Towards Flexible Device Participation in Federated Learning for Non-IID Data

Yichen Ruan, Xiaoxi Zhang, Shu-Che Liang, Carlee Joe-Wong
Department of Electrical and Computer Engineering
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. In this paper, we extend the current learning paradigm and consider devices that may become inactive, compute incomplete updates, and leave or join in the middle of training. We derive analytical results to illustrate how the flexible participation of devices could affect the convergence when data is not independently and identically distributed (IID), and when devices are heterogeneous. This paper proposes a new federated aggregation scheme that converges even when devices may be inactive or return incomplete updates. We finally discuss practical research questions an operator would encounter during the training, and provide answers based on our convergence analysis.

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. General 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:

- Synchronization: the coordinator synchronizes the latest *global model* with all devices.
- Local updates: each device trains a *local model* for a few *local epochs*, using samples from its *local dataset*.
- Aggregation: the coordinator aggregates some, or all, of the local models to produce the next global model.

Depending on the use-cases, federated learning can be divided into cross-device learning, where participating entities are mostly IoT devices, and cross-silo learning, where participating entities are different organizations that own separate enterprise databases [Kairouz et al., 2019]. Our work focuses on the cross-device learning, and we typically consider mobile devices such as smart phones and tablets. Theses devices generally have limited computing and communication resources, e.g., due to battery limitations, and have different training data distributions, i.e., data is not independently and identically distributed (non-IID) among devices [Li et al., 2020]. To relieve the computation and communication burden, in the last step of the training procedure, the federated learning coordinator may only aggregate a subset of local models, which is referred to as the partial device participation scheme. However, only a few device selection policies have been proved to work in the non-IID setting, and the selection must be independent of the status of devices [Li et al., 2020]. In other words, for the training to converge successfully, all selected devices must be able to train their local models

Preprint. Under review.

and upload the results whenever they are selected. This is why the traditional federated learning paradigm requires participating devices to be dedicated to the training during the entire federated learning period. For example, the popular *FedAvg* algorithm assumes mobile users will participate only when their phones are currently plugged-in, and have unlimited WI-FI access [McMahan et al., 2016].

Considering that federated learning typically takes thousands 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. Even for cross-silo applications, where more powerful computers or cloud servers may be adopted, devices' availability can still be an issue due to the increasingly popular usage of preemptive cloud services such as AWS's spot instances, where the user process can be interrupted unexpectedly [Zhang et al., 2020]. While many other methods have been proposed to mitigate the workload of individual devices, such as weight compression and federated dropout [Caldas et al., 2018], they cannot completely remove the probability of confronting undesirable device behaviors during the training. This probability, intuitively, increases as more devices join the training. Therefore, in large scale federated learning, many low-end 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, the existing training procedure does not specify how to react when confronting irregular 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 more devices to follow more flexible participation patterns. Specifically, the paper incorporates four situations that are not yet well discussed in the literature:

- In-completeness: devices submit only partially completed work in a round.
- In-activeness: devices do not respond to the coordinator in a round.
- Early departures: existing devices quit the training without finishing all training rounds.
- Late arrivals: new devices join after the training has already started.

The difference between in-activeness 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 are going to leave, and do not plan to rejoin the system. 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 leaving. A possible solution to incorporate flexible device participation is through the online learning framework, where datasets keep changing and the training can last forever. Though there are some preliminary explorations in this direction [Chen et al., 2019], it has not generally been verified that online learning can actually work in the federated setting. Thus, in this paper, we consider traditional fixed time learning where the training must terminate before some deadline, i.e., maximum number of rounds. The results of our paper can help to answer the following research questions.

- **How can we utilize the partially completed work?** Federated learning requires the participating devices to run a certain number of local epochs between two aggregations. Due to resource competition as mentioned above, some devices may not be able to complete all local epochs as expected. An easy solution is to simply discard the partial work and re-start this round, which slows down the training progress. We will show that devices' partial work can be reused without greatly affecting the convergence compared to full completion by all devices if we properly *scale the aggregating weight* in the aggregation step.
- **Should we remove an inactive device?** In the extreme case, a device may temporarily become inactive and fail to submit its update to the coordinator. We will show that a mild degree of in-activeness will not greatly affect the convergence. However, if a device is expected to be frequently inactive during the training period, we may want to kick it out and assume it quits the training, in order to prevent the device from interfering with the algorithm convergence. We answer the question: *what kind of devices should be removed from training*?
- **What to do if a device departs?** When an existing device leaves early, the coordinator will no longer receive updates from it. The trained model will inevitably move away from the leaving

device, which imposes additional bias on the training loss. This bias is less severe if the training is already close to the end. But if the leaving happens at an early stage, the bias may accumulate and cause the training to diverge. In this case, we can *shift the objective* to "forget" the influence of this device. After the shifting process, the trained model will no longer be applicable to the leaving dataset, but will have a higher accuracy on the remaining datasets. There thus exists a *threshold time* when we prefer shifting to sticking to the status quo.

- **When should we admit an arriving device?** When the training time is relatively long, more high-quality datasets can be discovered during the process. The operator may hope to include these devices to increase the generalizability of the trained model. Every time a device arrives, the current training progress will be interrupted, and we need to add a new term to the training objective that corresponds to the new device. In other words, the trained model, which is originally only intended for the old datasets, has to be updated so that it can work on the newly admitted data points. To establish the convergence of this process, sufficient training time must remain before the deadline so that the model can fully adapt to the new objective. In this work we will show how to find *the latest possible time to admit a new device*.

In Section 2, we will review relevant literature. In Section 3, we will give an convergence analysis that considers the four situations. Based on this analysis, we will answer our four research questions in Section 4, and compare our convergence rate analytically with other possible alternatives.

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 averages the updated parameters from a small set of end devices periodically. However, it lacks theoretical convergence guarantees in realistic settings, such as when the local data is not independent and identically distributed across devices (non-IID data). A few recent works provide theoretical results for the non-IID data case. For instance, in [Li et al., 2020], the authors analyze the convergence of FedAvg on non-IID data and establish an $O(\frac{1}{T})$ 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., 2020][Zhao et al., 2018], or consider alternative objective functions for the SGD algorithm to optimize [Smith et al., 2017].

The algorithm FedAvg with non-IID data across devices has also been modified in specific edge computing scenarios to reduce the communication overhead [Liu et al.] 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 is a possible way to enable flexible device participation in the federated learning scenario. For instance, [Chen et al., 2019] proposes an asynchronous federated learning algorithm to handle inbalanced 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 inactiveness or incompleteness and data heterogeneity.

3 Convergence Analysis

In this section, we establish the convergence bound for federated learning with flexible device participation patterns. Our analysis is based on the standard *FedAvg* algorithm where local devices use stochastic gradient descent (SGD) as the local optimizer. 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 3.3, and then we discuss the effects of device arrival and departure on the algorithm convergence in Section 3.4.

3.1 Algorithm Description

Suppose there are N devices, and that each device k wishes to minimize its local loss 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 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 the global objective F, and denote by F_k^* the optimal value of F_k. We quantify the degree to which data at each device k is distributed differently than that at other devices as $\Gamma_k = F_k(w^*) - F_k^*$, and let $\Gamma = \sum_{k=1}^{N} p^k \Gamma_k$ [Li et al., 2020].

