+i}] \leq \sum_{k=1}^{N} \left( p^k \right)^2 \alpha^k \| \bar{w}_{\tau E+i} - w^*\|^2 + 2(2 + \theta) L \sum_{k=1}^{N} p^k \alpha^k \| \bar{w}_{\tau E+i} - w^*\|^2
\]

\[
+ \left(2 + \frac{\mu}{2(1+\theta)L}\right) (E-1) G \left( \sum_{k=1}^{N} p^k s_k + \sum_{k=1}^{N} p^k \| \bar{g}_{\tau E+i} - g_{\tau E+i} \|^2 \right)
\]

(47)

For convenience we write \( \Delta_{\tau E+i} = \| \bar{w}_{\tau E+i} - w^*\|^2 \), and \( \bar{\Delta}_{\tau E+i} = \mathbb{E}[\Delta_{\tau E+i}] \), where the expectation is taken over all random variables up to \( \tau E+i \).

### A.1.4 Bounding \( \| \bar{w}_{\tau E} - w^* \| \)

Summing from \( \tau E \) to \((\tau + 1)E\) we have

\[
\sum_{i=0}^{E-1} \Delta_{\tau E+i} \leq \sum_{i=0}^{E-1} \left(1 - \frac{1}{2} \mu_\tau \sum_{k=1}^{N} p^k \alpha^k \right) \Delta_{\tau E+i} + \eta^2 C_{\tau} + 2 \eta_\tau \sum_{k=1}^{N} p^k s_k (F_k(w^*) - F_k(\bar{w}_{\tau E+i}))
\]

(48)

where \( C_{\tau} = \sum_{i=0}^{E-1} C_{\tau E+i} \), and \( \bar{w}_{\tau E+i} = \arg\min_{\bar{w}_{\tau E+i}} \sum_{k=1}^{N} p^k \alpha^k F_k(\bar{w}_{\tau E+i}) \).

Reorganize it we can get

\[
\sum_{i=0}^{E-1} \Delta_{(\tau + 1)E} \leq \Delta_{\tau E} - \frac{1}{2} \mu_\tau \sum_{k=1}^{N} p^k \alpha^k \Delta_{\tau E+i} + \eta^2 C_{\tau} + 2 \eta_\tau \sum_{k=1}^{N} p^k s_k (F_k(w^*) - F_k(\bar{w}_{\tau E+i}))
\]

(49)
We then seek to find a lower bound for $\Delta_{E+i}$.

$$
\sqrt{\Delta_{E+i+1}} = \|\bar{w}_{E+i+1} - w^*\| = \|\bar{w}_{E+i+1} - \bar{w}_{E+i} + \bar{w}_{E+i} - w^*\|
$$

$$
\leq \|\bar{w}_{E+i+1} - \bar{w}_{E+i}\| + \sqrt{\Delta_{E+i}}
$$

$$
= \|\eta_T \sum_{k=1}^{N} p_k^k \alpha_{E+i}^k \xi_{E+i}\| + \sqrt{\Delta_{E+i}}
$$

(50)

Define $h_{E+i} = \|\sum_{k=1}^{N} p_k^k \alpha_{E+i}^k \xi_{E+i}\|$.

Thus,

$$
\sqrt{\Delta_{(\tau+1)E}} \leq \sqrt{\Delta_{(\tau+1)E-1} + \eta_T h_{(\tau+1)E-1}}
$$

$$
\leq \cdots \leq \sqrt{\Delta_{E+i} + \sum_{j=i}^{E-1} \eta_T h_{E+j}}
$$

(51)

$$
\Delta_{(\tau+1)E} \leq \Delta_{E+i} + 2\sqrt{\Delta_{E+i} \left( \sum_{j=i}^{E-1} \eta_T h_{E+j} \right)} + \left( \sum_{j=i}^{E-1} \eta_T h_{E+j} \right)^2
$$

$$
\leq 2\Delta_{E+i} + 2\left( \sum_{j=i}^{E-1} \eta_T h_{E+j} \right)^2
$$

(52)

$$
\Delta_{E+i} \geq \frac{1}{2} \Delta_{(\tau+1)E} - \left( \sum_{j=i}^{E-1} \eta_T h_{E+j} \right)^2 \geq \frac{1}{2} \Delta_{(\tau+1)E} - \left( \sum_{j=0}^{E-1} \eta_T h_{E+j} \right)^2
$$

