A Additional Notations and Bounds for Sampling Schemes

In this section, we introduce additional notations that are used throughout the proofs. Following common practice, e.g. Stich (2019); Li et al. (2020b), we define two virtual sequences $\overline{\mathbf{v}}_t = \sum_{k=1}^N p_k \mathbf{v}_t^k$ and $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{w}_t^k$, where we recall the FedAvg updates from (2):

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \alpha_t \mathbf{g}_{t,k}, \quad \mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k & \text{if } t+1 \notin \mathcal{I}_E, \\ \sum_{k \in \mathcal{S}_{t+1}} q_k \mathbf{v}_{t+1}^k & \text{if } t+1 \in \mathcal{I}_E. \end{cases}$$

The following observations apply to FedAvg updates, while Nesterov accelerated FedAvg requires modifications. For full device participation or partial participation with $t \notin \mathcal{I}_E$, note that $\overline{\mathbf{v}}_t = \overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{v}_t^k$. For partial participation with $t \in \mathcal{I}_E$, $\overline{\mathbf{w}}_t \neq \overline{\mathbf{v}}_t$ since $\overline{\mathbf{v}}_t = \sum_{k=1}^N p_k \mathbf{v}_t^k$ while $\overline{\mathbf{w}}_t = \sum_{k \in \mathcal{S}_t} q_k \mathbf{w}_t^k$. However, we can use unbiased sampling strategies such that $\mathbb{E}_{\mathcal{S}_t} \overline{\mathbf{w}}_t = \overline{\mathbf{v}}_t$. Note that $\overline{\mathbf{v}}_{t+1}$ is one-step SGD from $\overline{\mathbf{w}}_t$.

$$\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_t - \alpha_t \mathbf{g}_t, \tag{3}$$

where $\mathbf{g}_t = \sum_{k=1}^N p_k \mathbf{g}_{t,k}$ is the one-step stochastic gradient averaged over all devices.

$$\mathbf{g}_{t,k} = \nabla F_k \left(\mathbf{w}_t^k, \xi_t^k \right),$$

Similarly, we denote the expected one-step gradient $\overline{\mathbf{g}}_t = \mathbb{E}_{\xi_t}[\mathbf{g}_t] = \sum_{k=1}^N p_k \mathbb{E}_{\xi_t^k} \mathbf{g}_{t,k}$, where

$$\mathbb{E}_{\xi_t^k} \mathbf{g}_{t,k} = \nabla F_k \left(\mathbf{w}_t^k \right), \tag{4}$$

and $\xi_t = \{\xi_t^k\}_{k=1}^N$ denotes random samples at all devices at time step t.
Since in this work we also consider the case of partial participation, the sampling strategy to approximate the system heterogeneity can also affect the convergence. Here we follow the prior works Li et al. (2020b) and Li et al. (2020a) and consider two types of sampling schemes that guarantee $\mathbb{E}_{\mathcal{S}_t} \overline{\mathbf{w}}_t = \overline{\mathbf{v}}_t$. The sampling scheme I establishes $\mathcal{S}_{t+1}$ by *i.i.d.* sampling the devices according to probabilities p_k with replacement, and setting $q_k = \frac{1}{K}$. In this case the upper bound of expected square norm of $\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}$ is given by (Li et al., 2020b, Lemma 5):

$$\mathbb{E}_{\mathcal{S}_{t+1}} \left\| \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} \right\|^2 \leq \frac{4}{K} \alpha_t^2 E^2 G^2. \tag{5}$$

The sampling scheme II establishes $\mathcal{S}_{t+1}$ by uniformly sampling all devices without replacement and setting $q_k = p_k \frac{N}{K}$, in which case we have

$$\mathbb{E}_{\mathcal{S}_{t+1}} \left\| \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} \right\|^2 \leq \frac{4(N-K)}{K(N-1)} \alpha_t^2 E^2 G^2. \tag{6}$$

We summarize these upper bounds as follows:

$$\mathbb{E}_{\mathcal{S}_{t+1}} \left\| \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} \right\|^2 \leq \frac{4}{K} \alpha_t^2 E^2 G^2. \tag{7}$$

and this bound will be used in the convergence proof of the partial participation result.

B Comparison of Convergence Rates with Related Works

In this section, we compare our convergence rate with the best-known results in the literature (see Table 2). In Haddadpour & Mahdavi (2019), the authors provide $\mathcal{O}(1/NT)$ convergence rate of non-convex problems under Polyak-Łojasiewicz (PL) condition, which means their results can directly apply to the strongly convex problems. However, their assumption is based on bounded gradient diversity, defined as follows:

$$\Lambda(\mathbf{w}) = \frac{\sum_k p_k \|\nabla F_k(\mathbf{w})\|_2^2}{\|\sum_k p_k \nabla F_k(\mathbf{w})\|_2^2} \leq B$$

This is a more restrictive assumption comparing to assuming bounded gradient under the case of target accuracy $\epsilon \to 0$ and PL condition. To see this, consider the gradient diversity at the global optimal

$\mathbf{w}^*$, i.e., $\Lambda(\mathbf{w}^*) = \frac{\sum_k p_k \|\nabla F_k(\mathbf{w})\|_2^2}{\|\sum_k p_k \nabla F_k(\mathbf{w})\|_2^2}$. For $\Lambda(\mathbf{w}^*)$ to be bounded, it requires $\|\nabla F_k(\mathbf{w}^*)\|_2^2 = 0, \forall\, k$. This indicates $\mathbf{w}^*$ is also the minimizer of each local objective, which contradicts to the practical setting of heterogeneous data. Therefore, their bound is not effective for arbitrary small ϵ-accuracy under general heterogeneous data while our convergence results still hold in this case.

In Karimireddy et al. (2019), the linear speedup convergence rate of FedAvg are provided for strongly convex, general convex, and non-convex problems under full participation setting. However, their rate does not enjoy linear speedup for any number of devices while our results apply to any valid $K \leq N$. For example, they provides an optimality gap of $\mathcal{O}\left((1 - \frac{K}{N})E/T\right)$ for the strongly convex case (Karimireddy et al., 2019, Theorem V). With partial participation, and when $K = O(1)$, their convergence rate is $\mathcal{O}(E/T)$ which does not have linear speedup. Under partial participation, the FedAvg analyses in Karimireddy et al. (2019) requires $E = \mathcal{O}(1)$. For example, the strongly convex result $\mathcal{O}((1 - \frac{K}{N})E/T)$ in Theorem V is $\mathcal{O}(E/T)$ when $K = O(1)$ and is $\mathcal{O}(E/NT)$ when $K = \mathcal{O}(N)$. In either case, to achieve a $\mathcal{O}(1/T)$ convergence rate, it requires $E = \mathcal{O}(1)$ as well. Similar conclusion also holds for the general convex problem.

Reference	Convergence rate	E	NonIID	Participation	Extra Assumptions	Setting
FedAvgLi et al. (2020b)	$\mathcal{O}(\frac{E^2}{T})$	$\mathcal{O}(1)$	✓	Partial	Bounded gradient	Strongly convex
FedAvgHaddadpour & Mahdavi (2019)	$\mathcal{O}(\frac{1}{KT})$	$\mathcal{O}(K^{-1/3}T^{2/3})^\dagger$	✓‡‡	Partial	Bounded gradient diversity	Strongly convex§
FedAvgKoloskova et al. (2020)	$\mathcal{O}(\frac{1}{NT})$	$\mathcal{O}(N^{-1/2}T^{1/2})$	✓	Full	Bounded gradient	Strongly convex
FedAvgKarimireddy et al. (2019)	$\mathcal{O}(\frac{1}{NT})^{\dagger\dagger}$	$\mathcal{O}(N^{-1/2}T^{1/2})^{\dagger\dagger}$	✓	Partial	Bounded gradient dissimilarity	Strongly convex
FedAvg/N-FedAvg (our work)	$\mathcal{O}(\frac{1}{KT})$	$\mathcal{O}(N^{-1/2}T^{1/2})^\ddagger$	✓	Partial	Bounded gradient	Strongly convex
FedAvgKhaled et al. (2020)	$\mathcal{O}(\frac{1}{\sqrt{NT}})$	$\mathcal{O}(N^{-3/2}T^{1/2})$	✓	Full	Bounded gradient	Convex
FedAvgKoloskova et al. (2020)	$\mathcal{O}(\frac{1}{\sqrt{NT}})$	$\mathcal{O}(N^{-3/4}T^{1/4})$	✓	Full	Bounded gradient	Convex
FedAvgKarimireddy et al. (2019)	$\mathcal{O}(\frac{1}{\sqrt{NT}})^{\dagger\dagger}$	$\mathcal{O}(N^{-3/4}T^{1/4})^{\dagger\dagger}$	✓	Partial	Bounded gradient dissimilarity	Convex
FedAvg/N-FedAvg (our work)	$\mathcal{O}\left(\frac{1}{\sqrt{KT}}\right)$	$\mathcal{O}(N^{-3/4}T^{1/4})^\ddagger$	✓	Partial	Bounded gradient	Convex
FedAvg	$\mathcal{O}\left(\exp(-\frac{NT}{E\kappa_1})\right)$	$\mathcal{O}(T^\beta)$	✓	Partial	Bounded gradient	Overparameterized LR
FedMass	$\mathcal{O}\left(\exp(-\frac{NT}{E\sqrt{\kappa_1\tilde{\kappa}}})\right)$	$\mathcal{O}(T^\beta)$	✓	Partial	Bounded gradient	Overparameterized LR

Table 2: A high-level summary of the convergence results in this paper compared to prior state-of-the-art FL algorithms. This table only highlights the dependence on T (number of iterations), E (the maximal number of local steps), N (the total number of devices), and $K \leq N$ the number of participated devices. κ is the condition number of the system and $\beta \in (0, 1)$. We denote Nesterov accelerated FedAvg as N-FedAvg in this table.

† This E is obtained under i.i.d. setting.

‡ This E is obtained under full participation setting.

§ In Haddadpour & Mahdavi (2019), the convergence rate is for non-convex smooth problems with PL condition, which also applies to strongly convex problems. Therefore, we compare it with our strongly convex results here.

‡‡ The bounded gradient diversity assumption is not applicable for general heterogeneous data when converging to arbitrarily small ϵ-accuracy (see discussions in Sec B).

†† Although the results in Karimireddy et al. (2019) is applicable for partial participation setting, their results only achieve linear speedup under full participation setting $K = N$ while we show linear speedup convergence for $K \leq N$ (see discussions in Sec B). The E in the table is obtained under full participation. Under partial participation, the communication complexity is $E = \mathcal{O}(1)$.

C A High-level Summary of FedAvg analysis

To facilitate the understanding of our analysis and highlight the improvement of our work comparing to prior arts, we summarize the general steps used in the proofs across the various settings. In this section, we take the strongly convex case as an example to illustrate our analysis. The corresponding proof for general convex functions follows the same framework.

One step progress bound

This step establishes the progress of distance ($\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2$) to optimal solution after one step SGD update (see line 9, Alg 1), as the following equation shows:

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathcal{O}(\eta_t \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2\sigma^2/N + \alpha_t^3 E^2 G^2).$$

Algorithm 1 FEDAVG: Federated Averaging

Server input: initial model $\mathbf{w}_0$, initial step size α_0, local steps E.
Client input:
for each round $r = 0, 1, ..., R$, where $r = t * E$ **do**
  Sample clients $\mathcal{S}_t \subseteq \{1, ..., N\}$
  Broadcast $\mathbf{w}$ to all clients $k \in \mathcal{S}_t$
  for each client $k \subseteq \mathcal{S}_t$ **do**
    initialize local model $\mathbf{w}_t^k = \mathbf{w}$
    for $t = r * E + 1, \ldots, (r + 1) * E$ **do**
      $\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \alpha_t \mathbf{g}_{t,k}$
    end for
  end for
  Average the local models at server end: $\overline{\mathbf{w}}_t = \sum_{k \in \mathcal{S}_t} \mathbf{w}_t^k$.
end for

The above bound consists of three main ingredients, the distance to optima in previous step (with $\eta_t \in (0, 1)$ to obtained a contraction bound), the variance of stochastic gradients in local clients (second term), the variance across different clients (third term). Notice that the third term in this bound is the primary source of improvement in the rate. Comparing to the bound in Li et al. (2020b), we improve the third term from $\mathcal{O}(\alpha_t^2 E^2 G^2)$ to $\mathcal{O}(\alpha_t^2 E^2 G^2)$, which enables the linear speedup in the convergence rate.

Iterative deduction
This step uses the *one step progress bound* iteratively to connect the the current distance to optimal solution with the initial distance ($\|\overline{\mathbf{w}}_0 - \mathbf{w}^*\|^2$), as follows:

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathcal{O}(\mathbb{E}\|\overline{\mathbf{w}}_0 - \mathbf{w}^*\|^2 \frac{1}{T}).$$

Then we can use the distance to optima to upper bound the optimality gap ($F(\mathbf{w}_t) - F^* \leq \mathcal{O}(1/T)$), as follows:

$$\mathbb{E}(F(\overline{\mathbf{w}}_t)) - F^* \leq \mathcal{O}(\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2).$$

The convergence rate of the optimality gap is equally obtained as the convergence rate of the distance to optima.

From full participation to partial participation
There are three sources of variances that affect the convergence rate. The first two sources come from the variances of within local clients and across clients (second and third term in one step progress bound). The partial participation, which involves a sampling procedure, is the third source of variance. Therefore, comparing to the rate in full participation, this will add another term of variance into the convergence rate, where we follow a similar derivation as in Li et al. (2020b).

D TECHNICAL LEMMAS

To facilitate reading, we first summarize some basic properties of L-smooth and μ-strongly convex functions, found in e.g. Rockafellar (1970), which are used in various steps of proofs in the appendix.

Lemma 1. *Let F be a convex L-smooth function. Then we have the following inequalities:*

1. Quadratic upper bound: $0 \leq F(\mathbf{w}) - F(\mathbf{w}') - \langle \nabla F(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle \leq \frac{L}{2}\|\mathbf{w} - \mathbf{w}'\|^2$.

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

3. Lower bound: $F(\mathbf{w}) \geq F(\mathbf{w}') + \langle \nabla F(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{1}{2L}\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\|^2$. *In particular,* $\|\nabla F(\mathbf{w})\|^2 \leq 2L(F(\mathbf{w}) - F(\mathbf{w}^*))$.

4. Optimality gap: $F(\mathbf{w}) - F(\mathbf{w}^*) \leq \langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^* \rangle$.

Lemma 2. *Let F be a μ-strongly convex function. Then*

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

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|^2$$

E Proof of Convergence Results for FedAvg

E.1 Strongly Convex Smooth Objectives

To organize our proofs more effectively and highlight the significance of our results compared to prior works, we first state the following key lemmas used in proofs of main results and defer their proofs to later.

Lemma 3 (One step progress, strongly convex). *Let $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{w}_t^k$, and suppose our functions satisfy Assumptions 1,2,3,4, and set step size $\alpha_t = \frac{4}{\mu(\gamma+t)}$ with $\gamma = \max\{32\kappa, E\}$ and $\kappa = \frac{L}{\mu}$, then the updates of FedAvg with full participation satisfy*

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq (1-\mu\alpha_t)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \frac{1}{N}\nu_{max}^2\sigma^2 + 6E^2L\alpha_t^3G^2.$$

We emphasize that the above lemma is the key step that allows us to obtain a bound that improves on the convergence result of Li et al. (2020b) with linear speedup. Its proof will make use of the following two results.

Lemma 4 (Bounding gradient variance (Lemma 2 Li et al. (2020b))). *Given Assumption 3, the upper bound of gradient variance is given as follows,*

$$\mathbb{E}\|\mathbf{g}_t - \overline{\mathbf{g}}_t\|^2 \leq \sum_{k=1}^N p_k^2\sigma_k^2.$$

Lemma 5 (Bounding the divergence of $\mathbf{w}_t^k$ (Lemma 3 Li et al. (2020b))). *Given Assumption 4, and assume that α_t is non-increasing and $\alpha_t \leq 2\alpha_{t+E}$ for all $t \geq 0$, we have*

$$\mathbb{E}\left[\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2\right] \leq 4E^2\alpha_t^2G^2.$$

We now restate Theorem 1 from the main text and then prove it using Lemma 3.

Theorem 1. *Let $\overline{\mathbf{w}}_T = \sum_{k=1}^N p_k \mathbf{w}_T^k$ in FedAvg, $\nu_{\max} = \max_k Np_k$, and set decaying learning rates $\alpha_t = \frac{4}{\mu(\gamma+t)}$ with $\gamma = \max\{32\kappa, E\}$ and $\kappa = \frac{L}{\mu}$. Then under Assumptions 1,2,3,4 with full device participation,*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) - F^* = \mathcal{O}\left(\frac{\kappa\nu_{\max}^2\sigma^2/\mu}{NT} + \frac{\kappa^2E^2G^2/\mu}{T^2}\right)$$

and with partial device participation with at most K sampled devices at each communication round,

$$\mathbb{E}F(\overline{\mathbf{w}}_T) - F^* = \mathcal{O}\left(\frac{\kappa E^2G^2/\mu}{KT} + \frac{\kappa\nu_{\max}^2\sigma^2/\mu}{NT} + \frac{\kappa^2E^2G^2/\mu}{T^2}\right)$$

Proof. The road map of the proof for full device participation contains three steps. First, we establish a recursive relationship between $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$ and $\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2$, upper bounding the progress of FedAvg from step t to step $t+1$. Second, we show that $\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 = \mathcal{O}(\frac{\nu_{max}^2\sigma^2/\mu}{tN} + \frac{E^2LG^2/\mu^2}{t^2})$ by induction using the recursive relationship from the previous step. Third, we use the property of L-smoothness to bound the optimality gap by $\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2$.