We consider discrete time steps $t = 0, 1, \dots$. 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 τth round), the following three steps are executed:

- First, the coordinator broadcasts the latest global weight $w_{\tau E}^{\mathcal{G}}$ to all devices. Each device updates its local weight so that: $w_{\tau E}^k = w_{\tau E}^{\mathcal{G}}$
- Second, each device runs SGD on its local objective F_k for $i = 0, \dots, s_\tau^k - 1$:

$$w_{\tau E+i+1}^k = w_{\tau E+i}^k - \eta_\tau g_{\tau E+i}^k \tag{1}$$

  Here η_τ is a staircase learning rate that decays with τ, $0 \le s_\tau^k \le E$ represents the number of local updates this device completes in this round, $g_t^k = \nabla F_k(w_t^k, \xi_t^k)$ is the stochastic gradient at device k, and ξ_t^k is a mini-batch sampled from device k's local dataset. We also define $\bar{g}_t^k = \nabla F_k(w_t^k)$ as the full batch gradient at device k, hence $\bar{g}_t^k = \mathbb{E}_{\xi_t^k}[g_t^k]$.
- Third, the coordinator aggregates the next global weight as

$$w_{(\tau+1)E}^{\mathcal{G}} = w_{\tau E}^{\mathcal{G}} + \sum_{k=1}^{N} p_\tau^k (w_{\tau E + s_\tau^k} - w_{\tau E}^{\mathcal{G}}) = w_{\tau E}^{\mathcal{G}} - \sum_{k=1}^{N} p_\tau^k \sum_{i=0}^{s_\tau^k} \eta_\tau g_{\tau E+i}^k \tag{2}$$

We define that a device k is *inactive* in round τ if $s_\tau^k = 0$ (i.e., it completes no local updates), and say it is *incomplete* if $0 < s_\tau^k < E$. We treat each s_τ^k as a random variable that can follow an arbitrary distribution. For all devices, s_τ^k can generally be time-varying, i.e., it may follow different distributions at different time steps. Devices are *heterogeneous* if they have different distributions (series) of s_τ^k, and otherwise they are *homogeneous*. We also allow the aggregation coefficients p_τ^k to vary with τ. In Section 4, we will discuss different schemes of choosing p_τ^k and their impacts on the training convergence.

As a special case, traditional *FedAvg* assumes all selected devices can complete all E local epochs, so that $s_\tau^k \equiv E$. Also, *FedAvg* with full device participation uses fixed aggregation coefficients $p_\tau^k \equiv p^k$, so that the right hand side of (2) can be written as $\sum_{k=1}^{N} p^k w_{\tau E}^k$, i.e., aggregating gradients is equivalent to aggregating the model parameters directly.

3.2 Assumptions

The analysis relies on the following five assumptions. The first four are standard [Li et al., 2020]. The last assumption excludes the extreme case of choosing infinitely large aggregation coefficients.

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

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

Assumption 3.3. *The variance of the stochastic gradients is bounded: $\mathbb{E}_\xi \|g_t^k - \bar{g}_t^k\|^2 \le \sigma_k^2$, $\forall k, t$.*

Assumption 3.4. *The expected squared norm of the stochastic gradients at each local device is uniformly bounded: $\mathbb{E}_\xi \|g_t^k\|^2 \le G^2$ for all k and t.*

Assumption 3.5. *There exists an upper bound for the aggregation coefficient $p_\tau^k / p^k \le \theta$.*

3.3 Convergence Bound

Assume the following expectations exist and can be computed or bounded: $\mathbb{E}[p_\tau^k]$, $\mathbb{E}[p_\tau^k s_\tau^k]$, $\mathbb{E}[(p_\tau^k)^2 s_\tau^k]$, $\mathbb{E}[(\sum_{k=1}^N p_\tau^k - 2)_+(\sum_{k=1}^N p_\tau^k s_\tau^k)]$ for all rounds τ and devices k, and assume $\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k] \neq 0$; intuitively, this last assumption ensures that some updates are aggregated in each round. Let $z_\tau \in \{0,1\}$ indicate the event that the ratio $\mathbb{E}[p_\tau^k s_\tau^k]/p^k$ take the same value for all k. We can obtain the convergence bound as follows:

Theorem 3.1. *By choosing the learning rate* $\eta_\tau = \frac{8}{\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]} \frac{2E}{(\tau+1)E+\gamma}$, *we can obtain*

$$\mathbb{E}\|w_{\tau E}^{\mathcal{G}} - w^*\|^2 \leq \frac{M_\tau D}{\tau E + \gamma} + \frac{V_\tau}{(\tau E + \gamma)^2} \tag{3}$$

Here $\gamma = \max\left\{\frac{32E(1+\theta)L}{\mu \min_\tau \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}, \frac{4E^2\theta}{\min_\tau \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}\right\}$, $M_\tau = \sum_{t=0}^{\tau-1}\mathbb{E}[z_t]$, $D = \max_\tau\left\{\frac{32E\sum_{k=1}^N \mathbb{E}[p_\tau^k s_\tau^k]}{\Gamma_k/(\mu\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}\right\}$, $V_\tau = \max\left\{\gamma^2 \mathbb{E}\|w_0^{\mathcal{G}} - w^*\|^2, \left(\frac{16E}{\mu}\right)^2 \sum_{t=0}^{\tau-1} \frac{\mathbb{E}[B_t]}{\left(\mathbb{E}[\sum_{k=1}^N p_t^k s_t^k]\right)^2}\right\}$, *and* $B_t = \sum_{k=1}^N (p_t^k)^2 s_t^k \sigma_k^2 + 2(2+\theta)L\sum_{k=1}^N p_t^k s_t^k \Gamma_k + \left(2 + \frac{\mu}{2(1+\theta)L}\right) E(E-1)G^2 \left(\sum_{k=1}^N p_t^k s_t^k + \theta(\sum_{k=1}^N p_t^k - 2)_+ \sum_{k=1}^N p_t^k s_t^k\right) + 2EG^2 \sum_{k=1}^N \frac{(p_t^k)^2}{p^k} s_t^k$

Note that $V_\tau = O(\tau)$, so $w_{\tau E}^{\mathcal{G}}$ will eventually converge to a globally optimal solution as $\tau \to \infty$ if M_τ increases sub-linearly with τ. In the original *FedAvg*, $p_\tau^k s_\tau^k \equiv p^k E$, thus $z_\tau \equiv 0$ and $M_\tau \equiv 0$ as per the definitions. When considering flexible device participation patterns, M_τ may increase with τ. But as we will see in Section 4, by smartly choosing the aggregation coefficients p_τ^k, the increase can be controlled to ensure convergence.

3.4 Shifts in the Global Objective

When a device k quits at some $\tau_0 < T$, no more updates will be received from it. Thus, $s_\tau^k = 0$ for all $\tau > \tau_0$. As a result, the ratio $\mathbb{E}[p_\tau^k s_\tau^k]/p^k$ must take different values for different k, when $\tau > \tau_0$, since otherwise all p_τ^k will equal zero. Therefore, $z_\tau = 1$ for all $\tau > \tau_0$. According to (3), $w_{\tau E}^{\mathcal{G}}$ then cannot converge to the global optimal w^* as $M_T \geq T - \tau_0$. This can be intuitively explained as follows: When the data distribution is non-IID, one device must contribute sufficiently more updates in order for its data's characteristics to be absorbed by the trained model. After a device leaves, the remaining training steps will wash out the model's previous memory about the leaving device, and push the model towards the data at the remaining devices. Thus, the output model may not still be applicable to the leaving device, especially when it leaves early in the training ($\tau_0 \ll T$). This result suggests that we may wish to discard a departing device if we cannot guarantee the trained model will perform well on it, and that the earlier a device quits, the more likely we are to discard it.

Furthermore, once we decide to discard the departing device (say the lth device) and exclude it from the training, sticking to the original learning objective $F = \sum_{k=1}^N p^k F_k$ is meaningless. Instead, it is more reasonable to update F as $\tilde{F} = \sum_{k=1, k\neq l}^N p^k F_k$, i.e., to remove F_l from the learning objective. As a result, the optimal weight w^* will shift to some $\tilde{w}^*$ that minimizes $\tilde{F}$. Generally, $\tilde{w}^*$ is different from w^*, which produces an additional offset to the convergence bound in Theorem 3.1 since $\|w_{\tau E}^{\mathcal{G}} - \tilde{w}^*\| \leq \|w_{\tau E}^{\mathcal{G}} - w^*\| + \|w^* - \tilde{w}^*\|$. A sufficient number of updates are then required for $w_{\tau E}^{\mathcal{G}}$, which is originally moving towards w^*, to converge to the new optimal $\tilde{w}^*$.