(53)

Plug (53) to (49) we can get

$$
(1 + \frac{1}{4} \mu \eta_T \sum_{k=1}^{N} p_k^k \xi_{E+i}^k) \Delta_{(\tau+1)E} \leq \Delta_{E} + \frac{1}{2} \mu \eta_T^2 \sum_{k=1}^{N} p_k^k s_k^k \left( \sum_{i=0}^{E-1} \eta_T h_{E+i} \right)^2 + \eta_T C_T
$$

$$
+ 2\eta_T \sum_{k=1}^{N} p_k^k s_k^k \left( F_k(w^*) - F_k(\bar{w}_{E+i}) \right)
$$

(54)

Define $H_T = (\sum_{i=0}^{E-1} h_{E+i})^2$. Apply Lemma A.2 Lemma A.3 and Assumption 3.4 we have

$$
\mathbb{E}_\xi [h_{E+i}^2] = \mathbb{E}_\xi \| \sum_{k=1}^{N} p_k^k \alpha_{E+i}^k \xi_{E+i}\|^2
$$

$$
\leq \sum_{k=1}^{N} \frac{(p_k^k)^2}{p_k^k} \mathbb{E}_\xi \| \alpha_{E+i}^k \xi_{E+i} \| ^2 \leq \sum_{k=1}^{N} \frac{(p_k^k)^2}{p_k^k} G^2 \alpha_{E+i}^k
$$

(55)

$$
\mathbb{E}_\xi [H_T] = \mathbb{E}_\xi [\sum_{i=0}^{E-1} h_{E+i}^2] \leq \mathbb{E}_\xi [E \sum_{i=0}^{E-1} h_{E+i}^2] \leq E G^2 \sum_{k=1}^{N} \frac{(p_k^k)^2}{p_k^k} s_k^k
$$

(56)

$$
\mathbb{E}_\xi [C_T] = \sum_{i=0}^{E-1} \mathbb{E}_\xi [C_{E+i}] = \sum_{i=0}^{E-1} (p_k^k)^2 s_k^\sigma_k + (2 + \theta)L \sum_{k=1}^{N} p_k^k \xi_{E+i}^k \Gamma_k
$$

$$
+ (2 + \frac{\mu}{2(1+\theta)L}) E(E-1) G^2 \left( \sum_{k=1}^{N} p_k^k s_k^\sigma_k + \theta \sum_{k=1}^{N} p_k^k - 2 \right) + \sum_{k=1}^{N} p_k^k s_k^\sigma_k
$$

(57)
Write $\bar{\Delta}_{E+i} = \mathbb{E}_\xi[\Delta_{E+i}], \bar{C}_\tau = \mathbb{E}_\xi[C_\tau], \bar{H}_\tau = \mathbb{E}_\xi[(\sum_{i=0}^{E-1} h_{E+i})^2]$, then

$$
(1 + \frac{1}{4} \mu \eta \sum_{k=1}^{N} p^k \cdot s^k \cdot \bar{\Delta}_{(\tau+1)E} \leq \bar{\Delta}_{E} + \frac{1}{2} \mu \eta^2 \sum_{k=1}^{N} p^k \cdot s^k \cdot \bar{H}_\tau + \eta^2 \bar{C}_\tau
+ 2\eta \mathbb{E}_\xi \sum_{k=1}^{N} p^k \cdot s^k \cdot (F_k(w^*) - F_k(\bar{w}_{E+i}))
$$

(58)