By Lemma 3, we have the following upper bound for the one step progress:

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq (1-\mu\alpha_t)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \frac{1}{N}\nu_{max}^2\sigma^2 + 6E^2L\alpha_t^3G^2.$$

We show next that $\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 = O(\frac{\nu_{max}^2\sigma^2/\mu}{tN} + \frac{E^2LG^2/\mu^2}{t^2})$ using induction. To simplify the presentation, we denote $C \equiv 6E^2LG^2$ and $D \equiv \frac{1}{N}\nu_{max}^2\sigma^2$. Suppose that we have the bound $\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 \le b \cdot (\alpha_t D + \alpha_t^2 C)$ for some constant b and learning rates α_t. Then the one step progress from Lemma 3 becomes:

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \le (b(1-\mu\alpha_t)+\alpha_t)\alpha_t D + (b(1-\mu\alpha_t)+\alpha_t)\alpha_t^2 C$$

To establish the result at step $t{+}1$, it remains to choose α_t and b such that $(b(1{-}\mu\alpha_t){+}\alpha_t)\alpha_t \le b\alpha_{t+1}$ and $(b(1-\mu\alpha_t)+\alpha_t)\alpha_t^2 \le b\alpha_{t+1}^2$. If we let $\alpha_t = \frac{4}{\mu(t+\gamma)}$ where $\gamma = \max\{E, 32\kappa\}$ (choice of γ required to guarantee the one step progress) and set $b = \frac{4}{\mu}$, we have:

$$\begin{aligned}(b(1-\mu\alpha_t)+\alpha_t)\alpha_t &= \left(b(1-\frac{4}{t+\gamma}) + \frac{4}{\mu(t+\gamma)}\right)\frac{4}{\mu(t+\gamma)} \le b\frac{4}{\mu(t+\gamma+1)} = b\alpha_{t+1}\\ (b(1-\mu\alpha_t)+\alpha_t)\alpha_t^2 &= b(\frac{t+\gamma-2}{t+\gamma})\frac{16}{\mu^2(t+\gamma)^2} \le b\frac{16}{\mu^2(t+\gamma+1)^2} = b\alpha_{t+1}^2\end{aligned}$$

where we have used the following inequalities:

$$\frac{t+\gamma-1}{(t+\gamma)^2} \le \frac{1}{(t+\gamma+1)} \qquad \frac{t+\gamma-2}{(t+\gamma)^3} \le \frac{1}{(t+\gamma+1)^2} \qquad \forall\, \gamma \ge 1$$

Thus we have established the result at step $t+1$ assuming the result is correct at step t:

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \le b \cdot (\alpha_{t+1}D + \alpha_{t+1}^2 C)$$

At step $t = 0$, we can ensure the following inequality by scaling b with $c\|\mathbf{w}_0 - \mathbf{w}^*\|^2$ for a sufficiently large constant c:

$$\|\mathbf{w}_0 - \mathbf{w}^*\|^2 \le b \cdot (\alpha_0 D + \alpha_0^2 C) = b \cdot (\frac{4}{\mu\gamma}D + \frac{16}{\mu^2\gamma^2}C)$$

It follows that

$$\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 \le c\|\mathbf{w}_0 - \mathbf{w}^*\|^2 \frac{4}{\mu}(D\alpha_t + C\alpha_t^2) \tag{8}$$

for all $t \ge 0$.

Finally, the L-smoothness of F implies

$$\begin{aligned}\mathbb{E}(F(\overline{\mathbf{w}}_T)) - F^* &\le \frac{L}{2}\mathbb{E}\|\overline{\mathbf{w}}_T - \mathbf{w}^*\|^2\\ &\le \frac{L}{2}c\|\mathbf{w}_0 - \mathbf{w}^*\|^2\frac{4}{\mu}(D\alpha_T + C\alpha_T^2)\\ &= 2c\|\mathbf{w}_0 - \mathbf{w}^*\|^2\kappa(D\alpha_T + C\alpha_T^2)\\ &\le 2c\|\mathbf{w}_0 - \mathbf{w}^*\|^2\kappa\left[\frac{4}{\mu(T+\gamma)}\cdot\frac{1}{N}\nu_{max}^2\sigma^2 + 6E^2LG^2\cdot(\frac{4}{\mu(T+\gamma)})^2\right]\\ &= \mathcal{O}(\frac{\kappa}{\mu}\frac{1}{N}\nu_{max}^2\sigma^2\cdot\frac{1}{T} + \frac{\kappa^2}{\mu}E^2G^2\cdot\frac{1}{T^2})\end{aligned}$$

where in the first line, we use the property of L-smooth function (see Lemma 1), and in the second line, we use the conclusion in Eq (8).

With partial participation, the update at each communication round is now given by weighted averages over a subset of sampled devices. When $t+1 \notin \mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_{t+1}$, while when $t+1 \in \mathcal{I}_E$, we have $\mathbb{E}\overline{\mathbf{w}}_{t+1} = \overline{\mathbf{v}}_{t+1}$ by design of the sampling schemes (Li et al. (2020b), Lemma 4), so that

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2\\ &= \mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2 + \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2\end{aligned}$$

This in particular implies $\mathbb{E}\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 \le \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2$ for all t. Since $\overline{\mathbf{v}}_t = \sum_{k=1}^N p_k\mathbf{v}_t^k$ always averages over all devices, the full participation one step progress result Lemma 3 applied to $\overline{\mathbf{v}}_t$ implies

$$\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \le \mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 6E^2L\alpha_t^3G^2 + \alpha_t^2\frac{1}{N}\nu_{max}^2\sigma^2$$

$$\leq \mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 6E^2 L\alpha_t^3 G^2 + \alpha_t^2 \frac{1}{N}\nu_{max}^2\sigma^2$$

The bound for $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2$ for the two sampling schemes we consider is provided in Eq (7), and applying it we can write the one step progress for partial participation as

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq (1-\mu\alpha_t)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2\frac{1}{N}\nu_{max}^2\sigma^2 + \frac{4}{K}\alpha_t^2 E^2 G^2 + 6E^2 L\alpha_t^3 G^2,$$

and the same arguments using induction and L-smoothness as the full participation case implies

$$\mathbb{E}F(\overline{\mathbf{w}}_T) - F^* = \mathcal{O}(\frac{\kappa\nu_{\max}^2\sigma^2/\mu}{NT} + \frac{\kappa E^2 G^2/\mu}{KT} + \frac{\kappa^2 E^2 G^2/\mu}{T^2})$$

□

E.1.1 DEFERRED PROOFS OF KEY LEMMAS

Here we first rewrite the proofs of lemmas 4 and 5 from Li et al. (2020b) with slight modifications for the consistency and completeness of this work, since later we will use modified versions of these results in the convergence proof for Nesterov accelerated FedAvg.

Proof of lemma 4.

$$\mathbb{E}\|\mathbf{g}_t - \overline{\mathbf{g}}_t\|^2 = \mathbb{E}\|\mathbf{g}_t - \mathbb{E}\mathbf{g}_t\|^2 = \sum_{k=1}^N p_k^2\|\mathbf{g}_{t,k} - \mathbb{E}\mathbf{g}_{t,k}\|^2 \leq \sum_{k=1}^N p_k^2\sigma_k^2$$

□

Proof of lemma 5. Now we bound $\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2$ following Li et al. (2020b). Since communication is done every E steps, for any $t \geq 0$, we can find a $t_0 \leq t$ such that $t - t_0 \leq E-1$ and $\mathbf{w}_{t_0}^k = \overline{\mathbf{w}}_{t_0}$for all k. Moreover, using α_t is non-increasing and $\alpha_{t_0} \leq 2\alpha_t$ for any $t - t_0 \leq E-1$, we have

$$\begin{aligned}
&\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \\
=&\mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_{t_0} - (\overline{\mathbf{w}}_t - \overline{\mathbf{w}}_{t_0})\|^2 \\
\leq&\mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_{t_0}\|^2 \\
=&\mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \mathbf{w}_{t_0}^k\|^2 \\
=&\mathbb{E}\sum_{k=1}^N p_k\| - \sum_{i=t_0}^{t-1}\alpha_i\mathbf{g}_{i,k}\|^2 \\
\leq&2\sum_{k=1}^N p_k\mathbb{E}\sum_{i=t_0}^{t-1}E\alpha_i^2\|\mathbf{g}_{i,k}\|^2 \\
\leq&2\sum_{k=1}^N p_k E^2\alpha_{t_0}^2 G^2 \\
\leq&4E^2\alpha_t^2 G^2
\end{aligned}$$

□

Based on the results of Lemma 4, 5, we now prove the upper bound of one step SGD progress. This proof improves on the previous work Li et al. (2020b) and is the first to reveal the linear speedup of convergence of FedAvg.

Proof of lemma 3. We have

$$\begin{aligned}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 &= \|(\overline{\mathbf{w}}_t-\alpha_t\mathbf{g}_t)-\mathbf{w}^*\|^2 = \|(\overline{\mathbf{w}}_t-\alpha_t\overline{\mathbf{g}}_t-\mathbf{w}^*)-\alpha_t(\mathbf{g}_t-\overline{\mathbf{g}}_t)\|^2 \\ &= \underbrace{\|\overline{\mathbf{w}}_t-\mathbf{w}^*-\alpha_t\overline{\mathbf{g}}_t\|^2}_{A_1}+\underbrace{2\alpha_t\langle\overline{\mathbf{w}}_t-\mathbf{w}^*-\alpha_t\overline{\mathbf{g}}_t,\overline{\mathbf{g}}_t-\mathbf{g}_t\rangle}_{A_2}+\underbrace{\alpha_t^2\|\mathbf{g}_t-\overline{\mathbf{g}}_t\|^2}_{A_3}\end{aligned}$$

where we denote:

$$\begin{aligned}A_1 &= \|\overline{\mathbf{w}}_t-\mathbf{w}^*-\alpha_t\overline{\mathbf{g}}_t\|^2 \\ A_2 &= 2\alpha_t\langle\overline{\mathbf{w}}_t-\mathbf{w}^*-\alpha_t\overline{\mathbf{g}}_t,\overline{\mathbf{g}}_t-\mathbf{g}_t\rangle \\ A_3 &= \alpha_t^2\|\mathbf{g}_t-\overline{\mathbf{g}}_t\|^2\end{aligned}$$

By definition of $\mathbf{g}_t$ and $\overline{\mathbf{g}}_t$ (see Eq (4)), we have $\mathbb{E}A_2 = 0$. For A_3, we have the following upper bound (see Lemma 4):

$$\alpha_t^2\mathbb{E}\|\mathbf{g}_t-\overline{\mathbf{g}}_t\|^2 \leq \alpha_t^2\sum_{k=1}^N p_k^2\sigma_k^2$$

Next we bound A_1:

$$\|\overline{\mathbf{w}}_t-\mathbf{w}^*-\alpha_t\overline{\mathbf{g}}_t\|^2 = \|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2+2\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,-\alpha_t\overline{\mathbf{g}}_t\rangle+\|\alpha_t\overline{\mathbf{g}}_t\|^2$$

and we will show that the third term $\|\alpha_t\overline{\mathbf{g}}_t\|^2$ can be canceled by an upper bound of the second term, which is one of major improvement comparing to prior art Li et al. (2020b). The upper bound of second term can be derived as follows, using the strong convexity and L-smoothness of F_k:

$$\begin{aligned}&-2\alpha_t\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\overline{\mathbf{g}}_t\rangle \\ =&-2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\nabla F_k(\mathbf{w}_t^k)\rangle \\ =&-2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}_t^k,\nabla F_k(\mathbf{w}_t^k)\rangle-2\alpha_t\sum_{k=1}^N p_k\langle\mathbf{w}_t^k-\mathbf{w}^*,\nabla F_k(\mathbf{w}_t^k)\rangle \\ \leq&-2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}_t^k,\nabla F_k(\mathbf{w}_t^k)\rangle+2\alpha_t\sum_{k=1}^N p_k(F_k(\mathbf{w}^*)-F_k(\mathbf{w}_t^k))-\alpha_t\mu\sum_{k=1}^N p_k\|\mathbf{w}_t^k-\mathbf{w}^*\|^2 \\ \leq&2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}_t^k)-F_k(\overline{\mathbf{w}}_t)+\frac{L}{2}\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2+F_k(\mathbf{w}^*)-F_k(\mathbf{w}_t^k)\right]-\alpha_t\mu\|\sum_{k=1}^N p_k\mathbf{w}_t^k-\mathbf{w}^*\|^2 \\ =&\alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2+2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*)-F_k(\overline{\mathbf{w}}_t)\right]-\alpha_t\mu\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2\end{aligned}$$

We record the bound we have obtained so far, as it will also be used in the proof for convex case:

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 \leq&\mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2+\alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2 \\ &+2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*)-F_k(\overline{\mathbf{w}}_t)\right]+\alpha_t^2\sum_{k=1}^N p_k^2\sigma_k^2+\alpha_t^2\|\overline{\mathbf{g}}_t\|^2 \qquad (9)\end{aligned}$$

For the term $2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*)-F_k(\overline{\mathbf{w}}_t)\right]$, which is negative, we can ignore it, but this yields a suboptimal bound that fails to provide the desired linear speedup. Instead, we upper bound it using the following derivation:

$$2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*)-F_k(\overline{\mathbf{w}}_t)\right]$$

$$
\begin{aligned}
&\leq 2\alpha_t \left[F(\overline{\mathbf{w}}_{t+1}) - F(\overline{\mathbf{w}}_t)\right] \\
&\leq 2\alpha_t \mathbb{E}\langle \nabla F(\overline{\mathbf{w}}_t), \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{w}}_t \rangle + \alpha_t L \mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{w}}_t\|^2 \\
&= -2\alpha_t^2 \mathbb{E}\langle \nabla F(\overline{\mathbf{w}}_t), \mathbf{g}_t \rangle + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&= -2\alpha_t^2 \mathbb{E}\langle \nabla F(\overline{\mathbf{w}}_t), \overline{\mathbf{g}}_t \rangle + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&= -\alpha_t^2 \left[\|\nabla F(\overline{\mathbf{w}}_t)\|^2 + \|\overline{\mathbf{g}}_t\|^2 - \|\nabla F(\overline{\mathbf{w}}_t) - \overline{\mathbf{g}}_t\|^2\right] + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&= -\alpha_t^2 \left[\|\nabla F(\overline{\mathbf{w}}_t)\|^2 + \|\overline{\mathbf{g}}_t\|^2 - \|\nabla F(\overline{\mathbf{w}}_t) - \sum_k p_k \nabla F(\mathbf{w}_t^k)\|^2\right] + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&\leq -\alpha_t^2 \left[\|\nabla F(\overline{\mathbf{w}}_t)\|^2 + \|\overline{\mathbf{g}}_t\|^2 - \sum_k p_k \|\nabla F(\overline{\mathbf{w}}_t) - \nabla F(\mathbf{w}_t^k)\|^2\right] + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&\leq -\alpha_t^2 \left[\|\nabla F(\overline{\mathbf{w}}_t)\|^2 + \|\overline{\mathbf{g}}_t\|^2 - L^2 \sum_k p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2\right] + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 \\
&\leq -\alpha_t^2 \|\overline{\mathbf{g}}_t\|^2 + \alpha_t^2 L^2 \sum_k p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 - \alpha_t^2 \|\nabla F(\overline{\mathbf{w}}_t)\|^2
\end{aligned}
$$

where we have used the smoothness of F twice.

Note that the term $-\alpha_t^2\|\overline{\mathbf{g}}_t\|^2$ exactly cancels the $\alpha_t^2\|\overline{\mathbf{g}}_t\|^2$ in the bound in Eq (9), so that plugging in the bound for $-2\alpha_t\langle \overline{\mathbf{w}}_t - \mathbf{w}^*, \overline{\mathbf{g}}_t\rangle$, we have so far proved

$$
\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 \\
+ \alpha_t^2 L^2 \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 - \alpha_t^2 \|\nabla F(\overline{\mathbf{w}}_t)\|^2 \qquad (10)
\end{aligned}
$$

Under Assumption 4, we have $\mathbb{E}\|\mathbf{g}_t\|^2 \leq G^2$. Furthermore, we can check that our choice of α_t satisfies α_t is non-increasing and $\alpha_t \leq 2\alpha_{t+E}$, so we may plug in the bound $\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \leq 4E^2\alpha_t^2 G^2$ to the above inequality (see Lemma 5).