The same argument holds when a new device arrives; admitting this device will require changing the global objective function to include the loss on its data. The learning rate also needs to be increased after shifting the objective. Intuitively, if the shift happens at a large time τ_0, when $w_{\tau_0 E}^{\mathcal{G}}$ is close to the old optimal w^* and η_{τ_0} is close to zero, reducing the latest difference $\|w_{\tau_0 E}^{\mathcal{G}} - \tilde{w}^*\| \approx \|w^* - \tilde{w}^*\|$ with such a small learning rate is unreasonable. Thus, a greater learning rate is required, which is equivalent to initiate a fresh start after the shift. Nevertheless, even with a larger learning rate, there still must be enough time left for the model to fully absorb the new device.

The following theorem bounds the offset of the optimal weight due to the objective shift:

Theorem 3.2. *Consider the objective shift $F \to \tilde{F}$, $w^* \to \tilde{w}^*$. Let $\tilde{\Gamma}_k = F_k(\tilde{w}^*) - F_k^*$ quantify the degree of non-IID data with respect to the new objective. Then in the device departure case*

$$\|w^* - \tilde{w}^*\| \leq \frac{2\sqrt{2L}}{\mu} \frac{n_l}{n} \sqrt{\tilde{\Gamma}_l} \tag{4}$$

and in the arrival case

$$\|w^* - \tilde{w}^*\| \leq \frac{2\sqrt{2L}}{\mu} \frac{n_l}{n + n_l} \sqrt{\Gamma_l} \tag{5}$$

Here l indexes the departing/arriving device, and n is the total number of data points before the shift.

As we can intuitively expect, the difference reduces when the data becomes more IID ($\Gamma_l \to 0$), and when the departing/arriving device owns fewer data points ($n_l \to 0$).

4 Discussion

Based on the convergence analysis in Section 3, in this Section, we will answer the research questions in Section 1 associated with the four flexible participation patterns.

4.1 Schemes of Choosing p_τ^k

According to Theorem 3.1, the convergence bound is controlled by expectation of p_τ^k and its functions. Below we discuss three schemes of choosing p_τ^k, and compare their convergence rates in Table 1.

- Scheme A: Only aggregate parameters from clients who complete all E local epochs, with weight $p_\tau^k = \frac{Np^k}{K_\tau} q_\tau^k$, where K_τ is the number of complete devices, $q_\tau^k \in \{0, 1\}$ denotes if client k is complete, i.e., it completes E epochs. If $K_\tau = 0$, this round is discarded.
- Scheme B: Allow clients to upload incomplete work, with fixed weight $p_\tau^k = p^k$.
- Scheme C: Accept partial work, with adaptive weight $p_\tau^k = \frac{E}{s_\tau^k} p^k$, or 0 if $s_\tau^k = 0$.

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 seems counter-intuitive, as fewer local updates might lead to less optimal parameters, it turns out to be the only scheme that guarantees convergence when device participation is heterogeneous.

Corollary 4.0.1. *Table 1 gives the convergence rates of Schemes A, B, and C when device updates may be incomplete.*

	Homogeneous Devices	Heterogeneous Devices
A	$O\left(\frac{E^3 N^2 \sum_{t=0}^{\tau-1} \mathbb{E}[\frac{1}{K_\tau} \| K_\tau \neq 0]}{(\tau E + EN)^2}\right)$	N/A
B	$O\left(\frac{\left(E^4 + E^2 \sum_k^N (p^k \sigma_k)^2\right) \sum_{t=0}^{\tau-1} \frac{1}{\mathbb{E}[s_t]}}{(\tau E + E)^2}\right)$	N/A
C	$O\left(\frac{\left(E^5 + E^2 \sum_k^N (p^k \sigma_k)^2\right) \sum_{t=0}^{\tau-1} \mathbb{E}\left[\frac{1}{s_t}\right]}{(\tau E + E^2)^2}\right)$	$O\left(\frac{E^5 \sum_{t=0}^{\tau-1} \sum_k^N p^k \mathbb{E}\left[\frac{1}{s_t^k}\right] + E^2 \sum_{t=0}^{\tau-1} \sum_k^N (p^k \sigma_k)^2 \mathbb{E}\left[\frac{1}{s_t^k}\right]}{(\tau E + E^2)^2}\right)$

Table 1: Convergence bound for different schemes with incomplete devices. The bound for Scheme A assumes there is at least one complete device. While the three schemes have similar performance in the homogeneous setting, only Scheme C guarantees convergence in the heterogeneous setting.

The reason for enlarging the aggregation coefficients in Scheme C can be understood by observing from (2) that increasing p_τ^k 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.

Next we consider inactive devices. In the homogeneous setting, inactive devices do not affect the convergence bound in Theorem 3.1, except that in-activeness should be considered when calculating the expectations of s_t and its functions. In the heterogeneous setting, the convergence rate for Scheme C includes an offset term that equals $O\left(\frac{\sum_{t=0}^{\tau-1} y_t D}{\tau E+E^2}\right)$. Here y_t indicates if there are any inactive devices in round t. This term converges to zero if $y_t < O(\tau)$. In other words, a mild degree of device in-activeness will not impede the convergence.

In practice, due to network or battery outage, a device can frequently become inactive. In this case, permanently removing the device may lead to a more accurate model as the training will be guaranteed to converge. Specifically, we will remove the device if doing so will result in a smaller training loss when the training terminates at the deadline T. The decision entails predicting future probability of in-activeness for the target device, and the participation patterns of other devices. Thus, we assume that devices' participation is invariant with time, and that all other devices are active.

With Scheme C, suppose in each round, a device l becomes inactive with probability $0 \ll y^l < 1$. Let $f_0(\tau)$ be the convergence bound if we keep the device, and $f_1(\tau)$ be the bound if it is removed at τ_0. For f_0, with sufficiently more steps, the second term in (3) shrinks to zero, and the first term converges to $y^l D$, so $f_0 \approx y^l D$. For f_1, as is discussed in Section 3.4, removing the device is equivalent to initiating a fresh start after the objective shift, hence $f_1(\tau) = \frac{\tilde{V}_\tau}{(E(\tau-\tau_0)+\tilde{\gamma})^2}$ for some $\tilde{V}_\tau$ and $\tilde{\gamma}$. The first term on D does not appear in f_1 because other in-activeness were omitted. We thus have the following corollary:

Corollary 4.0.2. *A frequently inactive device l should be removed if*

$$y^l D > f_1(T) \tag{6}$$

If we further assume $V_\tau \approx \tilde{V}_\tau = \tau V$ and $\gamma \approx \tilde{\gamma}$ (so the departing device does not significantly affect the overall SGD variance and the degree of non-IID), then (6) becomes

$$y^l > O\left(\frac{V/D}{TE}\right) \tag{7}$$

Note that TE represents the total number of local epochs a device is expected to run in the training. Thus, the more epochs the training runs, the more sensitive it is to the in-activeness.

4.2 Departure and Arrival: The Time of Change

As is discussed in Section 3.4, when a device leaves, one can decide to either discard this device and shift the objective, or keep it and stick to the old objective. The decision depends on the time at which the device leaves. When keeping the device, the first term in Theorem 3.1's convergence bound (3) will increase and the second term will decrease. Thus, the convergence bound will first decrease then increase, or strictly increase as the time of departure increases. On the other hand, if the device is discarded, 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 objective.