Let $z_{\tau} = 0$ indicate the event that for all $k$, $\mathbb{E}[p^k \cdot s^k] = c_{\tau} \cdot p^k$ for some constant $c_{\tau}$ that does not depend on $k$, otherwise $s_{\tau}^k = 1$. Note that if $z_{\tau} = 0$, then $\sum_{k=1}^{N} p^k \cdot s^k \cdot (F_k(w^*) - F_k(\bar{w}_{E+i})) = c_{\tau} (F(w^*) - F(\bar{w}_{E+i})) \leq 0$. Otherwise, we have

$$
\sum_{k=1}^{N} p^k \cdot s^k \cdot (F_k(w^*) - F_k(\bar{w}_{E+i})) = \sum_{k=1}^{N} p^k \cdot s^k \cdot (F_k(w^*) - F_k^* + F_k^* - F_k(\bar{w}_{E+i}))
\leq \sum_{k=1}^{N} p^k \cdot s^k \cdot \Gamma_k
$$

(59)

Put it together

$$
\sum_{k=1}^{N} p^k \cdot s^k \cdot (F_k(w^*) - F_k(\bar{w}_{E+i})) \leq z_{\tau} \sum_{k=1}^{N} p^k \cdot s^k \cdot \Gamma_k
$$

(60)

Assume $\eta \leq \frac{4}{\mu \mathbb{E}_\xi} < \frac{4}{\mu \sum_{k=1}^{N} p^k \cdot s^k}$, divide both sides with $1 + \frac{1}{4} \mu \eta \sum_{k=1}^{N} p^k \cdot s^k$, in (58) we can get

$$
\bar{\Delta}_{(\tau+1)E} \leq \left(1 - \frac{1}{4} \mu \eta \sum_{k=1}^{N} p^k \cdot s^k \right) \bar{\Delta}_{E} + 2\eta^2 \bar{H}_\tau + \eta^2 \bar{C}_\tau
+ 2\eta \mathbb{E}_\xi \sum_{k=1}^{N} p^k \cdot s^k \cdot \Gamma_k
$$

(61)

Note that $p^k, s^k$ are independent with $\bar{\Delta}_{E}$. Taking expectation over $p^k$ and $s^k$ we get

$$
\mathbb{E}[\bar{\Delta}_{(\tau+1)E}] \leq \left(1 - \frac{1}{8} \mu \eta \mathbb{E}[\sum_{k=1}^{N} p^k \cdot s^k] \right) \bar{\Delta}_{E} + \eta^2 \mathbb{E} [B_\tau] + 2\eta \mathbb{E}_\xi \sum_{k=1}^{N} \mathbb{E}[p^k \cdot s^k \cdot \Gamma_k]
$$

(62)

### A.2 Proof of Theorem 3.1

When the distributions of $s^k$ do not change with time, we have $B_\tau = B$. We prove the convergence by induction. Let $\eta = \frac{8}{\mu \mathbb{E}[\sum_{k=1}^{N} p^k \cdot s^k] \cdot 2\mathbb{E}_\xi \cdot \tau / \mathbb{E}[\bar{\Delta}_0] \cdot \mathbb{E}[B_\tau]}. Initially, \frac{\tau}{\mathbb{E}[\bar{\Delta}_0]} \geq \mathbb{E}[\Delta_0].$ Suppose $\mathbb{E}[\bar{\Delta}_E] \leq \frac{M_{D+V}}{\tau \mathbb{E}[\bar{\Delta}_E]}$, then
\[
\mathbb{E}[\tilde{\Delta}_{(\tau+1)E}] \leq \frac{\tau E + \gamma - 2E M_r D + V}{\tau E + \gamma} + \left( \frac{16E}{\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]} \right)^2 \frac{B}{(\tau E + \gamma)^2} + \frac{\frac{1}{2} \tau E + D}{\tau E + \gamma}
\]

\[
\leq \frac{\tau E + \gamma - E}{(\tau E + \gamma)^2} (M_r D + V) + \frac{1}{2} \tau E + \gamma + \left( \frac{16E}{\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]} \right)^2 \frac{B}{(\tau E + \gamma)^2} - \frac{E(M_r D + V)}{(\tau E + \gamma)^2} \leq 0
\]

Thus \( \tilde{\Delta}_{(\tau+1)E} \leq \frac{M_r D + V}{(\tau + 1)E + \gamma} \)

We can check it satisfies the previous assumptions regarding \( \eta_r \):

\[
\eta_r \leq \eta_0 = \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])}{E + \gamma} \leq \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])}{32E(1 + \theta)L/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])] = \frac{1}{2(1 + \theta)L} \]

\[
\eta_r \leq \eta_0 = \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])}{E + \gamma} \leq \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])}{4E^2\theta/(\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b])} = \frac{4}{\mu E\theta}
\]

A.2.1 Extension to Time-Varying Distributions

When the distribution of \( s_k^b \) vary with time, we can still establish a convergence with slightly different definitions.