Therefore, we can conclude that, with $\nu_{max} := N \cdot \max_k p_k$ and $\nu_{min} := N \cdot \min_k p_k$,

$$
\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \\
\leq &\mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 4E^2 L\alpha_t^3 G^2 + 4E^2 L^2 \alpha_t^4 G^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 + \alpha_t^3 L G^2 \\
= &\mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 4E^2 L\alpha_t^3 G^2 + 4E^2 L^2 \alpha_t^4 G^2 + \alpha_t^2 \frac{1}{N^2} \sum_{k=1}^N (p_k N)^2 \sigma_k^2 + \alpha_t^3 L G^2 \\
\leq &\mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 4E^2 L\alpha_t^3 G^2 + 4E^2 L^2 \alpha_t^4 G^2 + \alpha_t^2 \frac{1}{N^2} \nu_{max}^2 \sum_{k=1}^N \sigma_k^2 + \alpha_t^3 L G^2 \\
\leq &\mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 6E^2 L\alpha_t^3 G^2 + \alpha_t^2 \frac{1}{N} \nu_{max}^2 \sigma^2
\end{aligned}
$$

where in the last inequality we use $\sigma^2 = \sum_{k=1}^N p_k \sigma_k^2$, and that by construction α_t satisfies $L\alpha_t \leq \frac{1}{8}$.

□

One may ask whether the dependence on E in the term $\frac{\kappa E^2 G^2/\mu}{KT}$ can be removed, or equivalently whether $\sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 = \mathcal{O}(1/T^2)$ can be independent of E. We provide a simple counterexample that shows that this is not possible in general.

Proposition 1. *There exists a dataset such that if $E = \mathcal{O}(T^\beta)$ for any $\beta > 0$ then $\sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 = \Omega(\frac{1}{T^{2-2\beta}})$.*

Proof. Suppose that we have an even number of devices and each $F_k(\mathbf{w}) = \frac{1}{n_k}\sum_{j=1}^{n_k}(\mathbf{x}_k^j - \mathbf{w})^2$ contains data points $\mathbf{x}_k^j = \mathbf{w}^{*,k}$, with $n_k \equiv n$. Moreover, the $\mathbf{w}^{*,k}$'s come in pairs around the origin. As a result, the global objective F is minimized at $\mathbf{w}^* = 0$. Moreover, if we start from $\overline{\mathbf{w}}_0 = 0$, then by design of the dataset the updates in local steps exactly cancel each other at each iteration, resulting in $\overline{\mathbf{w}}_t = 0$ for all t. On the other hand, if $E = T^\beta$, then starting from any $t = \mathcal{O}(T)$ with constant step size $\mathcal{O}(\frac{1}{T})$, after E iterations of local steps, the local parameters are updated towards $\mathbf{w}^{*,k}$ with $\|\mathbf{w}_{t+E}^k\|^2 = \Omega((T^\beta \cdot \frac{1}{T})^2) = \Omega(\frac{1}{T^{2-2\beta}})$. This implies that

$$\begin{aligned}\sum_k p_k\|\mathbf{w}_{t+E}^k - \overline{\mathbf{w}}_{t+E}\|^2 &= \sum_k p_k\|\mathbf{w}_{t+E}^k\|^2 \\ &= \Omega(\frac{1}{T^{2-2\beta}})\end{aligned}$$

which is at a slower rate than $\frac{1}{T^2}$ for any $\beta > 0$. Thus the sampling variance $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2 = \Omega(\sum_k p_k \mathbb{E}\|\mathbf{w}_{t+1}^k - \overline{\mathbf{w}}_{t+1}\|^2)$ decays at a slower rate than $\frac{1}{T^2}$, resulting in a convergence rate slower than $\mathcal{O}(\frac{1}{T})$ with partial participation. □

E.2 Convex Smooth Objectives

In this section we provide the proof of the convergence result for FedAvg with convex and smooth objectives. The key step is a one step progress result analogous to that in the strongly convex case, and their proofs share identical components as well.

Lemma 6 (**One step progress, convex case**). *Let $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{w}_t^k$ in FedAvg. Under assumptions 1,3,4, the following bound holds for all t:*

$$\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \alpha_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \frac{1}{N}\nu_{\max}^2\sigma^2 + 6\alpha_t^3 E^2 L G^2$$

Proof. The first part of the proof follows directly from Eq (9) in the proof of Lemma 3. Setting $\mu = 0$ in Eq (9) (since we are in the convex setting instead of strongly convex), we obtain

$$\begin{aligned}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t L \sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \\ + 2\alpha_t \sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right] + \alpha_t^2\|\overline{\mathbf{g}}_t\|^2 + \alpha_t^2\sum_{k=1}^N p_k^2\sigma_k^2\end{aligned}$$

The difference of this bound with that in the strongly convex case is that we no longer have a contraction factor of $1 - \mu\alpha_t$ in front of $\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2$. In the strongly convex case, we were able to cancel $\alpha_t^2\|\overline{\mathbf{g}}_t\|^2$ with $2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right]$ and obtain only lower order terms. In the convex case, we use a different strategy and preserve $\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right]$ in order to obtain the desired optimality gap.

We have

$$\begin{aligned}\|\overline{\mathbf{g}}_t\|^2 &= \|\sum_k p_k\nabla F_k(\mathbf{w}_t^k)\|^2 \\ &= \|\sum_k p_k\nabla F_k(\mathbf{w}_t^k) - \sum_k p_k\nabla F_k(\overline{\mathbf{w}}_t) + \sum_k p_k\nabla F_k(\overline{\mathbf{w}}_t)\|^2 \\ &\leq 2\|\sum_k p_k\nabla F_k(\mathbf{w}_t^k) - \sum_k p_k\nabla F_k(\overline{\mathbf{w}}_t)\|^2 + 2\|\sum_k p_k\nabla F_k(\overline{\mathbf{w}}_t)\|^2 \\ &\leq 2L^2\sum_k p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 + 2\|\sum_k p_k\nabla F_k(\overline{\mathbf{w}}_t)\|^2 \\ &= 2L^2\sum_k p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 + 2\|\nabla F(\overline{\mathbf{w}}_t)\|^2\end{aligned}$$

using $\nabla F(\mathbf{w}^*) = 0$. Now using the L smoothness of F, we have $\|\nabla F(\overline{\mathbf{w}}_t)\|^2 \leq 2L(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*))$, so that

$$\begin{aligned}
&\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \\
\leq &\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + 2\alpha_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right] \\
&+ 2\alpha_t^2 L^2 \sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 + 4\alpha_t^2 L(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 \\
= &\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + (2\alpha_t^2 L^2 + \alpha_t L) \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right] \\
&+ \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 + \alpha_t(1 - 4\alpha_t L)(F(\mathbf{w}^*) - F(\overline{\mathbf{w}}_t))
\end{aligned}$$

Since $F(\mathbf{w}^*) \leq F(\overline{\mathbf{w}}_t)$, as long as $4\alpha_t L \leq 1$, we can ignore the last term, and rearrange the inequality to obtain

$$\begin{aligned}
&\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \alpha_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \\
\leq &\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + (2\alpha_t^2 L^2 + \alpha_t L) \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 \\
\leq &\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \frac{3}{2}\alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2
\end{aligned}$$

The same argument as before yields $\mathbb{E}\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \leq 4E^2\alpha_t^2 G^2$ which gives

$$\begin{aligned}
\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \alpha_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) &\leq \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 + 6\alpha_t^3 E^2 L G^2 \\
&\leq \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \frac{1}{N}\nu_{\max}^2 \sigma^2 + 6\alpha_t^3 E^2 L G^2
\end{aligned}$$

□

With the one step progress result, we can now prove the convergence result in the convex setting, which we restate below.

Theorem 2. *Under assumptions 1,3,4 and constant learning rate $\alpha_t = \mathcal{O}(\sqrt{\frac{N}{T}})$, FedAvg satisfies*

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\nu_{\max}\sigma^2}{\sqrt{NT}} + \frac{NE^2LG^2}{T}\right)$$

with full participation, and with partial device participation with K sampled devices at each communication round and learning rate $\alpha_t = \mathcal{O}(\sqrt{\frac{K}{T}})$,

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\nu_{\max}\sigma^2}{\sqrt{KT}} + \frac{E^2G^2}{\sqrt{KT}} + \frac{KE^2LG^2}{T}\right)$$

Proof. We first prove the bound for full participation. Applying Lemma 6, we have

$$\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \alpha_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t^2 \frac{1}{N}\nu_{\max}^2 \sigma^2 + 6\alpha_t^3 E^2 L G^2$$

Summing the inequalities from $t = 0$ to $t = T$, we obtain

$$\sum_{t=0}^T \alpha_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^T \alpha_t^2 \cdot \frac{1}{N}\nu_{\max}^2 \sigma^2 + \sum_{t=0}^T \alpha_t^3 \cdot 6E^2LG^2$$

so that

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) \leq \frac{1}{\sum_{t=0}^T \alpha_t}\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^T \alpha_t^2 \cdot \frac{1}{N}\nu_{\max}^2\sigma^2 + \sum_{t=0}^T \alpha_t^3 \cdot 6E^2LG^2\right)$$

By setting the constant learning rate $\alpha_t \equiv \sqrt{\frac{N}{T}}$, we have

$$\begin{aligned}
\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) &\leq \frac{1}{\sqrt{NT}} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{1}{\sqrt{NT}} T \cdot \frac{N}{T} \cdot \frac{1}{N}\nu_{\max}^2\sigma^2 + \frac{1}{\sqrt{NT}} T (\sqrt{\frac{N}{T}})^3 6E^2LG^2 \\
&\leq \frac{1}{\sqrt{NT}} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{1}{\sqrt{NT}} T \cdot \frac{N}{T} \cdot \frac{1}{N}\nu_{\max}^2\sigma^2 + \frac{N}{T} 6E^2LG^2 \\
&= (\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \nu_{\max}^2\sigma^2)\frac{1}{\sqrt{NT}} + \frac{N}{T} 6E^2LG^2 \\
&= \mathcal{O}(\frac{\nu_{\max}^2\sigma^2}{\sqrt{NT}} + \frac{NE^2LG^2}{T})
\end{aligned}$$

Similarly, for partial participation, we have

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) \leq \frac{1}{\sum_{t=0}^T \alpha_t}\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^T \alpha_t^2 \cdot (\frac{1}{N}\nu_{\max}\sigma^2 + C) + \sum_{t=0}^T \alpha_t^3 \cdot 6E^2LG^2\right)$$

where $C = \frac{4}{K}E^2G^2$ or $\frac{N-K}{N-1}\frac{4}{K}E^2G^2$ depending on the sampling scheme, so that with $\alpha_t = \sqrt{\frac{K}{T}}$, we have

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) = \mathcal{O}(\frac{\nu_{\max}\sigma^2}{\sqrt{KT}} + \frac{E^2G^2}{\sqrt{KT}} + \frac{KE^2LG^2}{T})$$

□

F PROOF OF CONVERGENCE RESULTS FOR NESTEROV ACCELERATED FEDAVG

F.1 STRONGLY CONVEX SMOOTH OBJECTIVES

Recall that the Nesterov accelerated FedAvg follows the updates

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \alpha_t \mathbf{g}_{t,k}, \quad \mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k + \beta_t(\mathbf{v}_{t+1}^k - \mathbf{v}_t^k) & \text{if } t+1 \notin \mathcal{I}_E, \\ \sum_{k\in\mathcal{S}_{t+1}} q_k \left[\mathbf{v}_{t+1}^k + \beta_t(\mathbf{v}_{t+1}^k - \mathbf{v}_t^k)\right] & \text{if } t+1 \in \mathcal{I}_E. \end{cases}$$

The proofs of convergence results for Nesterov Accelerated FedAvg consists of components that are direct analogues of the FedAvg case. We first state these analogue results before proving the main theorem. Like before, the proofs of the lemmas are deferred to after the main proof.

Lemma 7 (One step progress, Nesterov). *Let $\overline{\mathbf{v}}_t = \sum_{k=1}^N p_k \mathbf{v}_t^k$ in Nesterov accelerated FedAvg, and suppose our functions satisfy Assumptions 1,2,3,4, and set step sizes $\alpha_t = \frac{6}{\mu}\frac{1}{t+\gamma}$, $\beta_{t-1} = \frac{3}{14(t+\gamma)(1-\frac{6}{t+\gamma})\max\{\mu,1\}}$ with $\gamma = \max\{32\kappa, E\}$ and $\kappa = \frac{L}{\mu}$, the updates of Nesterov accelerated FedAvg satisfy*

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\mu\alpha_t)(1+\beta_{t-1})^2\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 20E^2L\alpha_t^3G^2 + (1-\alpha_t\mu)\beta_{t-1}^2\|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \\
+ \alpha_t^2\frac{1}{N}\nu_{\max}\sigma^2 + 2\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\|
\end{aligned}$$

The one step progress result makes use of the same bound on the gradient variance in Lemma 4, as well as a divergence bound analogous to Lemma 5, which we state below.

Lemma 8 (**Bounding the divergence of $\mathbf{w}_t^k$, Nesterov**). *Given Assumption 4, and assume that α_t is non-increasing, $\alpha_t \leq 2\alpha_{t+E}$, and $2\beta_{t-1}^2 + 2\alpha_t^2 \leq 1/2$ for all $t \geq 0$, $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{w}_t^k$ in Nesterov accelerated FedAvg satisfies*

$$\mathbb{E}\left[\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2\right] \leq 16(E-1)^2\alpha_t^2 G^2.$$

Theorem 3. *Let $\overline{\mathbf{v}}_T = \sum_{k=1}^N p_k \mathbf{v}_T^k$ in Nesterov accelerated FedAvg and set learning rates $\alpha_t = \frac{6}{\mu}\frac{1}{t+\gamma}$, $\beta_{t-1} = \frac{3}{14(t+\gamma)(1-\frac{6}{t+\gamma})\max\{\mu,1\}}$. Then under Assumptions 1,2,3,4 with full device participation,*

$$\mathbb{E}F(\overline{\mathbf{v}}_T) - F^* = \mathcal{O}\left(\frac{\kappa\nu_{\max}\sigma^2/\mu}{NT} + \frac{\kappa^2 E^2 G^2/\mu}{T^2}\right),$$

and with partial device participation with K sampled devices at each communication round,

$$\mathbb{E}F(\overline{\mathbf{v}}_T) - F^* = \mathcal{O}\left(\frac{\kappa\nu_{\max}\sigma^2/\mu}{NT} + \frac{\kappa E^2 G^2/\mu}{KT} + \frac{\kappa^2 E^2 G^2/\mu}{T^2}\right).$$

Proof. We first prove the result for full participation. Applying the one step progress bound in Lemma 7, we have

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\mu\alpha_t)(1+\beta_{t-1})^2\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 20E^2L\alpha_t^3G^2 + (1-\alpha_t\mu)\beta_{t-1}^2\|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \\ + \alpha_t^2\frac{1}{N}\nu_{\max}\sigma^2 + 2\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\|\end{aligned}$$

Recall that we require $\alpha_{t_0} \leq 2\alpha_t$ for any $t - t_0 \leq E-1$, $L\alpha_t \leq \frac{1}{5}$, and $2\beta_{t-1}^2 + 2\alpha_t^2 \leq 1/2$ in order for Lemmas 8 and 7 to hold, which we can check by definition of α_t and β_t.
We show next that $\mathbb{E}\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 = \mathcal{O}(\frac{\nu_{max}^2\sigma^2/\mu}{tN} + \frac{E^2LG^2/\mu^2}{t^2})$ by induction. Assume that we have shown

$$\mathbb{E}\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 \leq b(C\alpha_t^2 + D\alpha_t)$$

for all iterations until t, where $C = 20E^2LG^2$, $D = \frac{1}{N}\nu_{max}^2\sigma^2$, and b is some constant to be chosen later. For step sizes recall that we choose $\alpha_t = \frac{6}{\mu}\frac{1}{t+\gamma}$ and $\beta_{t-1} = \frac{3}{14(t+\gamma)(1-\frac{6}{t+\gamma})\max\{\mu,1\}}$ where $\gamma = \max\{32\kappa, E\}$, so that $\beta_{t-1} \leq \alpha_t$ and

$$\begin{aligned}(1-\mu\alpha_t)(1+14\beta_{t-1}) &\leq (1-\frac{6}{t+\gamma})(1+\frac{3}{(t+\gamma)(1-\frac{6}{t+\gamma})}) \\ &= 1 - \frac{6}{t+\gamma} + \frac{3}{t+\gamma} = 1 - \frac{3}{t+\gamma} = 1 - \frac{\mu\alpha_t}{2}\end{aligned}$$

Moreover, $\mathbb{E}\|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\|^2 \leq b(C\alpha_{t-1}^2 + D\alpha_{t-1}) \leq 4b(C\alpha_t^2 + D\alpha_t)$ with the chosen step sizes. Therefore the bound for $\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2$ can be further simplified with

$$2\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu)\mathbb{E}\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\| \leq 4\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu) \cdot b(C\alpha_t^2 + D\alpha_t)$$

and

$$(1-\alpha_t\mu)\beta_{t-1}^2\mathbb{E}\|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \leq 4(1-\alpha_t\mu)\beta_{t-1}^2 \cdot b(C\alpha_t^2 + D\alpha_t)$$

so that

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 &\leq (1-\mu\alpha_t)((1+\beta_{t-1})^2 + 4\beta_{t-1}(1+\beta_{t-1}) + 4\beta_{t-1}^2) \cdot b(C\alpha_t^2 + D\alpha_t) \\ &\quad + 20E^2L\alpha_t^3G^2 + \alpha_t^2\frac{1}{N}\nu_{\max}\sigma^2 \\ &\leq \mathbb{E}(1-\mu\alpha_t)(1+14\beta_{t-1}) \cdot b(C\alpha_t^2 + D\alpha_t) + 20E^2L\alpha_t^3G^2 + \alpha_t^2\frac{1}{N}\nu_{\max}\sigma^2\end{aligned}$$

$$\leq b(1-\frac{\mu\alpha_t}{2})(C\alpha_t^2+D\alpha_t)+C\alpha_t^3+D\alpha_t^2$$
$$=(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_t^2C+(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_tD$$

and so it remains to choose b such that

$$(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_t\leq b\alpha_{t+1}$$
$$(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_t^2\leq b\alpha_{t+1}^2$$

from which we can conclude $\mathbb{E}\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\|^2\leq\alpha_{t+1}^2C+\alpha_{t+1}D$.