Suppose a device leaves at $\tau_0 < T$. Assume there is no in-activeness and no other departing/arriving devices, and suppose devices' participation is invariant with time. Let $f_0(\tau)$ be the convergence bound if we keep the device, and $f_1(\tau)$ be the bound if it is discarded. We can obtain $f_0(\tau) = \frac{(\tau-\tau_0)D}{\tau E+\gamma} + \frac{V_\tau}{(\tau E+\gamma)^2}$, $f_1(\tau) = \frac{\tilde{V}_\tau}{(E(\tau-\tau_0)+\tilde{\gamma})^2}$. Here $\tilde{M}_\tau, \tilde{V}_\tau, \tilde{\gamma}$ are defined analogously to M_τ, V_τ, γ but they exclude the departing device. A device is discarded if by doing so, a smaller training loss can be obtained at the deadline T, which is concluded in the following corollary:

Corollary 4.0.3. *A leaving device should be discarded if*

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

Further assume $\tilde{V}_\tau = V_\tau = \tau V, \tilde{\gamma} = \gamma$ as in Corollary 4.0.2. (8) can then be written as:

$$T - \tau_0 \geq O\left(\sqrt{\frac{V}{D}\tau_0}\right) \tag{9}$$

From (9), in order to discard the leaving device, the remaining training time $T - \tau_0$ must be at least $O(\sqrt{\tau_0})$. The duration changes with τ_0 because as τ_0 increases, the learning rate without shift gets smaller, mitigating the increase of the training loss from the leaving device.

Since device arrival and departure are symmetric, the same analysis can also be applied to the arriving devices. Thus, we need to guarantee at least $O(\tau_0)$ of training time in order to admit an arriving device. Solving $T - \tau_0 = \sqrt{\frac{V}{D}\tau_0}$, the latest time to admit a device is $O(T - \sqrt{T})$.

5 Conclusion

This paper extends the classic federated learning paradigm to incorporate more flexible device participation patterns. The analysis shows that incomplete local device updates can be utilized by scaling the corresponding aggregation coefficients. Furthermore, a mild degree of device in-activeness will not impact the convergence, but the tolerance to in-activeness diminishes as the training time increases. Finally, the paper provides the number of rounds one must reserve in order to discard a departing device, or admit an arriving device, without impacting model convergence. We plan to empirically validate these results in training on image recognition datasets.

References

Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

Yujing Chen, Yue Ning, and Huzefa Rangwala. Asynchronous online federated learning for edge devices. *arXiv preprint arXiv:1911.02134*, 2019.

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

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *In Proc. of ICLR*, 2020.

Lumin Liu, Jun Zhang, S. H. Song, and Khaled B. Letaief. Client-edge-cloud hierarchical federated learning. `https://arxiv.org/abs/1905.06641`.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-tasking learning. In *In Proc. of NIPS*, 2017.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. `https://arxiv.org/abs/1804.05271`, 2019.

Xiaoxi Zhang, Jianyu Wang, Gauri Joshi, and Carlee Joe-Wong. Machine learning on volatile instances. *arXiv preprint arXiv:2003.05649*, 2020.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning on non-iid data. `https://arxiv.org/abs/1806.00582`, 2018.

A Proof of Theorem 3.1

A.1 Equivalent View

For ease of the analysis, we introduce for each client k and each global round τ 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 $\sum_{i=0}^{E} \alpha^k_{\tau E+i} = s^k_\tau$. If s^k_τ is a random variable, then α^k_t's are also random variables, and the distributions of α^k_t's determine the distribution of s^k_τ. For example, if $\alpha^k_t \overset{iid}{\sim}$ 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 α^k_t's. Our results are thus valid for any realization of s^k_τ.

With the definition of α^k_t's, we can rewrite (1)(2) as:

$$w^k_{\tau E+i+1} = w^k_{\tau E+i} - \eta_\tau g^k_{\tau E+i}\alpha^k_{\tau E+i} \tag{10}$$

$$w^g_{(\tau+1)E} = w^g_{\tau E} - \sum_{k=1}^{N} p^k_\tau \sum_{i=0}^{E} \eta_\tau g^k_{\tau E+i}\alpha^k_{\tau E+i} \tag{11}$$

Note that w^g_t 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

$$\bar{w}_{\tau E+i+1} = \bar{w}_{\tau E+i} - \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{\tau E+i}\alpha_{\tau E+i} \tag{12}$$

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 τ, $\bar{w}_{\tau E} = w^g_{\tau E}$.*

Proof. We will prove by induction. By definition, $\bar{w}_0 = w^g_0$. Suppose $\bar{w}_{\tau E} = w^g_{\tau E}$, then

$$\begin{aligned}
\bar{w}_{(\tau+1)E} &= \bar{w}_{(\tau+1)E-1} - \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{(\tau+1)E-1}\alpha_{(\tau+1)E-1} \\
&= \cdots = \bar{w}_{\tau E} - \sum_{i=0}^{E-1} \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{\tau E+i}\alpha_{\tau E+i} \\
&= w^g_{\tau E} - \sum_{k=1}^{N} p^k_\tau \sum_{i=0}^{E-1} \eta_\tau g^k_{\tau E+i}\alpha^k_{\tau E+i} = w^g_{(\tau+1)E}
\end{aligned} \tag{13}$$

□

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

A.2 Key Lemmas

We first present a couple of important lemmas:

Lemma A.2.

$$\mathbb{E}_\xi \|\sum_{k=1}^{N} p^k_\tau (g^k_t - \bar{g}^k_t)\|^2 \le \sum_{k=1}^{N} (p^k_\tau)^2 \sigma^2_k \tag{14}$$

Proof.

$$\|\sum_{k=1}^{N} p^k_\tau (g^k_t - \bar{g}^k_t)\|^2 = \sum_{k=1}^{N} \|p^k_\tau (g^k_t - \bar{g}^k_t)\|^2 + \sum_{j\neq k} \|p^k_\tau p^j_\tau (g^k_t - \bar{g}^k_t)(g^j_t - \bar{g}^j_t)\|^2 \tag{15}$$

Since each client is running independently, the covariance

$$\mathbb{E}_\xi \|(g_t^k - \bar{g}_t^k)(g_t^j - \bar{g}_t^j)\|^2 = 0 \tag{16}$$

Thus,

$$\mathbb{E}_\xi \| \sum_{k=1}^N p_\tau^k (g_t^k - \bar{g}_t^k)\|^2 = \sum_{k=1}^N \mathbb{E}_\xi \|p_\tau^k (g_t^k - \bar{g}_t^k)\|^2 \le \sum_{k=1}^N (p_\tau^k)^2 \sigma_k^2 \tag{17}$$

□

Lemma A.3.

$$\mathbb{E}_\xi [\sum_{k=1}^N p_\tau^k \|\bar{w}_{\tau E+i} - w_{\tau E+i}^k\|^2] \le (E-1)G^2\eta_\tau^2 \Big(\sum_{k=1}^N p_\tau^k s_\tau^k + (\sum_{k=1}^N p_\tau^k - 2)_+ \sum_{k=1}^N \frac{(p_\tau^k)^2}{p^k} s_\tau^k\Big) \tag{18}$$

Proof. Note that $w_{\tau E}^k = \bar{w}_{\tau E}$ for all k.

$$\begin{aligned} \|\bar{w}_{\tau E+i} - w_{\tau E+i}^k\|^2 &= \|(\bar{w}_{\tau E+i} - \bar{w}_{\tau E}) - (w_{\tau E+i}^k - \bar{w}_{\tau E})\|^2 \\ &= \|\bar{w}_{\tau E+i} - \bar{w}_{\tau E}\|^2 - 2\langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, w_{\tau E+i}^k - \bar{w}_{\tau E}\rangle + \|w_{\tau E+i}^k - \bar{w}_{\tau E}\|^2 \end{aligned} \tag{19}$$