Redefine \( \gamma = \max \left\{ \frac{32E(1+\theta)L}{\mu \min \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b], \min \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]], V_r = \max \left\{ \gamma^2 \mathbb{E}\|w^* - w^0\|^2, \left( \frac{16E}{\mu E} \right)^2 \sum_{i=0}^{\tau-1} \mathbb{E}[B_i] \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b] \right\} \right\} \]

We now prove by induction that with this definition, we can obtain

\[
\mathbb{E}[\tilde{\Delta}_{\tau E}] \leq \frac{M_r D}{\tau E + \gamma} + \frac{V_r}{(\tau E + \gamma)^2} \]

Let \( \eta_r = \frac{8}{\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]} \mathbb{E}[\tilde{\Delta}_{(\tau + 1)E}] \geq \mathbb{E}[\tilde{\Delta}] \). Initially, \( \frac{V_r}{\gamma} \geq \mathbb{E}[\tilde{\Delta}] \). Suppose \( \mathbb{E}[\tilde{\Delta}_{\tau E}] \leq \frac{M_r D}{\tau E + \gamma} + \frac{V_r}{(\tau E + \gamma)^2} \), then

\[
\mathbb{E}[\tilde{\Delta}_{(\tau+1)E}] \leq \frac{\tau E + \gamma - E}{(\tau + 1)E + \gamma} (M_r D + V) + \left( \frac{16E}{\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]} \right)^2 \frac{B}{(\tau E + \gamma)^2} + \frac{\frac{1}{2} \tau E + D}{\tau E + \gamma}
\]

\[
\leq \frac{\tau E + \gamma - E}{(\tau E + \gamma)^2} (M_r D + V) + \left( \frac{16E}{\mu \mathbb{E}[\sum_{k=1}^N p_k^b s_k^b]} \right)^2 \frac{B}{(\tau E + \gamma)^2} + \frac{\frac{1}{2} \tau E + D}{\tau E + \gamma}
\]

Thus \( \tilde{\Delta}_{(\tau+1)E} \leq \frac{M_{r+1} D V_{\tau+1}}{(\tau + 1)E + \gamma} \frac{V_{\tau+1}}{(\tau + 1)E + \gamma)^2} \)

Easy to check previous assumptions regarding \( \eta_r \) are all satisfied.
A.3 Proof of Theorem 3.2

- Departure Case: \( \tilde{n} = n - n_l \)

\[
\|w^* - \hat{w}^*\| \leq \frac{2}{\mu} \|\nabla F(\hat{w}^*)\| = \frac{2}{\mu} \left\| \nabla F(\hat{w}^*) - \nabla \hat{F}(\hat{w}^*) \right\| = 0
\]

\[
= \frac{2}{\mu} \sum_{k \neq l} (p^k - \bar{p}^k) \nabla F_k(\hat{w}^*) + p^l \nabla F_l(\hat{w}^*)
\]

\[
= \frac{2}{\mu} \sum_{k \neq l} \left( \frac{n_k}{n} - \frac{n_k}{n - n_l} \right) \nabla F_k(\hat{w}^*) + p^l \nabla F_l(\hat{w}^*)
\]

\[
= \frac{2}{\mu} \left( -p^l \sum_{k \neq l} \bar{p}^k \nabla F_k(\hat{w}^*) + p^l \nabla F_l(\hat{w}^*) \right)
\]

\[
= \frac{2p^l}{\mu} \|\nabla F_l(\hat{w}^*)\| \leq \frac{2p^l}{\mu} \sqrt{2L (F_l(\hat{w}^*) - F_l^*)} = \frac{2\sqrt{2L}}{\mu} \sqrt{\Gamma_l}
\]