With $b=\frac{6}{\mu}$, we have

$$\begin{aligned}(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_t&=(b(1-(\frac{3}{t+\gamma})+\frac{6}{\mu(t+\gamma)})\frac{6}{\mu(t+\gamma)}\\&=(b\frac{t+\gamma-3}{t+\gamma}+\frac{6}{\mu(t+\gamma)})\frac{6}{\mu(t+\gamma)}\\&\leq b(\frac{t+\gamma-1}{t+\gamma})\frac{6}{\mu(t+\gamma)}\\&\leq b\frac{6}{\mu(t+\gamma+1)}=b\alpha_{t+1}\end{aligned}$$

where we have used $\frac{t+\gamma-1}{(t+\gamma)^2}\leq\frac{1}{t+\gamma+1}$.

Similarly

$$\begin{aligned}(b(1-\frac{\mu\alpha_t}{2})+\alpha_t)\alpha_t^2&=(b(1-(\frac{3}{t+\gamma})+\frac{6}{\mu(t+\gamma)})(\frac{6}{\mu(t+\gamma)})^2\\&=(b\frac{t+\gamma-3}{t+\gamma}+\frac{6}{\mu(t+\gamma)})(\frac{6}{\mu(t+\gamma)})^2\\&=b(\frac{t+\gamma-2}{t+\gamma})(\frac{6}{\mu(t+\gamma)})^2\\&\leq b\frac{36}{\mu^2(t+\gamma+1)^2}=b\alpha_{t+1}^2\end{aligned}$$

where we have used $\frac{t+\gamma-2}{(t+\gamma)^3}\leq\frac{1}{(t+\gamma+1)^2}$.

Finally, to ensure $\|\mathbf{v}_0-\mathbf{w}^*\|^2\leq b(C\alpha_0^2+D\alpha_0)$, we can rescale b by $c\|\mathbf{v}_0-\mathbf{w}^*\|^2$ for some c. It follows that $\mathbb{E}\|\overline{\mathbf{v}}_t-\mathbf{w}^*\|^2\leq b(C\alpha_t^2+D\alpha_t)$ for all $t\geq 0$. Using the L-smooothness of F,

$$\begin{aligned}\mathbb{E}(F(\overline{\mathbf{v}}_T))-F^*&=\mathbb{E}(F(\overline{\mathbf{v}}_T)-F(\mathbf{w}^*))\\&\leq\frac{L}{2}\mathbb{E}\|\overline{\mathbf{v}}_T-\mathbf{w}^*\|^2\leq\frac{L}{2}c\|\mathbf{v}_0-\mathbf{w}^*\|^2\frac{6}{\mu}(D\alpha_T+C\alpha_T^2)\\&=3c\|\mathbf{v}_0-\mathbf{w}^*\|^2\kappa(D\alpha_T+C\alpha_T^2)\\&\leq 3c\|\mathbf{v}_0-\mathbf{w}^*\|^2\kappa\left[\frac{6}{\mu(T+\gamma)}\cdot\frac{1}{N}\nu_{\max}\sigma^2+20E^2LG^2\cdot(\frac{6}{\mu(T+\gamma)})^2\right]\\&=\mathcal{O}(\frac{\kappa}{\mu}\frac{1}{N}\nu_{\max}\sigma^2\cdot\frac{1}{T}+\frac{\kappa^2}{\mu}E^2G^2\cdot\frac{1}{T^2})\end{aligned}$$

With partial participation, the same argument with an added term for sampling error yields

$$\mathbb{E}F(\overline{\mathbf{w}}_T)-F^*=\mathcal{O}(\frac{\kappa\nu_{\max}\sigma^2/\mu}{NT}+\frac{\kappa E^2G^2/\mu}{KT}+\frac{\kappa^2E^2G^2/\mu}{T^2})$$

$\square$

F.1.1 DEFERRED PROOFS OF KEY LEMMAS

Proof of lemma 8. The proof of bound for $\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2$ in the Nesterov accelerated FedAvg follows a similar logic as in Lemma 5, but requires extra reasoning. Since communication is done every E steps, for any $t \geq 0$, we can find a $t_0 \leq t$ such that $t - t_0 \leq E - 1$ and $w_{t_0}^k = \overline{\mathbf{w}}_{t_0}$for all k. Moreover, using α_t is non-increasing, $\alpha_{t_0} \leq 2\alpha_t$, and $\beta_t \leq \alpha_t$ for any $t - t_0 \leq E - 1$, we have

$$
\begin{aligned}
\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 &= \mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_{t_0} - (\overline{\mathbf{w}}_t - \overline{\mathbf{w}}_{t_0})\|^2 \\
&\leq \mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \overline{\mathbf{w}}_{t_0}\|^2 \\
&= \mathbb{E}\sum_{k=1}^N p_k\|\mathbf{w}_t^k - \mathbf{w}_{t_0}^k\|^2 \\
&= \mathbb{E}\sum_{k=1}^N p_k\|\sum_{i=t_0}^{t-1}\beta_i(\mathbf{v}_{i+1}^k - \mathbf{v}_i^k) - \sum_{i=t_0}^{t-1}\alpha_i\mathbf{g}_{i,k}\|^2 \\
&\leq 2\sum_{k=1}^N p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\alpha_i^2\|\mathbf{g}_{i,k}\|^2 + 2\sum_{k=1}^N p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\beta_i^2\|(\mathbf{v}_{i+1}^k - \mathbf{v}_i^k)\|^2 \\
&\leq 2\sum_{k=1}^N p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\alpha_i^2(\|\mathbf{g}_{i,k}\|^2 + \|(\mathbf{v}_{i+1}^k - \mathbf{v}_i^k)\|^2) \\
&\leq 4\sum_{k=1}^N p_k\mathbb{E}\sum_{i=t_0}^{t-1}(E-1)\alpha_i^2 G^2 \\
&\leq 4(E-1)^2\alpha_{t_0}^2 G^2 \leq 16(E-1)^2\alpha_t^2 G^2
\end{aligned}
$$

where we have used $\mathbb{E}\|\mathbf{v}_t^k - \mathbf{v}_{t-1}^k\|^2 \leq G^2$. To see this identity for appropriate α_t, β_t, note the recursion

$$
\begin{aligned}
\mathbf{v}_{t+1}^k - \mathbf{v}_t^k &= \mathbf{w}_t^k - \mathbf{w}_{t-1}^k - (\alpha_t\mathbf{g}_{t,k} - \alpha_{t-1}\mathbf{g}_{t-1,k}) \\
\mathbf{w}_{t+1}^k - \mathbf{w}_t^k &= -\alpha_t\mathbf{g}_{t,k} + \beta_t(\mathbf{v}_{t+1}^k - \mathbf{v}_t^k)
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathbf{v}_{t+1}^k - \mathbf{v}_t^k &= -\alpha_{t-1}\mathbf{g}_{t-1,k} + \beta_{t-1}(\mathbf{v}_t^k - \mathbf{v}_{t-1}^k) - (\alpha_t\mathbf{g}_{t,k} - \alpha_{t-1}\mathbf{g}_{t-1,k}) \\
&= \beta_{t-1}(\mathbf{v}_t^k - \mathbf{v}_{t-1}^k) - \alpha_t\mathbf{g}_{t,k}
\end{aligned}
$$

Since the identity $\mathbf{v}_{t+1}^k - \mathbf{v}_t^k = \beta_{t-1}(\mathbf{v}_t^k - \mathbf{v}_{t-1}^k) - \alpha_t\mathbf{g}_{t,k}$ implies

$$
\mathbb{E}\|\mathbf{v}_{t+1}^k - \mathbf{v}_t^k\|^2 \leq 2\beta_{t-1}^2\mathbb{E}\|\mathbf{v}_t^k - \mathbf{v}_{t-1}^k\|^2 + 2\alpha_t^2 G^2
$$

as long as α_t, β_{t-1} satisfy $2\beta_{t-1}^2 + 2\alpha_t^2 \leq 1/2$, we can guarantee that $\mathbb{E}\|\mathbf{v}_t^k - \mathbf{v}_{t-1}^k\|^2 \leq G^2$ for all k by induction. This together with Jensen's inequality also gives $\mathbb{E}\|\overline{\mathbf{v}}_t - \overline{\mathbf{v}}_{t-1}\|^2 \leq G^2$ for all t. □

Now we are ready to prove the one step progress result for Nesterov accelerated FedAvg. The first part of the proof is identical to that of the FedAvg case, while the main recursion takes a different form.

Proof of lemma 7. We again have

$$
\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 = \|(\overline{\mathbf{w}}_t - \alpha_t\mathbf{g}_t) - \mathbf{w}^*\|^2
$$

and using exactly the same derivation as the FedAvg case, we can obtain the following bound (same as Eq (10) in the proof of Lemma 3):

$$
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\mu\alpha_t)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^2\sum_{k=1}^N p_k^2\sigma_k^2
$$

$$+ \alpha_t^2 L^2 \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^3 L \mathbb{E}\|\mathbf{g}_t\|^2 - \alpha_t^2 \|\nabla F(\overline{\mathbf{w}}_t)\|^2$$

Different from the FedAvg case, we no longer have $\overline{\mathbf{w}}_t = \overline{\mathbf{v}}_t$. Instead,

$$\begin{aligned}
\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &= \|\overline{\mathbf{v}}_t + \beta_{t-1}(\overline{\mathbf{v}}_t - \overline{\mathbf{v}}_{t-1}) - \mathbf{w}^*\|^2 \\
&= \|(1+\beta_{t-1})(\overline{\mathbf{v}}_t - \mathbf{w}^*) - \beta_{t-1}(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \\
&= (1+\beta_{t-1})^2 \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 - 2\beta_{t-1}(1+\beta_{t-1})\langle \overline{\mathbf{v}}_t - \mathbf{w}^*, \overline{\mathbf{v}}_{t-1} - \mathbf{w}^* \rangle + \beta_{t-1}^2 \|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \\
&\leq (1+\beta_{t-1})^2 \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 2\beta_{t-1}(1+\beta_{t-1})\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\| + \beta_{t-1}^2 \|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2
\end{aligned}$$

which gives a recursion involving both $\overline{\mathbf{v}}_t$ and $\overline{\mathbf{v}}_{t-1}$:

$$\begin{aligned}
\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \leq{}& (1-\alpha_t\mu)(1+\beta_{t-1})^2 \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 2(1-\alpha_t\mu)\beta_{t-1}(1+\beta_{t-1})\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\| + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 \\
&+ \beta_{t-1}^2(1-\alpha_t\mu)\|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 + \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^2 L^2 \sum_k p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \alpha_t^3 L G^2
\end{aligned}$$

and we will using this recursive relation to obtain the desired bound.

We can check that our choice of α_t and β_t satisfy α_t is non-increasing, $\alpha_t \leq 2\alpha_{t+E}$, and $2\beta_{t-1}^2 + 2\alpha_t^2 \leq 1/2$ for all $t \geq 0$, so that we can apply the bound from Lemma 8 on $\mathbb{E}\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2$ to conclude that, with $\nu_{\max} := N \cdot \max_k p_k$,

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \leq{}& \mathbb{E}(1-\mu\alpha_t)(1+\beta_{t-1})^2 \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 16E^2 L \alpha_t^3 G^2 + 16E^2 L^2 \alpha_t^4 G^2 + \alpha_t^3 L G^2 \\
&+ (1-\alpha_t\mu)\beta_{t-1}^2 \|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 + \alpha_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 + 2\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\| \\
\leq{}& \mathbb{E}(1-\mu\alpha_t)(1+\beta_{t-1})^2 \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2 + 20E^2 L \alpha_t^3 G^2 + (1-\alpha_t\mu)\beta_{t-1}^2 \|(\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*)\|^2 \\
&+ \alpha_t^2 \frac{1}{N} \nu_{\max} \sigma^2 + 2\beta_{t-1}(1+\beta_{t-1})(1-\alpha_t\mu)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\| \cdot \|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\|
\end{aligned}$$

where we have used $\sigma^2 = \sum_k p_k \sigma_k^2$, and by construction our α_t satisfies $L\alpha_t \leq \frac{1}{5}$. □

F.2 Convex Smooth Objectives

In this section we provide proof of the convergence result for Nesterov accelerated FedAvg with convex and smooth objectives. Unlike with the FedAvg algorithm, where convex and strongly convex results share identical components, the proof for the convergence result in the convex setting for Nesterov FedAvg uses a change of variables, although the general ideas are in the same vein: we have a one step progress bound for $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \eta_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*))$, which is then used to form a telescoping sum that gives an upper bound on $\min_{t \leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)$.

Lemma 9 (One step progress, convex case, Nesterov). *Let $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k \mathbf{w}_t^k$ in Nesterov accelerated FedAvg, and define $\eta_t = \frac{\alpha_t}{1-\beta_t}$. Under assumptions 1,3,4, the following bound holds for all t:*

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \eta_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 32LE^2\alpha_t^2\eta_t G^2 + \eta_t^2 \nu_{\max} \frac{1}{N}\sigma^2 + 2\eta_t \frac{\beta_t^2}{1-\beta_t} G^2$$

Theorem 4. *Set learning rates $\alpha_t = \beta_t = \mathcal{O}(\sqrt{\frac{N}{T}})$. Then under Assumptions 1,3,4 Nesterov accelerated FedAvg with full device participation has rate*

$$\min_{t \leq T} F(\overline{\mathbf{w}}_t) - F^* = \mathcal{O}\left(\frac{\nu_{\max}\sigma^2}{\sqrt{NT}} + \frac{NE^2LG^2}{T}\right),$$

and with partial device participation with K sampled devices at each communication round,

$$\min_{t \leq T} F(\overline{\mathbf{w}}_t) - F^* = \mathcal{O}\left(\frac{\nu_{\max}\sigma^2}{\sqrt{KT}} + \frac{E^2G^2}{\sqrt{KT}} + \frac{KE^2LG^2}{T}\right).$$

Proof. Applying the bound from Lemma 9, with $\eta_t = \frac{\alpha_t}{1-\beta_t}$ we have

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \eta_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 32LE^2\alpha_t^2\eta_t G^2 + \eta_t^2\nu_{\max}\frac{1}{N}\sigma^2 + 2\eta_t\frac{\beta_t^2}{1-\beta_t}G^2$$

Summing the inequalities from $t = 0$ to $t = T$, we obtain

$$\sum_{t=0}^{T}\eta_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^{T}\eta_t^2 \cdot \frac{1}{N}\nu_{\max}\sigma^2 + \sum_{t=0}^{T}\eta_t\alpha_t^2 \cdot 32LE^2G^2 + \sum_{t=0}^{T}2\eta_t\frac{\beta_t^2}{1-\beta_t}G^2$$

so that

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) \leq \frac{1}{\sum_{t=0}^{T}\eta_t}\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^{T}\eta_t^2 \cdot \frac{1}{N}\nu_{\max}\sigma^2 + \sum_{t=0}^{T}\eta_t\alpha_t^2 \cdot 32LE^2G^2 + \sum_{t=0}^{T}2\eta_t\frac{\beta_t^2}{1-\beta_t}G^2\right)$$