From (10)(12),

$$\begin{aligned} \sum_{k=1}^N p_\tau^k w_{\tau E+i}^k &= \sum_{k=1}^N p_\tau^k w_{\tau E+i-1}^k - \eta_\tau \sum_{k=1}^N p_\tau^k g_{\tau E+i-1}^k \alpha_{\tau E+i-1}^k \\ &= \sum_{k=1}^N p_\tau^k w_{\tau E+i-1}^k + \bar{w}_{\tau E+i} - \bar{w}_{\tau E+i-1} \\ &= \cdots = \sum_{k=1}^N p_\tau^k w_{\tau E}^k + \bar{w}_{\tau E+i} - \bar{w}_{\tau E} \end{aligned} \tag{20}$$

Thus,

$$\begin{aligned} &-2\sum_{k=1}^N p_\tau^k \langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, w_{\tau E+i}^k - \bar{w}_{\tau E}\rangle \\ =&-2\langle \bar{w}_{\tau E+i} - \bar{w}_{\tau E}, \sum_{k=1}^N p_\tau^k w_{\tau E}^k + \bar{w}_{\tau E+i} - \bar{w}_{\tau E} - \sum_{k=1}^N p_\tau^k \bar{w}_{\tau E}\rangle \\ =&-2\|\bar{w}_{\tau E+i} - \bar{w}_{\tau E}\|^2 \end{aligned} \tag{21}$$

$$\sum_{k=1}^N p_\tau^k \|\bar{w}_{\tau E+i} - w_{\tau E+i}^k\|^2 = (\sum_{k=1}^N p_\tau^k - 2)\|\bar{w}_{\tau E+i} - \bar{w}_{\tau E}\|^2 + \sum_{k=1}^N p_\tau^k \|w_{\tau E+i}^k - \bar{w}_{\tau E}\|^2 \tag{22}$$

$$\begin{aligned} \|\bar{w}_{\tau E+i} - \bar{w}_{\tau E}\|^2 &= \|\sum_{j=0}^{i-1} \eta_\tau \sum_{k=1}^N p_\tau^k g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 \\ =&\|\eta_\tau \sum_{k=1}^N p_\tau^k \Big(\sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k\Big)\|^2 = \eta_\tau^2 \|\sum_{k=1}^N p^k \Big(\frac{p_\tau^k}{p^k} \sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k\Big)\|^2 \\ \le&\eta_\tau^2 \sum_{k=1}^N \frac{(p_\tau^k)^2}{p^k} \|\sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 \end{aligned} \tag{23}$$

Here

$$
\begin{aligned}
\|\sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 &= \sum_{j=0}^{i-1} \|g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 + 2\sum_{p<q} \langle g_{\tau E+p}^k \alpha_{\tau E+p}^k, g_{\tau E+q}^k \alpha_{\tau E+q}^k \rangle \\
&\leq \sum_{j=0}^{i-1} \|g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 + 2\sum_{p<q} \|g_{\tau E+p}^k \alpha_{\tau E+p}^k\| \|g_{\tau E+q}^k \alpha_{\tau E+q}^k\| \\
&\leq \sum_{j=0}^{i-1} \|g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 + \sum_{p<q} \Big(\|g_{\tau E+p}^k \alpha_{\tau E+p}^k\|^2 + \|g_{\tau E+q}^k \alpha_{\tau E+q}^k\|^2 \Big) \\
&= i \sum_{j=0}^{i-1} \|g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2
\end{aligned} \tag{24}
$$

So

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

Plug (25) to (23) we have

$$
\mathbb{E}_\xi \|\bar{w}_{\tau E+i} - \bar{w}_{\tau E}\|^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^N \frac{(p_\tau^k)^2}{p^k} s_\tau^k \tag{26}
$$

Similarly

$$
\mathbb{E}_\xi \sum_{k=1}^N p_\tau^k \|w_{\tau E+i}^k - \bar{w}_{\tau E}\|^2 = \mathbb{E}_\xi \sum_{k=1}^N p_\tau^k \|\eta_\tau \sum_{j=0}^{i-1} g_{\tau E+j}^k \alpha_{\tau E+j}^k\|^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^N p_\tau^k s_\tau^k \tag{27}
$$

Plug (26)(27) to (22) we have

$$
\mathbb{E}_\xi [\sum_{k=1}^N p_\tau^k \|\bar{w}_{\tau E+i} - w_{\tau E+i}^k\|^2] \leq (E-1)G^2 \eta_\tau^2 \Big(\sum_{k=1}^N p_\tau^k s_\tau^k + (\sum_{k=1}^N p_\tau^k - 2)_+ \sum_{k=1}^N \frac{(p_\tau^k)^2}{p^k} s_\tau^k \Big) \tag{28}
$$

□

A.3 Bounding $\|\bar{w}_{\tau E+i+1} - w^*\|^2$

$$
\begin{aligned}
\|\bar{w}_{\tau E+i+1} - w^*\|^2 &= \|\bar{w}_{\tau E+i} - \eta_\tau \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k g_{\tau E+i}^k - w^* - \eta_\tau \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k \bar{g}_{\tau E+i}^k + \eta_\tau \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k \bar{g}_{\tau E+i}^k \|^2 \\
&= \underbrace{\|\bar{w}_{\tau E+i} - w^* - \eta_\tau \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k \bar{g}_{\tau E+i}^k\|^2}_{A_1} + \eta_\tau^2 \|\sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k (\bar{g}_{\tau E+i}^k - g_{\tau E+i}^k)\|^2 \\
&\quad \underbrace{+ 2\eta_\tau \langle \bar{w}_{\tau E+i} - w^* - \eta_\tau \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k \bar{g}_{\tau E+i}^k, \sum_{k=1}^N p_\tau^k \alpha_{\tau E+i}^k (\bar{g}_{\tau E+i}^k - g_{\tau E+i}^k) \rangle}_{A_2}
\end{aligned} \tag{29}
$$

Since $\mathbb{E}_\xi[g^k_{\tau E+i}] = \bar{g}^k_{\tau E+i}$, we have $\mathbb{E}_\xi[A_2] = 0$. We then bound A_1.

$$
\begin{aligned}
A_1 &= \|\bar{w}_{\tau E+i} - w^* - \eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\|^2 \\
&= \|\bar{w}_{\tau E+i} - w^*\|^2 \underbrace{-2\eta_t \langle \bar{w}_{\tau E+i} - w^*, \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle}_{B_1} + \underbrace{\eta_t^2 \|\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\|^2}_{B_2}
\end{aligned}
\tag{30}
$$

Since F_k is L-smooth,

$$
\|\alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\|^2 \leq 2L(F_k(w^k_{\tau E+i}) - F^*_k)\alpha^k_{\tau E+i} \tag{31}
$$

By the convexity of l_2 norm

$$
\begin{aligned}
B_2 &= \eta_\tau^2 \|\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\|^2 = \eta_\tau^2 \|\sum_{k=1}^N p^k (\frac{p^k_\tau}{p^k} \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i})\|^2 \\
&\leq \eta_\tau^2 \sum_{k=1}^N \frac{(p^k_\tau)^2}{p^k} \|\alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\|^2 \leq 2L\theta \eta_\tau^2 \sum_{k=1}^N p^k_\tau (F_k(w^k_{\tau E+i}) - F^*_k)\alpha^k_{\tau E+i}
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
B_1 &= -2\eta_\tau \langle \bar{w}_{\tau E+i} - w^*, \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle = -2\eta_\tau \sum_{k=1}^N p^k_\tau \langle \bar{w}_{\tau E+i} - w^*, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle \\
&= -2\eta_\tau \sum_{k=1}^N p^k_\tau \langle \bar{w}_{\tau E+i} - w^k_{\tau E+i}, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle - 2\eta_\tau \sum_{k=1}^N p_k \langle w^k_{\tau E+i} - w^*, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle
\end{aligned}
\tag{33}
$$