- Arrival Case: \( \hat{n} = n + n_l \)

\[
\|w^* - \hat{w}^*\| = \|w^* - \hat{w}^*\| \leq \frac{2}{\mu} \|\nabla \hat{F}(w^*)\| = \frac{2}{\mu} \left\| \nabla \hat{F}(w^*) - \nabla F(w^*) \right\| = 0
\]

\[
= \cdots = \frac{2}{\mu} \left( \frac{1}{\mu} \sum_{k \neq l} p^k \nabla F_k(w^*) + \bar{p}^l \nabla F_l(w^*) \right)
\]

\[
= \frac{2p^l}{\mu} \|\nabla F_l(w^*)\| = \frac{2\sqrt{2L}}{\mu} \bar{p}^l \sqrt{\Gamma_l}
\]

A.4 Proof of Corollary 4.0.1

A.4.1 Scheme A

In Scheme A, we only consider devices whose \( s_k^\tau = E \). Let \( q_k^\tau \) be an indicator denoting if client \( k \) is complete in round \( \tau \). Thus, \( K_\tau = \sum_{k=1}^{N} q_k^\tau \).

**Homogeneous participation.** Obviously \( q_k^\tau \)'s are homogeneous when \( s_k^\tau \)'s are homogeneous. Thus, \( E[q_k^\tau] = q_\tau \), where \( q_\tau = P(s_\tau = E) \). We then have \( P(K_\tau = 0) = (1 - q_\tau)^N \). When choosing \( p_k^\tau = \frac{Nk^\tau q_k^\tau}{N} \), \( \theta = N \). Note that by the definition of \( q_k^\tau \), we have \( q_k^\tau s_k^\tau = Eq_k^\tau \), so \( E[p_k^\tau s_k^\tau] = E[q_k^\tau] \). Similarly, we can replace all \( s_k^\tau \) terms with \( E \). Next we calculate \( E[p_\tau^\tau] \):
\[ E_q[p^k_r | K_r \neq 0] = N p^k q^k \sum_{i=1}^{N} \left(1 + \frac{1}{(1 + q_r)^i}\right) \left(\frac{q_r^i (1 - q_r)^N}{1 - (1 - q_r)^N}\right) \] 

\[ = N p^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

\[ = N p^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

\[ = p^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} = p^k \] 

Similarly,

\[ E_q[(p^k_r)^2 | K_r \neq 0] = (N p^k)^2 \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

\[ = N p^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

It is possible that \( \sum_{k=1}^{N} p^k r^k > 2 \), so we need to calculate \( E[p^k p^k_r | K_r \neq 0] \)

\[ E_q[p^k r^k_r | K_r \neq 0] = N^2 p^k r^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

\[ = N p^k \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

For all \( k \) and \( \tau \), \( E[p^k s^k_k | K_r \neq 0] = E p^k \), thus \( z_r = 0, M_r = 0 \) for all \( k, \tau \).

Therefore, \( E[B] = O(N^2 E[\frac{1}{K_r} | K_r \neq 0] + \sum_{k=1}^{N} (p^k \sigma_k)^2 + \Gamma), \gamma = O(N), \) hence \( V = O(N^2 E[\frac{1}{K_r} | K_r \neq 0] + \sum_{k=1}^{N} (p^k \sigma_k)^2 + \Gamma) \). Plug them into Theorem 3.1, we can get an asymptotic rate of \( O\left(\frac{E^2 \sigma^2}{\tau^2} + \sigma^2 + 1\right)\).

**Heterogeneous Participation.** When \( s^k_k \)'s (i.e., \( q^k_r \)'s) are heterogeneous, generally \( E[p^k] \neq p^k \), furthermore, we may have \( z_r = 1 \) for all \( \tau \). To see this, consider an example where a device \( k_0 \) has \( q^{k_0}_r = 1 \), i.e. \( P(s^k_r = E) = 1 \), whereas all the rest devices have \( E[q^k_r] = q_r \), then we can show that