By setting the constant learning rates $\alpha_t \equiv \sqrt{\frac{N}{T}}$ and $\beta_t \equiv c\sqrt{\frac{N}{T}}$ so that $\eta_t = \frac{\alpha_t}{1-\beta_t} = \frac{\sqrt{\frac{N}{T}}}{1-c\sqrt{\frac{N}{T}}} \leq 2\sqrt{\frac{N}{T}}$, we have

$$\begin{aligned}
&\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) \\
&\leq \frac{1}{2\sqrt{NT}} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{2}{\sqrt{NT}}T \cdot \frac{N}{T} \cdot \frac{1}{N}\nu_{\max}\sigma^2 + \frac{1}{\sqrt{NT}}T(\sqrt{\frac{N}{T}})^3 32LE^2G^2 + \frac{2}{\sqrt{NT}}T(\sqrt{\frac{N}{T}})^3G^2 \\
&= (\frac{1}{2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + 2\nu_{\max}\sigma^2)\frac{1}{\sqrt{NT}} + \frac{N}{T}(32LE^2G^2 + 2G^2) \\
&= O(\frac{\nu_{\max}\sigma^2}{\sqrt{NT}} + \frac{NE^2LG^2}{T})
\end{aligned}$$

Similarly, for partial participation, we have

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) \leq \frac{1}{\sum_{t=0}^{T}\alpha_t}\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sum_{t=0}^{T}\alpha_t^2 \cdot (\frac{1}{N}\nu_{\max}\sigma^2 + C) + \sum_{t=0}^{T}\alpha_t^3 \cdot 6E^2LG^2\right)$$

where $C = \frac{4}{K}E^2G^2$ or $\frac{N-K}{N-1}\frac{4}{K}E^2G^2$, so that with $\alpha_t \equiv \sqrt{\frac{K}{T}}$ and $\beta_t \equiv c\sqrt{\frac{K}{T}}$, we have

$$\min_{t\leq T} F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*) = \mathcal{O}(\frac{\nu_{\max}\sigma^2}{\sqrt{KT}} + \frac{E^2G^2}{\sqrt{KT}} + \frac{KE^2LG^2}{T})$$

□

F.2.1 DEFERRED PROOFS OF KEY LEMMAS

Proof of lemma 9. Define $\overline{\mathbf{p}}_t := \frac{\beta_t}{1-\beta_t}[\overline{\mathbf{w}}_t - \overline{\mathbf{w}}_{t-1} + \alpha_t\mathbf{g}_{t-1}] = \frac{\beta_t^2}{1-\beta_t}(\overline{\mathbf{v}}_t - \overline{\mathbf{v}}_{t-1})$ for $t \geq 1$ and 0 for $t = 0$. We can check that

$$\overline{\mathbf{w}}_{t+1} + \overline{\mathbf{p}}_{t+1} = \overline{\mathbf{w}}_t + \overline{\mathbf{p}}_t - \frac{\alpha_t}{1-\beta_t}\mathbf{g}_t$$

Now we define $\overline{\mathbf{z}}_t := \overline{\mathbf{w}}_t + \overline{\mathbf{p}}_t$ and $\eta_t = \frac{\alpha_t}{1-\beta_t}$ for all t, so that we have the recursive relation

$$\overline{\mathbf{z}}_{t+1} = \overline{\mathbf{z}}_t - \eta_t\mathbf{g}_t$$

Now

$$\begin{aligned}
\|\overline{\mathbf{z}}_{t+1} - \mathbf{w}^*\|^2 &= \|(\overline{\mathbf{z}}_t - \eta_t\mathbf{g}_t) - \mathbf{w}^*\|^2 \\
&= \|(\overline{\mathbf{z}}_t - \eta_t\overline{\mathbf{g}}_t - \mathbf{w}^*) - \eta_t(\mathbf{g}_t - \overline{\mathbf{g}}_t)\|^2 \\
&= A_1 + A_2 + A_3
\end{aligned}$$

where

$$
\begin{aligned}
A_1 &= \|\overline{\mathbf{z}}_t - \mathbf{w}^* - \eta_t \overline{\mathbf{g}}_t\|^2 \\
A_2 &= 2\eta_t \langle \overline{\mathbf{z}}_t - \mathbf{w}^* - \eta_t \overline{\mathbf{g}}_t, \overline{\mathbf{g}}_t - \mathbf{g}_t \rangle \\
A_3 &= \eta_t^2 \|\mathbf{g}_t - \overline{\mathbf{g}}_t\|^2
\end{aligned}
$$

where again $\mathbb{E}A_2 = 0$ and $\mathbb{E}A_3 \leq \eta_t^2 \sum_k p_k^2 \sigma_k^2$. For A_1 we have

$$\|\overline{\mathbf{z}}_t - \mathbf{w}^* - \eta_t \overline{\mathbf{g}}_t\|^2 = \|\overline{\mathbf{z}}_t - \mathbf{w}^*\|^2 + 2\langle \overline{\mathbf{z}}_t - \mathbf{w}^*, -\eta_t \overline{\mathbf{g}}_t \rangle + \|\eta_t \overline{\mathbf{g}}_t\|^2$$

Using the convexity and L-smoothness of F_k,

$$
\begin{aligned}
&-2\eta_t \langle \overline{\mathbf{z}}_t - \mathbf{w}^*, \overline{\mathbf{g}}_t \rangle \\
&= -2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \mathbf{w}^*, \nabla F_k(\mathbf{w}_t^k) \rangle \\
&= -2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k) \rangle - 2\eta_t \sum_{k=1}^N p_k \langle \mathbf{w}_t^k - \mathbf{w}^*, \nabla F_k(\mathbf{w}_t^k) \rangle \\
&= -2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle - 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k) \rangle - 2\eta_t \sum_{k=1}^N p_k \langle \mathbf{w}_t^k - \mathbf{w}^*, \nabla F_k(\mathbf{w}_t^k) \rangle \\
&\leq -2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle - 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k) \rangle + 2\eta_t \sum_{k=1}^N p_k (F_k(\mathbf{w}^*) - F_k(\mathbf{w}_t^k)) \\
&\leq 2\eta_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}_t^k) - F_k(\overline{\mathbf{w}}_t) + \frac{L}{2} \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + F_k(\mathbf{w}^*) - F_k(\mathbf{w}_t^k) \right] \\
&- 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle \\
&= \eta_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + 2\eta_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t) \right] - 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle
\end{aligned}
$$

which results in

$$
\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &+ \eta_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + 2\eta_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t) \right] \\
&+ \eta_t^2 \|\overline{\mathbf{g}}_t\|^2 + \eta_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 - 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle
\end{aligned}
$$

As before, $\|\overline{\mathbf{g}}_t\|^2 \leq 2L^2 \sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 + 4L(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*))$, so that

$$
\begin{aligned}
\eta_t^2 \|\overline{\mathbf{g}}_t\|^2 + \eta_t \sum_{k=1}^N p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t) \right] &\leq 2L^2 \eta_t^2 \sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2 + \eta_t (1 - 4\eta_t L)(F(\mathbf{w}^*) - F(\overline{\mathbf{w}}_t)) \\
&\leq 2L^2 \eta_t^2 \sum_k p_k \|\mathbf{w}_t^k - \overline{\mathbf{w}}_t\|^2
\end{aligned}
$$

for $\eta_t \leq 1/4L$. Using $\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \leq 16E^2 \alpha_t^2 G^2$ and $\sum_{k=1}^N p_k^2 \sigma_k^2 \leq \nu_{\max} \frac{1}{N} \sigma^2$, it follows that

$$
\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \eta_t (F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &+ (\eta_t L + 2L^2 \eta_t^2) \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \eta_t^2 \sum_{k=1}^N p_k^2 \sigma_k^2 \\
&- 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k) \rangle
\end{aligned}
$$

$$\leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 32LE^2\alpha_t^2\eta_t G^2 + \eta_t^2 \nu_{\max}\frac{1}{N}\sigma^2$$
$$- 2\eta_t \sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k)\rangle$$

if $\eta_t \leq \frac{1}{2L}$. It remains to bound $\mathbb{E}\sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k)\rangle$. Recall that $\overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t = \frac{\beta_t}{1-\beta_t}[\overline{\mathbf{w}}_t - \overline{\mathbf{w}}_{t-1} + \alpha_t \mathbf{g}_{t-1}] = \frac{\beta_t^2}{1-\beta_t}(\overline{\mathbf{v}}_t - \overline{\mathbf{v}}_{t-1})$ and $\mathbb{E}\|\overline{\mathbf{v}}_t - \overline{\mathbf{v}}_{t-1}\|^2 \leq G^2$, $\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k)\|^2 \leq G^2$. Cauchy-Schwarz gives

$$\mathbb{E}\sum_{k=1}^N p_k \langle \overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t, \nabla F_k(\mathbf{w}_t^k)\rangle \leq \sum_{k=1}^N p_k \sqrt{\mathbb{E}\|\overline{\mathbf{z}}_t - \overline{\mathbf{w}}_t\|^2} \cdot \sqrt{\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k)\|^2}$$
$$\leq \frac{\beta_t^2}{1-\beta_t}G^2$$

Thus

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \eta_t(F(\overline{\mathbf{w}}_t) - F(\mathbf{w}^*)) \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 32LE^2\alpha_t^2\eta_t G^2 + \eta_t^2\nu_{\max}\frac{1}{N}\sigma^2 + 2\eta_t\frac{\beta_t^2}{1-\beta_t}G^2$$

□

G GEOMETRIC CONVERGENCE OF FEDAVG IN THE OVERPARAMETERIZED SETTING

Overparameterization is a prevalent machine learning setting where the statistical model has much more parameters than the number of training samples and the existence of parameter choices with zero training loss is ensured Allen-Zhu et al. (2018); Zhang et al. (2016). Due to the property of *automatic variance reduction* in overparameterization, a line of recent works proved that SGD and accelerated methods achieve geometric convergence Ma et al. (2018); Moulines & Bach (2011); Needell et al. (2014); Schmidt & Roux (2013); Strohmer & Vershynin (2009). A natural question is whether such a result still holds in the federated learning setting. In this section, we provide the first geometric convergence rate of FedAvg for the overparameterized strongly convex and smooth problems, and show that it preserves linear speedup at the same time. We then sharpen this result in the special case of linear regression. Inspired by recent advances in accelerating SGD Liu et al. (2020); Jain et al. (2017), we further propose a novel momentum-based FedAvg algorithm, which enjoys an improved convergence rate over FedAvg. Detailed proofs are deferred to Appendix Section H. In particular, we do not need Assumptions 3 and 4 and use modified versions of Assumptions 1 and 2 detailed in this section.

G.1 GEOMETRIC CONVERGENCE OF FEDAVG IN THE OVERPARAMETERIZED SETTING

Recall the FL problem $\min_w \sum_{k=1}^N p_k F_k(\mathbf{w})$ with $F_k(\mathbf{w}) = \frac{1}{n_k}\sum_{j=1}^{n_k} \ell(\mathbf{w}; \mathbf{x}_k^j)$. In this section, we consider the standard Empirical Risk Minimization (ERM) setting where ℓ is non-negative, l-smooth, and convex, and as before, each $F_k(\mathbf{w})$ is L-smooth and μ-strongly convex. Note that $l \geq L$. This setup includes many important problems in practice. In the overparameterized setting, there exists $\mathbf{w}^* \in \arg\min_w \sum_{k=1}^N p_k F_k(\mathbf{w})$ such that $\ell(\mathbf{w}^*; \mathbf{x}_k^j) = 0$ for all $\mathbf{x}_k^j$. We first show that FedAvg achieves geometric convergence with linear speedup in the number of workers.

Theorem 5. *In the overparameterized setting, FedAvg with communication every E iterations and constant step size $\overline{\alpha} = \mathcal{O}(\frac{1}{E}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})})$ has geometric convergence:*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) \leq \frac{L}{2}(1-\overline{\alpha})^T\|\mathbf{w}_0 - \mathbf{w}^*\|^2 = \mathcal{O}\left(L\exp\left(-\frac{\mu}{E}\frac{NT}{l\nu_{\max} + L(N-\nu_{\min})}\right)\cdot\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right).$$

Linear speedup and Communication Complexity The linear speedup factor is on the order of $\mathcal{O}(N/E)$ for $N \leq \mathcal{O}(\frac{l}{L})$, i.e. FedAvg with N workers and communication every E iterations provides a geometric convergence speedup factor of $\mathcal{O}(N/E)$, for $N \leq \mathcal{O}(\frac{l}{L})$. When N is above this threshold, however, the speedup is almost constant in the number of workers. This matches the findings in Ma et al. (2018). Our result also illustrates that E can be taken $\mathcal{O}(T^{\beta})$ for any $\beta < 1$ to achieve geometric convergence, achieving better communication efficiency than the standard FL setting. We emphasize again that compared to the single-server results in Ma et al. (2018), the difference of our result lies in the factor of N in the speedup, which cannot be obtained if one simply applied the single-server result to each device in our problem.

G.2 Overparameterized Linear Regression Problems

We now turn to quadratic problems and show that the bound in Theorem 5 can be improved to $\mathcal{O}(\exp(-\frac{N}{E\kappa_1}t))$ for a larger range of N. We then propose a variant of FedAvg that has provable acceleration over FedAvg with SGD updates. The local device objectives are now given by the sum of squares $F_k(\mathbf{w}) = \frac{1}{2n_k}\sum_{j=1}^{n_k}(\mathbf{w}^T\mathbf{x}_k^j - z_k^j)^2$, and there exists $\mathbf{w}^*$ such that $F(\mathbf{w}^*) \equiv 0$. Two notions of condition number are important in our results: κ_1 which is based on local Hessians, and $\tilde{\kappa}$, which is termed the statistical condition number Liu & Belkin (2020); Jain et al. (2017). For their detailed definitions, please refer to Appendix Section H. Here we use the fact $\tilde{\kappa} \leq \kappa_1$. Recall $\nu_{\max} = \max_k p_k N$ and $\nu_{\min} = \min_k p_k N$.

Theorem 6. *For the overparamterized linear regression problem, FedAvg with communication every E iterations with constant step size $\overline{\alpha} = \mathcal{O}(\frac{1}{E}\frac{N}{l\nu_{\max}+\mu(N-\nu_{\min})})$ has geometric convergence:*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) \leq \mathcal{O}\left(L\exp(-\frac{NT}{E(\nu_{\max}\kappa_1 + (N-\nu_{\min}))})\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right).$$

When $N = \mathcal{O}(\kappa_1)$, the convergence rate is $\mathcal{O}((1-\frac{N}{E\kappa_1})^T) = \mathcal{O}(\exp(-\frac{NT}{E\kappa_1}))$, which exhibits linear speedup in the number of workers, as well as a $1/\kappa_1$ dependence on the condition number κ_1. Inspired by Liu & Belkin (2020), we propose the **MaSS accelerated FedAvg algorithm** (FedMaSS):

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{u}_t^k - \eta_1^k \mathbf{g}_{t,k} & \text{if } t+1 \notin \mathcal{I}_E, \\ \sum_{k\in\mathcal{S}_{t+1}}\left[\mathbf{u}_t^k - \eta_1^k\mathbf{g}_{t,k}\right] & \text{if } t+1 \in \mathcal{I}_E, \end{cases}$$
$$\mathbf{u}_{t+1}^k = \mathbf{w}_{t+1}^k + \gamma^k(\mathbf{w}_{t+1}^k - \mathbf{w}_t^k) + \eta_2^k\mathbf{g}_{t,k}.$$

When $\eta_2^k \equiv 0$, this algorithm reduces to the Nesterov accelerated FedAvg algorithm. In the next theorem, we demonstrate that FedMaSS improves the convergence to $\mathcal{O}(\exp(-\frac{NT}{E\sqrt{\kappa_1\tilde{\kappa}}}))$. To our knowledge, this is the first acceleration result of FedAvg with momentum updates over SGD updates.

Theorem 7. *For the overparamterized linear regression problem, FedMaSS with communication every E iterations and constant step sizes $\overline{\eta}_1 = \mathcal{O}(\frac{1}{E}\frac{N}{l\nu_{\max}+\mu(N-\nu_{\min})}), \overline{\eta}_2 = \frac{\overline{\eta}_1(1-\frac{1}{\tilde{\kappa}})}{1+\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}, \overline{\gamma} = \frac{1-\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}{1+\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}$ has geometric convergence:*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) \leq \mathcal{O}\left(L\exp(-\frac{NT}{E(\nu_{\max}\sqrt{\kappa_1\tilde{\kappa}} + (N-\nu_{\min}))})\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right).$$

Speedup of FedMaSS over FedAvg To better understand the significance of the above result, we briefly discuss related works on accelerating SGD. Nesterov and Heavy Ball updates are known to fail to accelerate over SGD in both the overparameterized and convex settings Liu & Belkin (2020); Kidambi et al. (2018); Liu et al. (2018); Yuan et al. (2016). Thus in general one cannot hope to obtain acceleration results for the FedAvg algorithm with Nesterov and Heavy Ball updates. Luckily, recent works in SGD Jain et al. (2017); Liu & Belkin (2020) introduced an additional compensation term to the Nesterov updates to address the non-acceleration issue. Surprisingly, we show the same approach can effectively improve the rate of FedAvg. Comparing the convergence rate of FedMass (Theorem 7) and FedAvg (Theorem 6), when $N = \mathcal{O}(\sqrt{\kappa_1\tilde{\kappa}})$, the convergence rate is $\mathcal{O}((1-\frac{N}{E\sqrt{\kappa_1\tilde{\kappa}}})^T) = \mathcal{O}(\exp(-\frac{NT}{E\sqrt{\kappa_1\tilde{\kappa}}}))$ as opposed to $\mathcal{O}(\exp(-\frac{NT}{E\kappa_1}))$. Since $\kappa_1 \geq \tilde{\kappa}$, this implies a speedup factor of $\sqrt{\frac{\kappa_1}{\tilde{\kappa}}}$ for FedMaSS. On the other hand, the same linear speedup in the number of workers holds for N in a smaller range of values.