Here

$$
\begin{aligned}
&-2\langle \bar{w}_{\tau E+i} - w^k_{\tau E+i}, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle \leq 2|\langle \bar{w}_{\tau E+i} - w^k_{\tau E+i}, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle| \\
\leq& 2\alpha^k_{\tau E+i} \|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\| \|\bar{g}^k_{\tau E+i}\| \leq \Big(\frac{1}{\eta_\tau} \|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2 + \eta_\tau \|\bar{g}^k_{\tau E+i}\|^2\Big)\alpha^k_{\tau E+i}
\end{aligned}
\tag{34}
$$

Since F_k is μ-strong convex

$$
\langle w^k_{\tau E+i} - w^*, \alpha^k_{\tau E+i} \bar{g}^k_{\tau E+i}\rangle \geq \big((F_k(w^k_{\tau E+i}) - F_k(w^*)) + \frac{\mu}{2}\|w^k_{\tau E+i} - w^*\|^2\big)\alpha^k_{\tau E+i} \tag{35}
$$

Plug (34)(35) to (33)

$$
B_1 \leq \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \Bigg(\|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2 + \eta_\tau^2 \|\bar{g}^k_{\tau E+i}\|^2 - 2\eta_\tau \big((F_k(w^k_{\tau E+i}) - F_k(w^*)) + \frac{\mu}{2}\|w^k_{\tau E+i} - w^*\|^2\big) \Bigg) \tag{36}
$$

Plug (32)(36) to (30)

$$
\begin{aligned}
A_1 \leq& \|\bar{w}_{\tau E+i} - w^*\|^2 + 2L\theta\eta_\tau^2 \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} (F_k(w^k_{\tau E+i}) - F^*_k) \\
&+ \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \Bigg(\|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2 + \underbrace{\eta_\tau^2 \|\bar{g}^k_{\tau E+i}\|^2}_{\leq 2\eta_\tau^2 L(F_k(w^k_{\tau E+i}) - F^*_k)} - 2\eta_\tau \big((F_k(w^k_{\tau E+i}) - F_k(w^*)) + \frac{\mu}{2}\|w^k_{\tau E+i} - w^*\|^2\big) \Bigg) \\
\leq& \|\bar{w}_{\tau E+i} - w^*\|^2 - \mu\eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \|w^k_{\tau E+i} - w^*\|^2 + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} \|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2 \\
&+ \underbrace{2(1+\theta)L\eta_\tau^2 \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} (F_k(w^k_{\tau E+i}) - F^*_k) - 2\eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i} (F_k(w^k_{\tau E+i}) - F_k(w^*))}_{C}
\end{aligned}
\tag{37}
$$

$$
\begin{aligned}
\|w^k_{\tau E+i} - w^*\|^2 &= \|w^k_{\tau E+i} - \bar{w}_{\tau E+i} + \bar{w}_{\tau E+i} - w^*\|^2 \\
&= \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 + 2\langle w^k_{\tau E+i} - \bar{w}_{\tau E+i}, \bar{w}_{\tau E+i} - w^*\rangle \\
&\geq \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 - 2\|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|\|\bar{w}_{\tau E+i} - w^*\| \\
&\geq \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2 + \|\bar{w}_{\tau E+i} - w^*\|^2 - (2\|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2 + \frac{1}{2}\|\bar{w}_{\tau E+i} - w^*\|^2) \\
&= \frac{1}{2}\|\bar{w}_{\tau E+i} - w^*\|^2 - \|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2
\end{aligned} \tag{38}
$$

Thus,

$$
A_1 \leq (1 - \frac{1}{2}\mu\eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i})\|\bar{w}_{\tau E+i} - w^*\|^2 + (1+\mu\eta_\tau)\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}\|\bar{w}_{\tau E+i} - w^k_{\tau E+i}\|^2 + C \tag{39}
$$

Let $\gamma_\tau = 2\eta_\tau(1-(1+\theta)L\eta_\tau)$. Assume $\eta_\tau \leq \frac{1}{2(1+\theta)L}$, hence $\eta_\tau \leq \gamma_\tau \leq 2\eta_\tau$.

$$
\begin{aligned}
C &= -2\eta_\tau(1-(1+\theta)L\eta_\tau)\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F^*_k) + 2\eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^*) - F^*_k) \\
&= -\gamma_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F^*_k + F_k(w^*) - F_k(w^*)) + 2\eta_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^*) - F^*_k) \\
&= -\gamma_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F_k(w^*)) + (2\eta_\tau - \gamma_\tau)\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^*) - F^*_k) \\
&\leq \underbrace{-\gamma_\tau \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F_k(w^*))}_{D} + 2(1+\theta)L\eta_\tau^2 \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}\Gamma_k
\end{aligned} \tag{40}
$$

Next we bound D

$$
\begin{aligned}
&\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F_k(w^*)) = \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(w^k_{\tau E+i}) - F_k(\bar{w}_{\tau E+i})) + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) \\
&\geq \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}\langle \nabla F_k(\bar{w}_{\tau E+i}), w^k_{\tau E+i} - \bar{w}_{\tau E+i}\rangle + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) \\
&\geq -\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}\|\nabla F_k(\bar{w}_{\tau E+i})\|\|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\| + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) \\
&\geq -\frac{1}{2}\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(\eta_\tau \underbrace{\|\nabla F_k(\bar{w}_{\tau E+i})\|^2}_{\leq 2L(F_k(\bar{w}_{\tau E+i}) - F^*_k)} + \frac{1}{\eta_\tau}\|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2) + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) \\
&\geq -\sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}\Big(\eta_\tau L(F_k(\bar{w}_{\tau E+i}) - F^*_k) + \frac{1}{2\eta_\tau}\|w^k_{\tau E+i} - \bar{w}_{\tau E+i}\|^2\Big) + \sum_{k=1}^N p^k_\tau \alpha^k_{\tau E+i}(F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
\end{aligned} \tag{41}
$$

Thus,

$$
\begin{aligned}
C &\leq \gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (\eta_\tau L \underbrace{(F_k(\bar{w}_{\tau E+i}) - F_k^*)}_{F_k(\bar{w}_{\tau E+i}) - F_k(w^*) + F_k(w^*) - F_k^*} + \frac{1}{2\eta_\tau} \| w_{\tau E+i}^k - \bar{w}_{\tau E+i} \|^2) \\
&\quad - \gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) + 2(1+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i} \Gamma_k \\
&= \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) + \underbrace{\frac{\gamma_\tau}{2\eta_\tau}}_{\leq 1} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| w_{\tau E+i}^k - \bar{w}_{\tau E+i} \|^2 \\
&\quad + 2(1+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i} \Gamma_k + \underbrace{\gamma_\tau}_{\leq 2\eta_\tau} \eta_\tau L \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \Gamma_k \\
&\leq \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (F_k(\bar{w}_{\tau E+i}) - F_k(w^*)) + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| w_{\tau E+i}^k - \bar{w}_{\tau E+i} \|^2 \\
&\quad + 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \Gamma_k
\end{aligned}
\tag{42}
$$

Plug (42) to (39) we have

$$
\begin{aligned}
A_1 &\leq \| \bar{w}_{\tau E+i} - w^* \|^2 - \mu \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| w_{\tau E+i}^k - w^* \|^2 + 2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| \bar{w}_{\tau E+i} - w_{\tau E+i}^k \|^2 \\
&\quad + 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \Gamma_k + \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (F_k(\bar{w}_{\tau E+i}) - F_k(w^*))
\end{aligned}
\tag{43}
$$