\[ E_q[p^k_{k_0} | K_r \neq 0] = E_q[p^k_{k_0}] = N p^k_{k_0} \sum_{i=1}^{N} \frac{1}{i} \frac{(N - 1)!}{(i - 1)!} \frac{(N - 1)!}{(N - i)!} \frac{(q_r)^i (1 - q_r)^{N-i}}{1 - (1 - q_r)^N} \] 

\[ = p^k_{k_0} \frac{1}{q_r} \frac{1}{1 - (1 - q_r)^N} \] 

\[ = p^k_{k_0} \frac{1}{q_r} \frac{1}{1 - (1 - q_r)^N} \] 

and for \( k \neq k_0 \)
Thus, different $k$ will have different ratio of $E[p_k^ks_k^*/p^k] = E[p_k^k/p^k]$, which indicates $z_r = 1$. Since this is true for all $\tau$, we have $M_r = \tau$. Therefore according to Theorem 3.1, the learning will not converge to the global optimal, and the remainder loss is bounded by $D/E$.

### A.4.2 Scheme B

In Scheme B, $p_k^k = p^k$ is a fixed number, so we only need to take expectation over $s_k^k$, and $c_p = 1$. Since $\sum_{k=1}^{N} p_k^k = 1 < 2$, we can bound $E[(\sum_{k=1}^{N} p_k^k - 2) + (\sum_{k=1}^{N} p_k^k s_k^k - Ns_r)] < 0$.

**Homogeneous Participation**. When $s_k^k$'s are homogeneous, i.e. $s_k^k \sim s_r$, then $E[p_k^k s_k^k/p^k] = E[s_r]$. This is the same for all $k$, thus $z_r = 0$ and $M_r = 0$. Moreover, we have $E[B] = O(E[s_r](\sigma_N^2 + \Gamma))$, $\gamma = O(1/E[s_r])$, $V = O \left( (\sigma_N^2 + \Gamma) \frac{1}{E[s_r]} \right)$, which yields an asymptotic convergence rate of $O \left( \frac{p_k^k + \Gamma}{E[s_r]} \right)$.

**Heterogeneous Participation**. When $s_k^k$'s are heterogeneous, $E[p_k^k s_k^k/p^k] = E[s_k^k]$ varies with $k$. Thus, $z_r = 1$ and $M_r = \tau$. Therefore, the algorithm will not converge to the global optimum according to Theorem 3.1.

### A.4.3 Scheme C

In Scheme C, $p_k^e = \frac{E_{p_k^e}}{s_k^e}$, so $\theta = E$. It is possible that $\sum_{k=1}^{K} p_k^k > 2$, so we need to calculate $E \left[ (\sum_{k=1}^{N} p_k^k) (\sum_{k=1}^{N} p_k^k s_k^k) \right]$.

**Homogeneous Participation**. When $s_k^k$'s are homogeneous, $E[p_k^k s_k^k/p^k] = E[s_k^k]$ for all $k$. Thus, $z_r = 0, M_r = 0$. Moreover, we have

$$E[\sum_{k=1}^{N} p_k^k] = E \left[ \frac{1}{s_r} \right]$$  
(73)

$$E[\sum_{k=1}^{N} (p_k^k)^2] = \left( E \left[ \frac{1}{s_r} \right] \right)^2 \sum_{k=1}^{N} (p_k^k)^2$$  
(74)

$$E[\sum_{k=1}^{N} (p_k^k)^2 s_k^k] = E^2 \left[ \frac{1}{s_r} \right] \sum_{k=1}^{N} (p_k^k)^2$$  
(75)

$$E \left[ \sum_{k=1}^{N} p_k^k \sum_{k=1}^{N} p_k^k s_k^k \right] = E^2 \left[ \frac{1}{s_r} \right]$$  
(76)

Therefore, we have $E[B] = O \left( E \left[ \frac{1}{s_r} \right] (\sigma_N + \Gamma) \right) = V$, which yields a convergence rate of $O \left( \frac{\sigma_N^2 + \Gamma}{\gamma E[s_r]} \right)$.

**Heterogeneous Participation**. Even when $s_k^k$'s are heterogeneous, we still have $E[p_k^k s_k^k/p^k] = E$ for active all
We first introduce the following lemma:

We are interested in the distance between $\tilde{w}$ and $w'$.