H PROOF OF GEOMETRIC CONVERGENCE RESULTS FOR OVERPARAMETERIZED PROBLEMS

H.1 GEOMETRIC CONVERGENCE OF FEDAVG FOR GENERAL STRONGLY CONVEX AND SMOOTH OBJECTIVES

Theorem 5. *For the overparameterized setting with general strongly convex and smooth objectives, FedAvg with local SGD updates and communication every E iterations with constant step size $\overline{\alpha} = \frac{1}{2E}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}$ gives the exponential convergence guarantee*

$$\mathbb{E}F(\overline{\mathbf{w}}_t) \leq \frac{L}{2}(1-\mu\overline{\alpha})^t\|\mathbf{w}_0-\mathbf{w}^*\|^2 = O(\exp(-\frac{\mu}{2E}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}t)\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2)$$

Proof. To illustrate the main ideas of the proof, we first present the proof for $E=2$. Let $t-1$ be a communication round, so that $\mathbf{w}_{t-1}^k = \overline{\mathbf{w}}_{t-1}$. We show that

$$\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 \leq (1-\alpha_t\mu)(1-\alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1}-\mathbf{w}^*\|^2$$

for appropriately chosen constant step sizes α_t, α_{t-1}. We have

$$\begin{aligned}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 &= \|(\overline{\mathbf{w}}_t-\alpha_t\mathbf{g}_t)-\mathbf{w}^*\|^2\\ &= \|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2 - 2\alpha_t\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\mathbf{g}_t\rangle + \alpha_t^2\|\mathbf{g}_t\|^2\end{aligned}$$

and the cross term can be bounded as usual using μ-convexity and L-smoothness of F_k:

$$\begin{aligned}
&-2\alpha_t\mathbb{E}_t\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\mathbf{g}_t\rangle\\
&= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\nabla F_k(\mathbf{w}_t^k)\rangle\\
&= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}_t^k,\nabla F_k(\mathbf{w}_t^k)\rangle - 2\alpha_t\sum_{k=1}^N p_k\langle\mathbf{w}_t^k-\mathbf{w}^*,\nabla F_k(\mathbf{w}_t^k)\rangle\\
&\leq -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t-\mathbf{w}_t^k,\nabla F_k(\mathbf{w}_t^k)\rangle + 2\alpha_t\sum_{k=1}^N p_k(F_k(\mathbf{w}^*)-F_k(\mathbf{w}_t^k)) - \alpha_t\mu\sum_{k=1}^N p_k\|\mathbf{w}_t^k-\mathbf{w}^*\|^2\\
&\leq 2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}_t^k)-F_k(\overline{\mathbf{w}}_t)+\frac{L}{2}\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2+F_k(\mathbf{w}^*)-F_k(\mathbf{w}_t^k)\right] - \alpha_t\mu\|\sum_{k=1}^N p_k(\mathbf{w}_t^k-\mathbf{w}^*)\|^2\\
&= \alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2 + 2\alpha_t\sum_{k=1}^N p_k\left[F_k(\mathbf{w}^*)-F_k(\overline{\mathbf{w}}_t)\right] - \alpha_t\mu\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2\\
&= \alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2 - 2\alpha_t\sum_{k=1}^N p_k F_k(\overline{\mathbf{w}}_t) - \alpha_t\mu\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2
\end{aligned}$$

and so

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 \leq \mathbb{E}(1-\alpha_t\mu)\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2 - 2\alpha_t F(\overline{\mathbf{w}}_t) + \alpha_t^2\|\mathbf{g}_t\|^2 + \alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2$$

Applying this recursive relation to $\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2$ and using $\|\overline{\mathbf{w}}_{t-1}-\mathbf{w}_{t-1}^k\|^2\equiv 0$, we further obtain

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2 \leq{}& \mathbb{E}(1-\alpha_t\mu)\left((1-\alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1}-\mathbf{w}^*\|^2 - 2\alpha_{t-1}F(\overline{\mathbf{w}}_{t-1}) + \alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2\right)\\ &- 2\alpha_t F(\overline{\mathbf{w}}_t) + \alpha_t^2\|\mathbf{g}_t\|^2 + \alpha_t L\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2\end{aligned}$$

Now instead of bounding $\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2$ using the arguments in the general convex case, we follow Ma et al. (2018) and use the fact that in the overparameterized setting, $\mathbf{w}^*$ is a minimizer of each $\ell(\mathbf{w}, x_k^j)$ and that each ℓ is l-smooth to obtain $\|\nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 \leq 2l(F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k) - F_k(\mathbf{w}^*, \xi_{t-1}^k))$, where recall $F_k(\mathbf{w}, \xi_{t-1}^k) = \ell(\mathbf{w}, \xi_{t-1}^k)$, so that

$$\begin{aligned}
\sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 &= \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_{t-1} - \alpha_{t-1}\mathbf{g}_{t-1} - \mathbf{w}_{t-1}^k + \alpha_{t-1}\mathbf{g}_{t-1,k}\|^2 \\
&= \sum_{k=1}^N p_k \alpha_{t-1}^2 \|\mathbf{g}_{t-1} - \mathbf{g}_{t-1,k}\|^2 \\
&= \alpha_{t-1}^2 \sum_{k=1}^N p_k (\|\mathbf{g}_{t-1,k}\|^2 - \|\mathbf{g}_{t-1}\|^2) \\
&= \alpha_{t-1}^2 \sum_{k=1}^N p_k \|\nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 - \alpha_{t-1}^2 \|\mathbf{g}_{t-1}\|^2 \\
&\leq \alpha_{t-1}^2 \sum_{k=1}^N p_k 2l(F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k) - F_k(\mathbf{w}^*, \xi_{t-1}^k)) - \alpha_{t-1}^2 \|\mathbf{g}_{t-1}\|^2
\end{aligned}$$

again using $\overline{\mathbf{w}}_{t-1} = \mathbf{w}_{t-1}^k$. Taking expectation with respect to ξ_{t-1}^k's and using the fact that $F(\mathbf{w}^*) = 0$, we have

$$\begin{aligned}
\mathbb{E}_{t-1} \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 &\leq 2l\alpha_{t-1}^2 \sum_{k=1}^N p_k F_k(\overline{\mathbf{w}}_{t-1}) - \alpha_{t-1}^2 \|\mathbf{g}_{t-1}\|^2 \\
&= 2l\alpha_{t-1}^2 F(\overline{\mathbf{w}}_{t-1}) - \alpha_{t-1}^2 \|\mathbf{g}_{t-1}\|^2
\end{aligned}$$

Note also that

$$\|\mathbf{g}_{t-1}\|^2 = \|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2$$

while

$$\begin{aligned}
\|\mathbf{g}_t\|^2 = \|\sum_{k=1}^N p_k \nabla F_k(\mathbf{w}_t^k, \xi_t^k)\|^2 &\leq 2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + 2\|\sum_{k=1}^N p_k (\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k) - \nabla F_k(\mathbf{w}_t^k, \xi_t^k))\|^2 \\
&\leq 2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + 2\sum_{k=1}^N p_k l^2 \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2
\end{aligned}$$

Substituting these into the bound for $\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &\leq \mathbb{E}(1 - \alpha_t\mu)((1 - \alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 - 2\alpha_{t-1}F(\overline{\mathbf{w}}_{t-1}) + \alpha_{t-1}^2 \|\mathbf{g}_{t-1}\|^2) \\
&\quad - 2\alpha_t F(\overline{\mathbf{w}}_t) + 2\alpha_t^2 \|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + \left(2l^2\alpha_{t-1}^2\alpha_t^2 + \alpha_t\alpha_{t-1}^2 L\right)\left(2lF(\overline{\mathbf{w}}_{t-1}) - \|\mathbf{g}_{t-1}\|^2\right) \\
&= \mathbb{E}(1 - \alpha_t\mu)(1 - \alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 \\
&\quad - 2\alpha_t(F(\overline{\mathbf{w}}_t) - \alpha_t\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) \\
&\quad - 2\alpha_{t-1}(1 - \alpha_t\mu)\left((1 - \frac{l\alpha_{t-1}(2l^2\alpha_t^2 + \alpha_t L)}{1 - \alpha_t\mu})F(\overline{\mathbf{w}}_{t-1}) - \frac{\alpha_{t-1}}{2}\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2\right)
\end{aligned}$$

from which we can conclude that

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq (1 - \alpha_t\mu)(1 - \alpha_{t-1}\mu)\mathbb{E}\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2$$

if we can choose α_t, α_{t-1} to guarantee

$$\mathbb{E}(F(\overline{\mathbf{w}}_t) - \alpha_t \| \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) \geq 0$$
$$\mathbb{E}\left((1 - \frac{l\alpha_{t-1}(2l^2\alpha_t^2 + \alpha_t L)}{1-\alpha_t\mu})F(\overline{\mathbf{w}}_{t-1}) - \frac{\alpha_{t-1}}{2}\| \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2\right) \geq 0$$

Note that

$$\begin{aligned}
\mathbb{E}_t\| \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 &= \mathbb{E}_t \langle \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k), \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + \sum_{k=1}^N \sum_{j\neq k} p_j p_k \mathbb{E}_t \langle \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k), \nabla F_j(\overline{\mathbf{w}}_t, \xi_t^j)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + \sum_{k=1}^N \sum_{j\neq k} p_j p_k \langle \nabla F_k(\overline{\mathbf{w}}_t), \nabla F_j(\overline{\mathbf{w}}_t)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + \sum_{k=1}^N \sum_{j=1}^N p_j p_k \langle \nabla F_k(\overline{\mathbf{w}}_t), \nabla F_j(\overline{\mathbf{w}}_t)\rangle - \sum_{k=1}^N p_k^2 \|\nabla F_k(\overline{\mathbf{w}}_t)\|^2 \\
&\leq \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + \| \sum_k p_k \nabla F_k(\overline{\mathbf{w}}_t)\|^2 - \frac{1}{N}\nu_{\min} \| \sum_k p_k \nabla F_k(\overline{\mathbf{w}}_t)\|^2 \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + (1 - \frac{1}{N}\nu_{\min})\|\nabla F(\overline{\mathbf{w}}_t)\|^2
\end{aligned}$$

and so following Ma et al. (2018) if we let $\alpha_t = \min\{\frac{qN}{2l\nu_{\max}}, \frac{1-q}{2L(1-\frac{1}{N}\nu_{\min})}\}$ for a $q \in [0,1]$ to be optimized later, we have

$$\begin{aligned}
&\mathbb{E}_t(F(\overline{\mathbf{w}}_t) - \alpha_t \| \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) \\
&\geq \mathbb{E}_t \sum_{k=1}^N p_k F_k(\overline{\mathbf{w}}_t) - \alpha_t \left[\sum_{k=1}^N p_k^2 \mathbb{E}_t \|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + (1 - \frac{1}{N}\nu_{\min})\|\nabla F(\overline{\mathbf{w}}_t)\|^2\right] \\
&\geq \mathbb{E}_t \sum_{k=1}^N p_k (qF_k(\overline{\mathbf{w}}_t, \xi_t^k) - \alpha_t \frac{1}{N}\nu_{\max}\|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) + ((1-q)F(\overline{\mathbf{w}}_t) - \alpha_t(1 - \frac{1}{N}\nu_{\min})\|\nabla F(\overline{\mathbf{w}}_t)\|^2) \\
&\geq q\mathbb{E}_t \sum_{k=1}^N p_k (F_k(\overline{\mathbf{w}}_t, \xi_t^k) - \frac{1}{2l}\|\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) + (1-q)(F(\overline{\mathbf{w}}_t) - \frac{1}{2L}\|\nabla F(\overline{\mathbf{w}}_t)\|^2) \\
&\geq 0
\end{aligned}$$

again using $\mathbf{w}^*$ optimizes $F_k(\mathbf{w}, \xi_t^k)$ with $F_k(\mathbf{w}^*, \xi_t^k) = 0$.

Maximizing $\alpha_t = \min\{\frac{qN}{2l\nu_{\max}}, \frac{1-q}{2L(1-\frac{1}{N}\nu_{\min})}\}$ over $q \in [0,1]$, we see that $q = \frac{l\nu_{\max}}{l\nu_{\max} + L(N-\nu_{\min})}$ results in the fastest convergence, and this translates to $\alpha_t = \frac{1}{2}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}$. Next we claim that $\alpha_{t-1} = c\frac{1}{2}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}$ also guarantees

$$\mathbb{E}(1 - \frac{l\alpha_{t-1}(2l^2\alpha_t^2 + \alpha_t L)}{1-\alpha_t\mu})F(\overline{\mathbf{w}}_{t-1}) - \frac{\alpha_{t-1}}{2}\| \sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 \geq 0$$

Note that by scaling α_{t-1} by a constant $c \leq 1$ if necessary, we can guarantee $\frac{l\alpha_{t-1}(2l^2\alpha_t^2+\alpha_t L)}{1-\alpha_t\mu} \leq \frac{1}{2}$, and so the condition is equivalent to

$$F(\overline{\mathbf{w}}_{t-1}) - \alpha_{t-1}\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 \geq 0$$

which was shown to hold with $\alpha_{t-1} \leq \frac{1}{2}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}$.

For the proof of general $E \geq 2$, we use the following two identities:

$$\|\mathbf{g}_t\|^2 \leq 2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 + 2\sum_{k=1}^N p_k l^2 \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2$$

$$\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \leq \mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 8\alpha_{t-1}^2 lF(\overline{\mathbf{w}}_{t-1}) - 2\alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2$$

where the first inequality has been established before. To establish the second inequality, note that

$$\begin{aligned}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 &= \sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \alpha_{t-1}\mathbf{g}_{t-1} - \mathbf{w}_{t-1}^k + \alpha_{t-1}\mathbf{g}_{t-1,k}\|^2 \\ &\leq 2\sum_{k=1}^N p_k\left(\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + \|\alpha_{t-1}\mathbf{g}_{t-1} - \alpha_{t-1}\mathbf{g}_{t-1,k}\|^2\right)\end{aligned}$$

and

$$\begin{aligned}&\sum_k p_k\|\mathbf{g}_{t-1,k} - \mathbf{g}_{t-1}\|^2 = \sum_k p_k(\|\mathbf{g}_{t-1,k}\|^2 - \|\mathbf{g}_{t-1}\|^2) \\ &= \sum_k p_k\|\nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k) + \nabla F_k(\mathbf{w}_{t-1}^k, \xi_{t-1}^k) - \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 - \|\mathbf{g}_{t-1}\|^2 \\ &\leq 2\sum_k p_k\left(\|\nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 + l^2\|\mathbf{w}_{t-1}^k - \overline{\mathbf{w}}_{t-1}\|^2\right) - \|\mathbf{g}_{t-1}\|^2\end{aligned}$$

so that using the l-smoothness of ℓ,

$$\begin{aligned}&\mathbb{E}\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \\ &\leq \mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 4\alpha_{t-1}^2\sum_k p_k\|\nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 - 2\alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2 \\ &\leq \mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 4\alpha_{t-1}^2 2l\sum_k p_k(F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k) - F_k(\mathbf{w}^*, \xi_{t-1}^k)) - 2\alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2 \\ &= \mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 8\alpha_{t-1}^2 lF(\overline{\mathbf{w}}_{t-1}) - 2\alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2\end{aligned}$$

Using the first inequality, we have

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &\leq \mathbb{E}(1-\alpha_t\mu)\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 \\ &\quad - 2\alpha_t F(\overline{\mathbf{w}}_t) + 2\alpha_t^2\|\sum_{k=1}^N p_k\nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2 \\ &\quad + (2\alpha_t^2 l^2 + \alpha_t L)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2\end{aligned}$$

and we choose α_t and α_{t-1} such that $\mathbb{E}(F(\overline{\mathbf{w}}_t) - \alpha_t\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_t, \xi_t^k)\|^2) \geq 0$ and $(2\alpha_t^2 l^2 + \alpha_t L) \leq (1-\alpha_t\mu)(2\alpha_{t-1}^2 l^2 + \alpha_{t-1}L)/3$. This gives

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\alpha_t\mu)[(1-\alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 - 2\alpha_{t-1}F(\overline{\mathbf{w}}_{t-1}) + 2\alpha_{t-1}^2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2$$
$$+ (2\alpha_{t-1}^2 l^2 + \alpha_{t-1}L)(\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + \sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2)/3]$$

Using the second inequality

$$\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \leq \mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 8\alpha_{t-1}^2 lF(\overline{\mathbf{w}}_{t-1}) - 2\alpha_{t-1}^2\|\mathbf{g}_{t-1}\|^2$$

and that $2(1+2l^2\alpha_{t-1}^2) \leq 3$, $2\alpha_{t-1}^2 l^2 + \alpha_{t-1}L \leq 1$, we have