Plug (43) to (29),

$$
\begin{aligned}
&\| \bar{w}_{\tau E+i+1} - w^* \|^2 \leq (1 - \frac{1}{2} \mu \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k) \| \bar{w}_{\tau E+i} - w^* \|^2 \\
&+ \eta_\tau^2 \| \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (\bar{g}_{\tau E+i}^k - g_{\tau E+i}^k) \|^2 + \underbrace{(2 + \mu \eta_\tau)}_{\leq 2 + \frac{\mu}{2(1+\theta)L}} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| \bar{w}_{\tau E+i} - w_{\tau E+i}^k \|^2 \\
&+ 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \Gamma_k + \underbrace{\gamma_\tau (1 - \eta_\tau L)}_{\leq 2\eta_\tau} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (F_k(w^*) - F_k(\bar{w}_{\tau E+i}))
\end{aligned}
\tag{44}
$$

Define

$$
\begin{aligned}
B_{\tau E+i} = (2 + \frac{\mu}{2(1+\theta)L}) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \| \bar{w}_{\tau E+i} - w_{\tau E+i}^k \|^2 + \| \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k (\bar{g}_{\tau E+i}^k - g_{\tau E+i}^k) \|^2 \\
+ 2(2+\theta) L \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k \Gamma_k
\end{aligned}
\tag{45}
$$

Thus,

$$\begin{aligned}&\|\bar{w}_{\tau E+i+1}-w^*\|^2 \le (1-\frac{1}{2}\mu\eta_\tau\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^k)\|\bar{w}_{\tau E+i}-w^*\|^2+\eta_\tau^2 B_{\tau E+i}\\ &+2\eta_\tau\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^k(F_k(w^*)-F_k(\bar{w}_{\tau E+i}))\end{aligned}\tag{46}$$

Apply the lemmas we have

$$\begin{aligned}\mathbb{E}_\xi[B_{\tau E+i}] &\le \sum_{k=1}^{N}(p_\tau^k)^2\alpha_{\tau E+i}^k\sigma_k^2+2(2+\theta)L\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^k\Gamma_k\\ &+(2+\frac{\mu}{2(1+\theta)L})(E-1)G^2\Big(\sum_{k=1}^{N}p_\tau^k s_\tau^k+(\sum_{k=1}^{N}p_\tau^k-2)_+\sum_{k=1}^{N}\frac{(p_\tau^k)^2}{p^k}s_\tau^k\Big)\end{aligned}\tag{47}$$

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.4 Bounding $\|\bar{w}_{\tau E}-w^*\|$

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

$$\sum_{i=1}^{E}\Delta_{\tau E+i}\le\sum_{i=0}^{E-1}(1-\frac{1}{2}\mu\eta_\tau\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^k)\Delta_{\tau E+i}+\eta_\tau^2B_\tau+2\eta_\tau\sum_{k=1}^{N}p_\tau^k s_\tau^k(F_k(w^*)-F_k(\bar{w}_{\tau E+l}))\tag{48}$$

where $B_\tau=\sum_{i=0}^{E-1}B_{\tau E+i}$, and $\bar{w}_{\tau E+l}=\text{argmin}_{\bar{w}_{\tau E+i}}\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^kF_k(\bar{w}_{\tau E+i})$.

Reorganize it we can get

$$\Delta_{(\tau+1)E}\le\Delta_{\tau E}-\frac{1}{2}\mu\eta_\tau\sum_{i=0}^{E-1}\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^k\Delta_{\tau E+i}+\eta_\tau^2B_\tau+2\eta_\tau\sum_{k=1}^{N}p_\tau^k s_\tau^k(F_k(w^*)-F_k(\bar{w}_{\tau E+l}))\tag{49}$$

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

$$\begin{aligned}\sqrt{\Delta_{\tau E+i+1}}&=\|\bar{w}_{\tau E+i+1}-w^*\|=\|\bar{w}_{\tau E+i+1}-\bar{w}_{\tau E+i}+\bar{w}_{\tau E+i}-w^*\|\\ &\le\|\bar{w}_{\tau E+i+1}-\bar{w}_{\tau E+i}\|+\sqrt{\Delta_{\tau E+i}}\\ &=\|\eta_\tau\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^kg_{\tau E+i}^k\|+\sqrt{\Delta_{\tau E+i}}\end{aligned}\tag{50}$$

Define $h_{\tau E+i}=\|\sum_{k=1}^{N}p_\tau^k\alpha_{\tau E+i}^kg_{\tau E+i}^k\|$.

Thus,

$$\begin{aligned}\sqrt{\Delta_{(\tau+1)E}}&\le\sqrt{\Delta_{(\tau+1)E-1}}+\eta_\tau h_{(\tau+1)E-1}\\ &\le\cdots\le\sqrt{\Delta_{\tau E+i}}+\sum_{j=i}^{E-1}\eta_\tau h_{\tau E+j}\end{aligned}\tag{51}$$

$$\begin{aligned}\Delta_{(\tau+1)E}&\le\Delta_{\tau E+i}+2\sqrt{\Delta_{\tau E+i}}(\sum_{j=i}^{E-1}\eta_\tau h_{\tau E+j})+(\sum_{j=i}^{E-1}\eta_\tau h_{\tau E+j})^2\\ &\le2\Delta_{\tau E+i}+2(\sum_{j=i}^{E-1}\eta_\tau h_{\tau E+j})^2\end{aligned}\tag{52}$$

$$\Delta_{\tau E+i} \geq \frac{1}{2}\Delta_{(\tau+1)E} - (\sum_{j=i}^{E-1} \eta_\tau h_{\tau E+j})^2 \geq \frac{1}{2}\Delta_{(\tau+1)E} - (\sum_{j=0}^{E-1} \eta_\tau h_{\tau E+j})^2 \tag{53}$$

Plug (53) to (49) we can get

$$\begin{aligned}(1+\frac{1}{4}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k)\Delta_{(\tau+1)E} \leq \Delta_{\tau E} + \frac{1}{2}\mu\eta_\tau^3 \sum_{k=1}^{N} p_\tau^k s_\tau^k (\sum_{i=0}^{E-1} h_{\tau E+i})^2 + \eta_\tau^2 B_\tau \\ + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k (F_k(w^*) - F_k(\bar{w}_{\tau E+l}))\end{aligned} \tag{54}$$

Apply Lemma A.2, Lemma A.3 and Assumption 3.4, we have

$$\begin{aligned}\mathbb{E}_\xi[h_{\tau E+i}^2] &= \mathbb{E}_\xi \| \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+i}^k g_{\tau E+i}^k \|^2 \\ &\leq \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} \mathbb{E}_\xi \|\alpha_{\tau E+i}^k g_{\tau E+i}^k\|^2 \leq \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} G^2 \alpha_{\tau E+i}^k\end{aligned} \tag{55}$$

$$\mathbb{E}_\xi[(\sum_{i=0}^{E-1} h_{\tau E+i})^2] \leq \mathbb{E}_\xi[E\sum_{i=0}^{E-1} h_{\tau E+i}^2] \leq EG^2 \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} s_\tau^k \tag{56}$$

$$\begin{aligned}\mathbb{E}_\xi[B_\tau] = \sum_{i=0}^{E-1} \mathbb{E}_\xi[B_{\tau E+i}] = \sum_{k=1}^{N} (p_\tau^k)^2 s_\tau^k \sigma_k^2 + 2(2+\theta)L \sum_{k=1}^{N} p_\tau^k s_\tau^k \Gamma_k \\ + (2 + \frac{\mu}{2(1+\theta)L})E(E-1)G^2 \Big(\sum_{k=1}^{N} p_\tau^k s_\tau^k + \theta(\sum_{k=1}^{N} p_\tau^k - 2)_+ \sum_{k=1}^{N} p_\tau^k s_\tau^k \Big)\end{aligned} \tag{57}$$