Next we investigate the effect of applying additional update from $l$. Suppose device $l$ arrives, then for any $w$, we have

$$F_l(w) = \frac{1}{p_l} \left( \frac{\bar{F}(w) - \frac{n}{N} F(w)}{\tau} \right)$$

**Proof.** We expand the right hand side expression and show it equals $F_l(w)$:

$$\frac{1}{p_l} \left( \frac{\bar{F}(w) - \frac{n}{N} F(w)}{\tau} \right) = \frac{1}{p_l} \left( \sum_{k=1}^{N} \hat{F}_k (w) + \frac{n}{n} F_l(w) - \sum_{k=1}^{N} \frac{n}{p_k} F_k(w) \right)$$

$$= \frac{1}{p_l} \left( \sum_{k=1}^{N} \hat{F}_k (w) + \frac{n}{n} F_l(w) - \sum_{k=1}^{N} \frac{n}{p_k} F_k(w) \right) = F_l(w) \quad \square$$

Next we investigate the effect of applying additional update from $l$. Suppose the current global weight is $w^{T}_{\tau E} = w$, and assume we perform full batch gradient for the additional update. After the update, it becomes

$$w' = w - \eta_{\tau} \delta_l' \nabla F_l(w)$$

We are interested in the distance between $w'$ and the new global optimum $\tilde{w}^*$:

$$\|w' - \tilde{w}\|^2 = \|w - \eta_{\tau} \delta_l' \nabla F_l(w) - \tilde{w}\|^2$$

$$= \|w - \tilde{w}\|^2 - 2 \eta_{\tau} \delta_l' \langle w - \tilde{w}, \nabla F_l(w) \rangle + \left( \eta_{\tau} \delta_l' \right)^2 \|\nabla F_l(w)\|^2$$

**Proof.** We expand the right hand side expression and show it equals $F_l(w)$:

$$A(w, \delta_l') = -2 \frac{\eta_{\tau} \delta_l'}{p_l} \langle w - \tilde{w}, \nabla \bar{F}(w) - \frac{n}{n} \nabla F(w) \rangle + \left( \eta_{\tau} \delta_l' \right)^2 \|\nabla F_l(w)\|^2$$

A.5 Proof of Corollary 4.0.2

We first introduce the following lemma:

**Lemma A.4.** Suppose device $l$ arrives, then for any $w$, we have

$$F_l(w) = \frac{1}{p_l} \left( \frac{\bar{F}(w) - \frac{n}{N} F(w)}{\tau} \right)$$

**Proof.** We expand the right hand side expression and show it equals $F_l(w)$:

$$\frac{1}{p_l} \left( \frac{\bar{F}(w) - \frac{n}{N} F(w)}{\tau} \right) = \frac{1}{p_l} \left( \sum_{k=1}^{N} \hat{F}_k (w) + \frac{n}{n} F_l(w) - \sum_{k=1}^{N} \frac{n}{p_k} F_k(w) \right)$$

$$= \frac{1}{p_l} \left( \sum_{k=1}^{N} \hat{F}_k (w) + \frac{n}{n} F_l(w) - \sum_{k=1}^{N} \frac{n}{p_k} F_k(w) \right) = F_l(w) \quad \square$$

Next we investigate the effect of applying additional update from $l$. Suppose the current global weight is $w^{T}_{\tau E} = w$, and assume we perform full batch gradient for the additional update. After the update, it becomes

$$w' = w - \eta_{\tau} \delta_l' \nabla F_l(w)$$

We are interested in the distance between $w'$ and the new global optimum $\tilde{w}^*$:

$$\|w' - \tilde{w}\|^2 = \|w - \eta_{\tau} \delta_l' \nabla F_l(w) - \tilde{w}\|^2$$

$$= \|w - \tilde{w}\|^2 - 2 \eta_{\tau} \delta_l' \langle w - \tilde{w}, \nabla F_l(w) \rangle + \left( \eta_{\tau} \delta_l' \right)^2 \|\nabla F_l(w)\|^2$$

Obviously, the additional update helps fast-reboot if $A(w, \delta_l') < 0$.