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq{}& \mathbb{E}(1-\alpha_t\mu)[(1-\alpha_{t-1}\mu)\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 \\
& - 2\alpha_{t-1}F(\overline{\mathbf{w}}_{t-1}) + 2\alpha_{t-1}^2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 + 8\alpha_{t-1}^2 lF(\overline{\mathbf{w}}_{t-1}) \\
& + (2\alpha_{t-1}^2 l^2 + \alpha_{t-1}L)(2\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2)]
\end{aligned}$$

and if α_{t-1} is chosen such that

$$(F(\overline{\mathbf{w}}_{t-1}) - 4\alpha_{t-1}lF(\overline{\mathbf{w}}_{t-1})) - \alpha_{t-1}\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-1}, \xi_{t-1}^k)\|^2 \geq 0$$

and

$$(2\alpha_{t-1}^2 l^2 + \alpha_{t-1}L)(1-\alpha_{t-1}\mu) \leq (2\alpha_{t-2}^2 l^2 + \alpha_{t-2}L)/3$$

we again have

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\alpha_t\mu)(1-\alpha_{t-1}\mu)[\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 + (2\alpha_{t-2}^2 l^2 + \alpha_{t-2}L)\cdot(2\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2)/3]$$

Applying the above derivation iteratively $\tau < E$ times, we have

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq{}& \mathbb{E}(1-\alpha_t\mu)\cdots(1-\alpha_{t-\tau+1}\mu)[(1-\alpha_{t-\tau}\mu)\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\|^2 \\
& - 2\alpha_{t-\tau}F(\overline{\mathbf{w}}_{t-\tau}) + 2\alpha_{t-\tau}^2\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-\tau}, \xi_{t-\tau}^k)\|^2 + 8\tau\alpha_{t-\tau}^2 lF(\overline{\mathbf{w}}_{t-\tau}) \\
& + (2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L)((\tau+1)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}_{t-\tau}^k\|^2)]
\end{aligned}$$

as long as the step sizes $\alpha_{t-\tau}$ are chosen such that the following inequalities hold

$$(2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L)(1-\alpha_{t-\tau}\mu) \leq (2\alpha_{t-\tau-1}^2 l^2 + \alpha_{t-\tau-1}L)/3$$
$$2(1+2l^2\alpha_{t-\tau}^2) \leq 3$$
$$2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L \leq 1$$
$$(F(\overline{\mathbf{w}}_{t-\tau}) - 4\tau\alpha_{t-\tau}lF(\overline{\mathbf{w}}_{t-\tau})) - \alpha_{t-\tau}\|\sum_{k=1}^N p_k \nabla F_k(\overline{\mathbf{w}}_{t-\tau}, \xi_{t-\tau}^k)\|^2 \geq 0$$

We can check that setting $\alpha_{t-\tau} = c\frac{1}{\tau+1}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})}$ for some small constant c satisfies the requirements.

Since communication is done every E iterations, $\overline{\mathbf{w}}_{t_0} = \mathbf{w}_{t_0}^k$ for some $t_0 > t - E$, from which we can conclude that

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &\leq (\prod_{\tau=1}^{t-t_0-1}(1-\mu\alpha_{t-\tau}))\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2 \\ &\leq (1 - c\frac{\mu}{E}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})})^{t-t_0}\|\mathbf{w}_{t_0} - \mathbf{w}^*\|^2\end{aligned}$$

and applying this inequality to iterations between each communication round,

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &\leq (1 - c\frac{\mu}{E}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})})^{t}\|\mathbf{w}_{0} - \mathbf{w}^*\|^2 \\ &= O(\exp(\frac{\mu}{E}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})}t))\|\mathbf{w}_{0} - \mathbf{w}^*\|^2\end{aligned}$$

With partial participation, we note that

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \\ &= \mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2 + \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \\ &= \frac{1}{K}\sum_k p_k\mathbb{E}\|\mathbf{w}_{t+1}^k - \overline{\mathbf{w}}_{t+1}\|^2 + \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2\end{aligned}$$

and so the recursive identity becomes

$$\begin{aligned}\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}(1-\alpha_t\mu)\cdots(1-\alpha_{t-\tau+1}\mu)[&(1-\alpha_{t-\tau}\mu)\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\|^2 \\ &- 2\alpha_{t-\tau}F(\overline{\mathbf{w}}_{t-\tau}) + 2\alpha_{t-\tau}^2\|\sum_{k=1}^N p_k\nabla F_k(\overline{\mathbf{w}}_{t-\tau}, \xi_{t-\tau}^k)\|^2 + 8\tau\alpha_{t-\tau}^2 lF(\overline{\mathbf{w}}_{t-\tau}) \\ &+ (2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L + \frac{1}{K})((\tau+1)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}_{t-\tau}^k\|^2)]\end{aligned}$$

which requires

$$\begin{aligned}(2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L + \frac{1}{K})(1-\alpha_{t-\tau}\mu) &\leq (2\alpha_{t-\tau-1}^2 l^2 + \alpha_{t-\tau-1}L + \frac{1}{K})/3 \\ 2(1+2l^2\alpha_{t-\tau}^2) &\leq 3 \\ 2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L + \frac{1}{K} &\leq 1 \\ (F(\overline{\mathbf{w}}_{t-\tau}) - 4\tau\alpha_{t-\tau}lF(\overline{\mathbf{w}}_{t-\tau})) - \alpha_{t-\tau}\|\sum_{k=1}^N p_k\nabla F_k(\overline{\mathbf{w}}_{t-\tau}, \xi_{t-\tau}^k)\|^2 &\geq 0\end{aligned}$$

to hold. Again setting $\alpha_{t-\tau} = c\frac{1}{\tau+1}\frac{N}{l\nu_{\max}+L(N-\nu_{\min})}$ for a possibly different constant from before satisfies the requirements.

Finally, using the L-smoothness of F,

$$F(\overline{\mathbf{w}}_T) - F(\mathbf{w}^*) \leq \frac{L}{2}\mathbb{E}\|\overline{\mathbf{w}}_T - \mathbf{w}^*\|^2 = O(L\exp(-\frac{\mu}{E}\frac{N}{l\nu_{\max} + L(N-\nu_{\min})}T))\|\mathbf{w}_0 - \mathbf{w}^*\|^2$$

□

H.2 Geometric Convergence of FedAvg for Overparameterized Linear Regression

We first provide details on quantities used in the proof of results on linear regression in Section G. The local device objectives are now given by the sum of squares $F_k(\mathbf{w}) = \frac{1}{2n_k}\sum_{j=1}^{n_k}(\mathbf{w}^T\mathbf{x}_k^j - \mathbf{z}_k^j)^2$, and there exists $\mathbf{w}^*$ such that $F(\mathbf{w}^*) \equiv 0$. Define the local Hessian matrix as $\mathbf{H}^k := \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbf{x}_k^j(\mathbf{x}_k^j)^T$, and the stochastic Hessian matrix as $\tilde{\mathbf{H}}_t^k := \xi_t^k(\xi_t^k)^T$, where ξ_t^k is the stochastic sample on the kth device at time t. Define l to be the smallest positive number such that $\mathbb{E}\|\xi_t^k\|^2\xi_t^k(\xi_t^k)^T \preceq l\mathbf{H}^k$ for all k. Note that $l \leq \max_{k,j}\|\mathbf{x}_k^j\|^2$. Let L and μ be lower and upper bounds of non-zero eigenvalues of $\mathbf{H}^k$. Define $\kappa_1 := l/\mu$ and $\kappa := L/\mu$.

Following Liu & Belkin (2020); Jain et al. (2017), we define the statistical condition number $\tilde{\kappa}$ as the smallest positive real number such that $\mathbb{E}\sum_k p_k\tilde{\mathbf{H}}_t^k\mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k \leq \tilde{\kappa}\mathbf{H}$. The condition numbers κ_1 and $\tilde{\kappa}$ are important in the characterization of convergence rates for FedAvg algorithms. Note that $\kappa_1 > \kappa$ and $\kappa_1 > \tilde{\kappa}$.

Let $\mathbf{H} = \sum_k p_k\mathbf{H}^k$. In general $\mathbf{H}$ has zero eigenvalues. However, because the null space of $\mathbf{H}$ and range of $\mathbf{H}$ are orthogonal, in our subsequence analysis it suffices to project $\overline{\mathbf{w}}_t - \mathbf{w}^*$ onto the range of $\mathbf{H}$, thus we may restrict to the non-zero eigenvalue of $\mathbf{H}$.

A useful observation is that we can use $\mathbf{w}^{*T}\mathbf{x}_k^j - \mathbf{z}_k^j \equiv 0$ to rewrite the local objectives as $F_k(\mathbf{w}) = \frac{1}{2}\langle\mathbf{w}-\mathbf{w}^*, \mathbf{H}^k(\mathbf{w}-\mathbf{w}^*)\rangle \equiv \frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{H}^k}^2$:

$$\begin{aligned}F_k(\mathbf{w}) &= \frac{1}{2n_k}\sum_{j=1}^{n_k}(\mathbf{w}^T\mathbf{x}_{k,j} - \mathbf{z}_{k,j} - (\mathbf{w}^{*T}\mathbf{x}_{k,j} - \mathbf{z}_{k,j}))^2 = \frac{1}{2n_k}\sum_{j=1}^{n_k}((\mathbf{w}-\mathbf{w}^*)^T\mathbf{x}_{k,j})^2 \\ &= \frac{1}{2}\langle\mathbf{w}-\mathbf{w}^*, \mathbf{H}^k(\mathbf{w}-\mathbf{w}^*)\rangle = \frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{H}^k}^2\end{aligned}$$

so that $F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|_H^2$.

Finally, note that $\mathbb{E}\tilde{\mathbf{H}}_t^k = \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbf{x}_k^j(\mathbf{x}_k^j)^T = \mathbf{H}^k$ and $\mathbf{g}_{t,k} = \nabla F_k(\mathbf{w}_t^k, \xi_t^k) = \tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \mathbf{w}^*)$ while $\mathbf{g}_t = \sum_{k=1}^N p_k\nabla F_k(\mathbf{w}_t^k, \xi_t^k) = \sum_{k=1}^N p_k\tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \mathbf{w}^*)$ and $\overline{\mathbf{g}}_t = \sum_{k=1}^N p_k\mathbf{H}^k(\mathbf{w}_t^k - \mathbf{w}^*)$

Theorem 6. *For the overparamterized linear regression problem, FedAvg with communication every E iterations with constant step size $\overline{\alpha} = \mathcal{O}(\frac{1}{E}\frac{N}{l\nu_{\max}+\mu(N-\nu_{\min})})$ has geometric convergence:*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) \leq \mathcal{O}\left(L\exp(-\frac{NT}{E(\nu_{\max}\kappa_1 + (N-\nu_{\min}))})\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right).$$

Proof. We again show the result first when $E = 2$ and $t-1$ is a communication round. We have

$$\begin{aligned}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \|(\overline{\mathbf{w}}_t - \alpha_t\mathbf{g}_t) - \mathbf{w}^*\|^2 \\ &= \|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 - 2\alpha_t\langle\overline{\mathbf{w}}_t - \mathbf{w}^*, \mathbf{g}_t\rangle + \alpha_t^2\|\mathbf{g}_t\|^2\end{aligned}$$

and

$$\begin{aligned}&-2\alpha_t\mathbb{E}_t\langle\overline{\mathbf{w}}_t - \mathbf{w}^*, \mathbf{g}_t\rangle \\ &= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t - \mathbf{w}^*, \nabla F_k(\mathbf{w}_t^k)\rangle \\ &= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k)\rangle - 2\alpha_t\sum_{k=1}^N p_k\langle\mathbf{w}_t^k - \mathbf{w}^*, \nabla F_k(\mathbf{w}_t^k)\rangle \\ &= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k)\rangle - 2\alpha_t\sum_{k=1}^N p_k\langle\mathbf{w}_t^k - \mathbf{w}^*, \mathbf{H}^k(\mathbf{w}_t^k - \mathbf{w}^*)\rangle \\ &= -2\alpha_t\sum_{k=1}^N p_k\langle\overline{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla F_k(\mathbf{w}_t^k)\rangle - 4\alpha_t\sum_{k=1}^N p_kF_k(\mathbf{w}_t^k) \\ &\leq 2\alpha_t\sum_{k=1}^N p_k(F_k(\mathbf{w}_t^k) - F_k(\overline{\mathbf{w}}_t) + \frac{L}{2}\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2) - 4\alpha_t\sum_{k=1}^N p_kF_k(\mathbf{w}_t^k)\end{aligned}$$

$$= \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 - 2\alpha_t \sum_{k=1}^N p_k F_k(\overline{\mathbf{w}}_t) - 2\alpha_t \sum_{k=1}^N p_k F_k(\mathbf{w}_t^k)$$
$$= \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 - \alpha_t \sum_{k=1}^N p_k \langle (\overline{\mathbf{w}}_t - \mathbf{w}^*), \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle - 2\alpha_t \sum_{k=1}^N p_k F_k(\mathbf{w}_t^k)$$

and

$$\begin{aligned}
\|\mathbf{g}_t\|^2 &= \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \mathbf{w}^*)\|^2 \\
&= \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*) + \sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \overline{\mathbf{w}}_t)\|^2 \\
&\leq 2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + 2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \overline{\mathbf{w}}_t)\|^2
\end{aligned}$$

which gives

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 - \alpha_t \sum_{k=1}^N p_k \langle \overline{\mathbf{w}}_t - \mathbf{w}^*, \mathbf{H}^k \overline{\mathbf{w}}_t - \mathbf{w}^* \rangle + 2\alpha_t^2 \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
+ \alpha_t L \sum_{k=1}^N p_k \|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + 2\alpha_t^2 \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k - \overline{\mathbf{w}}_t)\|^2 - 2\alpha_t \sum_{k=1}^N p_k F_k(\mathbf{w}_t^k)
\end{aligned}$$

following Ma et al. (2018) we first prove that

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 - \alpha_t \sum_{k=1}^N p_k \langle (\overline{\mathbf{w}}_t - \mathbf{w}^*), \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle + 2\alpha_t^2 \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
\leq (1 - \frac{N}{8(\nu_{\max}\kappa_1 + (N - \nu_{\min}))}) \mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2
\end{aligned}$$

with appropriately chosen α_t. Compared to the rate $O(\frac{\mu N}{l\nu_{\max} + L(N-\nu_{\min})}) = O(\frac{N}{\nu_{\max}\kappa_1 + (N-\nu_{\min})\kappa})$ for general strongly convex and smooth objectives, this is an improvement as linear speedup is now available for a larger range of N.

We have

$$\begin{aligned}
&\mathbb{E}_t \|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
&= \mathbb{E}_t \langle \sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*), \sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + \sum_{k=1}^N \sum_{j \neq k} p_j p_k \mathbb{E}_t \langle \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*), \tilde{\mathbf{H}}_t^j(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + \sum_{k=1}^N \sum_{j \neq k} p_j p_k \mathbb{E}_t \langle \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*), \mathbf{H}^j(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + \sum_{k=1}^N \sum_{j=1}^N p_j p_k \mathbb{E}_t \langle \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*), \mathbf{H}^j(\overline{\mathbf{w}}_t - \mathbf{w}^*)\rangle - \sum_{k=1}^N p_k^2 \|\mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
&= \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + \|\sum_k p_k \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 - \sum_{k=1}^N p_k^2 \|\mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
&\leq \sum_{k=1}^N p_k^2 \mathbb{E}_t \|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 + \|\sum_k p_k \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 - \frac{1}{N}\nu_{\min} \|\sum_k p_k \mathbf{H}^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2
\end{aligned}$$

$$\leq \frac{1}{N}\nu_{\max}\sum_{k=1}^{N}p_k\mathbb{E}_t\|\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\|^2+(1-\frac{1}{N}\nu_{\min})\|\sum_k p_k\mathbf{H}^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\|^2$$

$$\leq \frac{1}{N}\nu_{\max}l\sum_{k=1}^{N}p_k\langle(\overline{\mathbf{w}}_t-\mathbf{w}^*),\mathbf{H}^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\rangle+(1-\frac{1}{N}\nu_{\min})\|\sum_k p_k\mathbf{H}^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\|^2$$

$$= \frac{1}{N}\nu_{\max}l\langle(\overline{\mathbf{w}}_t-\mathbf{w}^*),\mathbf{H}(\overline{\mathbf{w}}_t-\mathbf{w}^*)\rangle+(1-\frac{1}{N}\nu_{\min})\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,\mathbf{H}^2(\overline{\mathbf{w}}_t-\mathbf{w}^*)\rangle$$

using $\|\tilde{\mathbf{H}}_t^k\|\leq l$.