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

$$\begin{aligned}(1+\frac{1}{4}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k)\bar{\Delta}_{(\tau+1)E} \leq \bar{\Delta}_{\tau E} + \frac{1}{2}\mu\eta_\tau^3 \sum_{k=1}^{N} p_\tau^k s_\tau^k \bar{H}_\tau + \eta_\tau^2 \bar{B}_\tau \\ + 2\eta_\tau \mathbb{E}_\xi \sum_{k=1}^{N} p_\tau^k s_\tau^k (F_k(w^*) - F_k(\bar{w}_{\tau E+l}))\end{aligned} \tag{58}$$

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

$$\begin{aligned}\sum_{k=1}^{N} p_\tau^k s_\tau^k (F_k(w^*) - F_k(\bar{w}_{\tau E+l})) &= \sum_{k=1}^{N} p_\tau^k s_\tau^k (\underbrace{F_k(w^*) - F_k^*}_{\Gamma_k} + \underbrace{F_k^* - F_k(\bar{w}_{\tau E+l})}_{\leq 0}) \\ &\leq \sum_{k=1}^{N} p_\tau^k s_\tau^k \Gamma_k\end{aligned} \tag{59}$$

Put it together

$$\sum_{k=1}^{N} p_\tau^k s_\tau^k (F_k(w^*) - F_k(\bar{w}_{\tau E+l})) \leq z_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k \Gamma_k \tag{60}$$

Assume $\eta_\tau \leq \frac{4}{\mu E\theta} \leq \frac{4}{\mu \sum_{k=1}^N p_\tau^k s_\tau^k}$, divide both sides with $1 + \frac{1}{4}\mu\eta_\tau \sum_{k=1}^N p_\tau^k s_\tau^k$ in (58) we can get

$$\begin{aligned}
\bar{\Delta}_{(\tau+1)E} &\leq \Big(1 - \frac{\frac{1}{4}\mu\eta_\tau \sum_{k=1}^N p_\tau^k s_\tau^k}{1 + \frac{1}{4}\mu\eta_\tau \sum_{k=1}^N p_\tau^k s_\tau^k}\Big)\bar{\Delta}_{\tau E} + 2\eta_\tau^2 \bar{H}_\tau + \eta_\tau^2 \bar{B}_\tau \\
&+ 2\eta_\tau z_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k \Gamma_k \\
&\leq \Big(1 - \frac{1}{8}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k\Big)\bar{\Delta}_{\tau E} + \eta_\tau^2(\bar{B}_\tau + 2\bar{H}_\tau) + 2\eta_\tau z_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k \Gamma_k
\end{aligned} \tag{61}$$

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

$$\mathbb{E}[\bar{\Delta}_{(\tau+1)E}] \leq \Big(1 - \frac{1}{8}\mu\eta_\tau \mathbb{E}[\sum_{k=1}^{N} p_\tau^k s_\tau^k]\Big)\bar{\Delta}_{\tau E} + \eta_\tau^2 \mathbb{E}[\bar{B}_\tau + 2\bar{H}_\tau] + 2\eta_\tau z_\tau \sum_{k=1}^{N} \mathbb{E}[p_\tau^k s_\tau^k]\Gamma_k \tag{62}$$

A.5 Proof of Theorem 3.1

We prove by induction. Let $\eta_\tau = \frac{8}{\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]} \frac{2E}{(\tau+1)E+\gamma}$. Initially, $\frac{V_0}{\gamma^2} \geq \mathbb{E}[\bar{\Delta}_0]$. Suppose $\mathbb{E}[\bar{\Delta}_{\tau E}] \leq \frac{M_\tau D}{\tau E+\gamma} + \frac{V_\tau}{(\tau E+\gamma)^2}$, then

$$\begin{aligned}
\mathbb{E}[\bar{\Delta}_{(\tau+1)E}] &\leq \frac{\tau E+\gamma-E}{(\tau+1)E+\gamma}\left(\frac{M_\tau D}{\tau E+\gamma} + \frac{V_\tau}{(\tau E+\gamma)^2}\right) + \frac{(16E)^2 \mathbb{E}[\bar{B}_\tau + 2\bar{H}_\tau]}{(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2 \left((\tau+1)E+\gamma\right)^2} + \frac{z_\tau D}{(\tau+1)E+\gamma} \\
\leq& \frac{(\tau E+\gamma-E)M_\tau D}{(\tau E+\gamma)^2 - E^2} + \frac{\tau E+\gamma-E}{(\tau E+\gamma)^2 - E^2}\frac{V_\tau}{(\tau+1)E+\gamma} + \frac{(16E)^2 \mathbb{E}[\bar{B}_\tau + 2\bar{H}_\tau]}{(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2 \left((\tau+1)E+\gamma\right)^2} + \frac{z_\tau D}{(\tau+1)E+\gamma} \\
\leq& \frac{M_\tau D}{(\tau+1)E+\gamma} + \frac{V_\tau}{\left((\tau+1)E+\gamma\right)^2} + \frac{(16E)^2 \mathbb{E}[\bar{B}_\tau + 2\bar{H}_\tau]}{(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2 \left((\tau+1)E+\gamma\right)^2} + \frac{z_\tau D}{(\tau+1)E+\gamma} \\
=& \frac{M_{\tau+1} D}{(\tau+1)E+\gamma} + \frac{V_{\tau+1}}{\left((\tau+1)E+\gamma\right)^2}
\end{aligned} \tag{63}$$

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

We can check it satisfies the previous assumptions regarding η_τ:

$$\begin{aligned}
\eta_\tau \leq \eta_0 =& \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])}{E+\gamma} \\
\leq& \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])}{32E(1+\theta)L/(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])} = \frac{1}{2(1+\theta)L}
\end{aligned} \tag{64}$$

$$\eta_\tau \leq \eta_0 = \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])}{E+\gamma} \leq \frac{16E/(\mu \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])}{4E^2\theta/(\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])} = \frac{4}{\mu E\theta} \tag{65}$$

B Proof of Theorem 3.2

- Departure Case: $\tilde{n} = n - n_l$

$$\|\tilde{w}^* - w^*\| \le \frac{2}{\mu}\|\nabla F(\tilde{w}^*)\| = \frac{2}{\mu}\left\|\nabla F(\tilde{w}^*) - \underbrace{\nabla \tilde{F}(\tilde{w}^*)}_{=0}\right\|$$

$$= \frac{2}{\mu}\left\|\sum_{k\neq l}(p^k - \tilde{p}^k)\nabla F_k(\tilde{w}^*) + p^l \nabla F_l(\tilde{w}^*)\right\|$$

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

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

$$= \frac{2}{\mu}\left\|-p^l \underbrace{\sum_{k\neq l}\tilde{p}^k \nabla F_k(\tilde{w}^*)}_{=\nabla \tilde{F}(\tilde{w}^*)=0} + p^l \nabla F_l(\tilde{w}^*)\right\|$$

$$= \frac{2p^l}{\mu}\|\nabla F_l(\tilde{w}^*)\| \le \frac{2p^l}{\mu}\sqrt{2L\left(F_l(\tilde{w}^*) - F_l^*\right)} = \frac{2\sqrt{2L}}{\mu}p^l\sqrt{\tilde{\Gamma}_l}$$

- Arrival Case: $\tilde{n} = n + n_l$

$$\|\tilde{w}^* - w^*\| = \|w^* - \tilde{w}^*\| \le \frac{2}{\mu}\|\nabla \tilde{F}(w^*)\| = \frac{2}{\mu}\left\|\nabla \tilde{F}(w^*) - \underbrace{\nabla F(w^*)}_{=0}\right\|$$

$$= \cdots = \frac{2}{\mu}\left\|-\tilde{p}^l \underbrace{\sum_{k\neq l}p^k \nabla F_k(w^*)}_{=\nabla F(w^*)=0} + \tilde{p}^l \nabla F_l(w^*)\right\|$$

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