Applying Lemma A.4 we can get

$$A(w, \delta_l') = -2 \frac{\eta_{\tau} \delta_l'}{p_l} \langle w - \tilde{w}, \nabla \bar{F}(w) - \frac{n}{n} \nabla F(w) \rangle + \left( \eta_{\tau} \delta_l' \right)^2 \|\nabla F_l(w)\|^2$$
Write $b = w - w^*$, and use the mean value theorem we have

\[-\langle w - w^*, \nabla F_1(w) \rangle = -\langle b + w^* - w^*, \nabla F_1(w^*) + \nabla^2 F_1(\xi)b \rangle
\]
\[= -\langle w^* - w^*, \nabla F_1(w^*) \rangle - \langle w^* - w^*, \nabla^2 F_1(\xi)b \rangle - \langle b, \nabla F_1(w) \rangle
\]
\[\leq -\langle w^* - w^*, \nabla F_1(w^*) \rangle + \|w^* - w^*\|\|\nabla^2 F_1(\xi)\|_2\|b\| + \|\nabla F_1(w)\|\|b\|
\]
\[\leq -\langle w^* - w^*, \nabla F_1(w^*) \rangle + (\|w^* - w^*\| + 1)W\|b\|
\]
\[\leq -\frac{1}{\tilde{p}} \langle w^* - w^*, \nabla \tilde{F}(w^*) \rangle - \frac{n}{n} \nabla F(w^*) + \left(\frac{2\sqrt{2L}}{\mu} \tilde{p}^l \sqrt{\Gamma_i + 1}\right) W\|b\|
\]
\[\leq -\frac{1}{\tilde{p}} \left(\tilde{F}(w^*) - \tilde{F}(\tilde{w}^*)\right) + \left(\frac{2\sqrt{2L}}{\mu} \tilde{p}^l \sqrt{\Gamma_i + 1}\right) W\|b\|
\]

Therefore,

\[A(w, \delta^l) \leq 2\frac{n\eta\delta^l}{\tilde{p}} \left(\left(\frac{2\sqrt{2L}}{\mu} \tilde{p}^l \sqrt{\Gamma_i + 1}\right) W\|b\| - \left(\tilde{F}(w^*) - \tilde{F}(\tilde{w}^*)\right)\right) + (\eta_c\delta^l)^2 W^2
\]

For $\delta^l > 0$, the right hand side can be negative if and only if $\|b\| < \frac{\tilde{F}(w^*) - \tilde{F}(\tilde{w}^*)}{\left(\frac{2\sqrt{2L}}{\mu} \tilde{p}^l \sqrt{\Gamma_i + 1}\right) \tilde{p}^l W}$.

### A.6 Proof of Corollary 4.0.3

The loss bound without objective shift is $f_0(\tau) = \frac{(\tau - \tau_0)D + V}{\tau E + \gamma}$, and the bound with shift is $f_1(\tau) = \frac{\tau \gamma + \Gamma_i}{(\tau - \tau_0)E + \gamma}$.

Note that $f_0(\tau)$ is a monotonic function. When it is increasing, we just need $f_0(\tau_0) = f_1(\tau)$, which yields

\[\tau - \tau_0 = 1 - \gamma + \frac{\Gamma_i(\tau_0 E + \gamma)}{E V} = O\left(\frac{\Gamma_i \tau_0}{V}\right)
\]

Now we consider monotonically decreasing $f_0(\tau)$, which is more commonly observed in experiments. Let $C_1 = DE, C_2 = \gamma D + \gamma E - E \Gamma_i, C_3 = V(\gamma - 1)$, the only possible root for the quadratic equation $f_0(\tau) = f_1(\tau)$ is

\[\tau - \tau_0 = \frac{EV}{\tau_0 E + \gamma} - C_2 + \sqrt{4C_1 \Gamma_i(\tau_0 E + \gamma) + \left(\frac{EV}{\tau_0 E + \gamma}\right)^2 - \frac{2C_2 EV}{\tau_0 E + \gamma} + (C_2^2 - 4C_1 C_3)}
\]

\[= O(\sqrt{\tau_0 \Gamma_i})
\]