Now we have

$$\mathbb{E}\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2-\alpha_t\sum_{k=1}^{N}p_k\langle(\overline{\mathbf{w}}_t-\mathbf{w}^*),\mathbf{H}^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\rangle+2\alpha_t^2\|\sum_{k=1}^{N}p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t-\mathbf{w}^*)\|^2=$$
$$\langle\overline{\mathbf{w}}_t-\mathbf{w}^*,(I-\alpha_t\mathbf{H}+2\alpha_t^2(\frac{\nu_{\max}l}{N}\mathbf{H}+\frac{N-\nu_{\min}}{N}\mathbf{H}^2))(\overline{\mathbf{w}}_t-\mathbf{w}^*)\rangle$$

and it remains to bound the maximum eigenvalue of

$$(I-\alpha_t\mathbf{H}+2\alpha_t^2(\frac{\nu_{\max}l}{N}\mathbf{H}+\frac{N-\nu_{\min}}{N}\mathbf{H}^2))$$

and we bound this following Ma et al. (2018). If we choose $\alpha_t<\frac{N}{2(\nu_{\max}l+(N-\nu_{\min})L)}$, then

$$-\alpha_t\mathbf{H}+2\alpha_t^2(\frac{\nu_{\max}l}{N}\mathbf{H}+\frac{N-\nu_{\min}}{N}\mathbf{H}^2)\prec 0$$

and the convergence rate is given by the maximum of $1-\alpha_t\lambda+2\alpha_t^2(\frac{\nu_{\max}l}{N}\lambda+\frac{N-\nu_{\min}}{N}\lambda^2)$ maximized over the non-zero eigenvalues λ of $\mathbf{H}$. To select the step size α_t that gives the smallest upper bound, we then minimize over α_t, resulting in

$$\min_{\alpha_t<\frac{N}{2(\nu_{\max}l+(N-\nu_{\min})L)}}\ \max_{\lambda>0:\exists v,\mathbf{H}v=\lambda v}\left\{1-\alpha_t\lambda+2\alpha_t^2(\frac{\nu_{\max}l}{N}\lambda+\frac{N-\nu_{\min}}{N}\lambda^2)\right\}$$

Since the objective is quadratic in λ, the maximum is achieved at either the largest eigenvalue $\lambda_{\max}$ of $\mathbf{H}$ or the smallest non-zero eigenvalue $\lambda_{\min}$ of $\mathbf{H}$.

When $N\leq\frac{4\nu_{\max}l}{L-\lambda_{\min}}+4\nu_{\min}$, i.e. when $N=O(l/\lambda_{\min})=O(\kappa_1)$, the optimal objective value is achieved at $\lambda_{\min}$ and the optimal step size is given by $\alpha_t=\frac{N}{4(\nu_{\max}l+(N-\nu_{\min})\lambda_{\min})}$. The optimal convergence rate (i.e. the optimal objective value) is equal to $1-\frac{1}{8}\frac{N\lambda_{\min}}{(\nu_{\max}l+(N-\nu_{\min})\lambda_{\min})}=1-\frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1+(N-\nu_{\min}))}$. This implies that when $N=O(\kappa_1)$, the optimal convergence rate has a linear speedup in N. When N is larger, this step size is no longer optimal, but we still have $1-\frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1+(N-\nu_{\min}))}$ as an upper bound on the convergence rate.

Now we have proved

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\|^2\leq(1-\frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1+(N-\nu_{\min}))})\mathbb{E}\|\overline{\mathbf{w}}_t-\mathbf{w}^*\|^2$$
$$+\alpha_tL\sum_{k=1}^{N}p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2+2\alpha_t^2\|\sum_{k=1}^{N}p_k\tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k-\overline{\mathbf{w}}_t)\|^2-2\alpha_t\sum_{k=1}^{N}p_kF_k(\mathbf{w}_t^k)$$

Next we bound terms in the second line using a similar argument as the general case. We have

$$2\alpha_t^2\|\sum_{k=1}^{N}p_k\tilde{\mathbf{H}}_t^k(\mathbf{w}_t^k-\overline{\mathbf{w}}_t)\|^2\leq 2\alpha_t^2l^2\sum_{k=1}^{N}p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2$$

and

$$\mathbb{E}\sum_{k=1}^{N}p_k\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\|^2\leq\mathbb{E}2(1+2l^2\alpha_{t-1}^2)\sum_{k=1}^{N}p_k\|\overline{\mathbf{w}}_{t-1}-\mathbf{w}_{t-1}^k\|^2+8\alpha_{t-1}^2lF(\overline{\mathbf{w}}_{t-1})$$

$$= 4\alpha_{t-1}^2 l\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\rangle$$

and if α_t, α_{t-1} satisfy

$$\begin{aligned}
\alpha_t L + 2\alpha_t^2 &\leq (1 - \frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})(\alpha_{t-1}L + 2\alpha_{t-1}^2)/3 \\
2(1 + 2l^2\alpha_{t-1}^2) &\leq 3 \\
\alpha_t L + 2\alpha_t^2 &\leq 1
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \\
&\leq (1 - \frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})[\mathbb{E}\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 - \alpha_t\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\rangle + 2\alpha_t^2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{w}}_t - \mathbf{w}^*)\|^2 \\
&+ (\alpha_{t-1}L + 2\alpha_{t-1}^2)\cdot 2\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}_{t-1}^k\|^2 + 4\alpha_{t-1}^2 l\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\rangle]
\end{aligned}$$

and again by choosing $\alpha_{t-1} = c\frac{N}{8(\nu_{\max}l + (N - \nu_{\min})\lambda_{\min})}$ for a small constant c, we can guarantee that

$$\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 - \alpha_{t-1}\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\rangle \\
&+2\alpha_{t-1}^2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_{t-1}^k(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\|^2 + 4\alpha_{t-1}^2 l\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\rangle \\
&\leq (1 - c\frac{N}{16(\nu_{\max}l + (N - \nu_{\min})\lambda_{\min})})\mathbb{E}\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2
\end{aligned}$$

For general E, we have the recursive relation

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &\leq \mathbb{E}(1 - c\frac{1}{8}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})\cdots(1 - c\frac{1}{8\tau}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})[\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\|^2 \\
&- \alpha_{t-\tau}\langle \overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*, \mathbf{H}\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\rangle + 2\alpha_{t-\tau}^2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_{t-\tau}^k(\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*)\|^2 \\
&+ 4\tau\alpha_{t-1}^2 l\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\rangle \\
&+ (2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L)((\tau+1)\sum_{k=1}^N p_k\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}_{t-\tau}^k\|^2)]
\end{aligned}$$

as long as the step sizes are chosen $\alpha_{t-\tau} = c\frac{N}{4\tau(\nu_{\max}l + (N - \nu_{\min})\lambda_{\min})}$ such that the following inequalities hold

$$\begin{aligned}
(2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L) &\leq (1 - \alpha_{t-\tau}\mu)(2\alpha_{t-\tau-1}^2 l^2 + \alpha_{t-\tau-1}L)/3 \\
2(1 + 2l^2\alpha_{t-\tau}^2) &\leq 3 \\
2\alpha_{t-\tau}^2 l^2 + \alpha_{t-\tau}L &\leq 1
\end{aligned}$$

and

$$\begin{aligned}
&\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\|^2 - \alpha_{t-\tau}\langle \overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*, \mathbf{H}\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\rangle \\
&+ 2\alpha_{t-\tau}^2\|\sum_{k=1}^N p_k \tilde{\mathbf{H}}_{t-\tau}^k(\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*)\|^2 + 4\tau\alpha_{t-1}^2 l\langle \overline{\mathbf{w}}_{t-1} - \mathbf{w}^*, \mathbf{H}(\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*)\rangle \\
&\leq (1 - c\frac{N}{8(\tau+1)(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})\mathbb{E}\|\overline{\mathbf{w}}_{t-\tau} - \mathbf{w}^*\|^2
\end{aligned}$$

which gives

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 &\leq (1 - c\frac{1}{8E}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))})^t\|\mathbf{w}_0 - \mathbf{w}^*\|^2 \\
&= O(\exp(-\frac{1}{E}\frac{N}{(\nu_{\max}\kappa_1 + (N - \nu_{\min}))}t))\|\mathbf{w}_0 - \mathbf{w}^*\|^2
\end{aligned}$$

and with partial participation, the same bound holds with a possibly different choice of c. □

H.3 Geometric Convergence of FedMaSS for Overparameterized Linear Regression

Theorem 7. *For the overparamterized linear regression problem, FedMaSS with communication every E iterations and constant step sizes $\overline{\eta}_1 = \mathcal{O}(\frac{1}{E}\frac{N}{l\nu_{\max}+\mu(N-\nu_{\min})}), \overline{\eta}_2 = \frac{\overline{\eta}_1(1-\frac{1}{\tilde{\kappa}})}{1+\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}, \overline{\gamma} = \frac{1-\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}{1+\frac{1}{\sqrt{\kappa_1\tilde{\kappa}}}}$ has geometric convergence:*

$$\mathbb{E}F(\overline{\mathbf{w}}_T) \leq \mathcal{O}\left(L\exp(-\frac{NT}{E(\nu_{\max}\sqrt{\kappa_1\tilde{\kappa}} + (N-\nu_{\min}))})\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right).$$

Proof. The proof is based on results in Liu & Belkin (2020) which originally proposed the MaSS algorithm. Note that the update can equivalently be written as

$$\begin{aligned}
\mathbf{v}_{t+1}^k &= (1-\alpha^k)\mathbf{v}_t^k + \alpha^k\mathbf{u}_t^k - \delta^k\mathbf{g}_{t,k} \\
\mathbf{w}_{t+1}^k &= \begin{cases} \mathbf{u}_t^k - \eta^k\mathbf{g}_{t,k} & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k=1}^N p_k\left[\mathbf{u}_t^k - \eta^k\mathbf{g}_{t,k}\right] & \text{if } t+1 \in \mathcal{I}_E \end{cases} \\
\mathbf{u}_{t+1}^k &= \frac{\alpha^k}{1+\alpha^k}\mathbf{v}_{t+1}^k + \frac{1}{1+\alpha^k}\mathbf{w}_{t+1}^k
\end{aligned}$$

where there is a bijection between the parameters $\frac{1-\alpha^k}{1+\alpha^k} = \gamma^k, \eta^k = \eta_1^k, \frac{\eta^k - \alpha^k\delta^k}{1+\alpha^k} = \eta_2^k$, and we further introduce an auxiliary parameter $\mathbf{v}_t^k$, which is initialized at $\mathbf{v}_0^k$. We also note that when $\delta^k = \frac{\eta^k}{\alpha^k}$, the update reduces to the Nesterov accelerated SGD. This version of the FedAvg algorithm with local MaSS updates is used for analyzing the geometric convergence.

As before, define the virtual sequences $\overline{\mathbf{w}}_t = \sum_{k=1}^N p_k\mathbf{w}_t^k, \overline{\mathbf{v}}_t = \sum_{k=1}^N p_k\mathbf{v}_t^k, \overline{\mathbf{u}}_t = \sum_{k=1}^N p_k\mathbf{u}_t^k$, and $\overline{\mathbf{g}}_t = \sum_{k=1}^N p_k\mathbb{E}\mathbf{g}_{t,k}$. We have $\mathbb{E}\mathbf{g}_t = \overline{\mathbf{g}}_t$ and $\overline{\mathbf{w}}_{t+1} = \overline{\mathbf{u}}_t - \eta_t\mathbf{g}_t$, $\overline{\mathbf{v}}_{t+1} = (1-\alpha^k)\overline{\mathbf{v}}_t + \alpha^k\overline{\mathbf{w}}_t - \delta^k\mathbf{g}_t$, and $\overline{\mathbf{u}}_{t+1} = \frac{\alpha^k}{1+\alpha^k}\overline{\mathbf{v}}_{t+1} + \frac{1}{1+\alpha^k}\overline{\mathbf{w}}_{t+1}$.

We first prove the theorem with $E=2$ and $t-1$ being a communication round. We have

$$\begin{aligned}
&\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 \\
&= \|(1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \delta\sum_k p_k\tilde{\mathbf{H}}_t^k(\mathbf{u}_t^k - \mathbf{w}^*) - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 \\
&= \|(1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 + \delta^2\|\sum_k p_k\tilde{\mathbf{H}}_t^k(\mathbf{u}_t^k - \mathbf{w}^*)\|_{\mathbf{H}^{-1}}^2 \\
&- 2\delta\langle\sum_k p_k\tilde{\mathbf{H}}_t^k(\mathbf{u}_t^k - \mathbf{w}^*), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle_{\mathbf{H}^{-1}} \\
&\leq \underbrace{\|(1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2}_{A} + \underbrace{2\delta^2\|\sum_k p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{w}^*)\|_{\mathbf{H}^{-1}}^2}_{B} + 2\delta^2\|\sum_k p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{u}_t^k)\|_{\mathbf{H}^{-1}}^2 \\
&\underbrace{-2\delta\langle\sum_k p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{w}^*), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle_{\mathbf{H}^{-1}}}_{C} \\
&- 2\delta\langle\sum_k p_k\tilde{\mathbf{H}}_t^k(\mathbf{u}_t^k - \overline{\mathbf{u}}_t), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle_{\mathbf{H}^{-1}}
\end{aligned}$$

Following the proof in Liu & Belkin (2020),

$$\begin{aligned}
\mathbb{E}A &\leq \mathbb{E}(1-\alpha)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 + \alpha\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 \\
&\leq \mathbb{E}(1-\alpha)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 + \frac{\alpha}{\mu}\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2
\end{aligned}$$

using the convexity of the norm $\|\cdot\|_{\mathbf{H}^{-1}}$ and that μ is the smallest non-zero eigenvalue of H.

Now

$$\mathbb{E}B \leq 2\delta^2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N})\|(\overline{\mathbf{u}}_t - \mathbf{w}^*)\|_H^2$$

using the folowing bound:

$$\begin{aligned}
\mathbb{E}\left(\sum_k p_k \tilde{\mathbf{H}}_t^k\right)\mathbf{H}^{-1}\left(\sum_k p_k \tilde{\mathbf{H}}_t^k\right) &= \mathbb{E}\sum_k p_k^2 \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \sum_{k\neq j} p_k p_j \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^j \\
&\preceq \nu_{\max}\frac{1}{N}\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \sum_{k\neq j} p_k p_j \mathbf{H}^k \mathbf{H}^{-1}\mathbf{H}^j \\
&= \nu_{\max}\frac{1}{N}\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \sum_{k,j} p_k p_j \mathbf{H}^k \mathbf{H}^{-1}\mathbf{H}^j - \sum_k p_k^2 \mathbf{H}^k \mathbf{H}^{-1}\mathbf{H}^k \\
&\preceq \nu_{\max}\frac{1}{N}\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \mathbf{H} - \frac{1}{N}\nu_{\min}\sum_k p_k \mathbf{H}^k \mathbf{H}^{-1}\mathbf{H}^k \\
&\preceq \nu_{\max}\frac{1}{N}\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \mathbf{H} - \frac{1}{N}\nu_{\min}(\sum_k p_k \mathbf{H}^k)\mathbf{H}^{-1}(\sum_k p_k \mathbf{H}^k) \\
&= \nu_{\max}\frac{1}{N}\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k + \frac{N-\nu_{\min}}{N}\mathbf{H} \\
&\preceq \nu_{\max}\frac{1}{N}\tilde{\kappa}\mathbf{H} + \frac{N-\nu_{\min}}{N}\mathbf{H}
\end{aligned}$$

where we have used $\mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \mathbf{H}^{-1}\tilde{\mathbf{H}}_t^k \leq \tilde{\kappa}\mathbf{H}$ by definition of $\tilde{\kappa}$ and the operator convexity of the mapping $W \to W\mathbf{H}^{-1}W$.

Finally,

$$\begin{aligned}
\mathbb{E}C &= -\mathbb{E}2\delta\langle\sum_k p_k \tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{w}^*), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle_{\mathbf{H}^{-1}} \\
&= -2\delta\langle\sum_k p_k \mathbf{H}^k(\overline{\mathbf{u}}_t - \mathbf{w}^*), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle_{\mathbf{H}^{-1}} \\
&= -2\delta\langle(\overline{\mathbf{u}}_t - \mathbf{w}^*), (1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t - \mathbf{w}^*\rangle \\
&= -2\delta\langle(\overline{\mathbf{u}}_t - \mathbf{w}^*), \overline{\mathbf{u}}_t - \mathbf{w}^* + \frac{1-\alpha}{\alpha}(\overline{\mathbf{u}}_t - \overline{\mathbf{w}}_t)\rangle \\
&= -2\delta\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 + \frac{1-\alpha}{\alpha}\delta(\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 - \|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 - \|\overline{\mathbf{w}}_t - \overline{\mathbf{u}}_t\|^2) \\
&\leq \frac{1-\alpha}{\alpha}\delta\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 - \frac{1-\alpha}{\alpha}\delta\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2
\end{aligned}$$

where we have used

$$\begin{aligned}
&(1-\alpha)\overline{\mathbf{v}}_t + \alpha\overline{\mathbf{u}}_t \\
&= (1-\alpha)((1+\alpha)\overline{\mathbf{u}}_t - \overline{\mathbf{w}}_t)/\alpha + \alpha\overline{\mathbf{u}}_t \\
&= \frac{1}{\alpha}\overline{\mathbf{u}}_t - \frac{1-\alpha}{\alpha}\overline{\mathbf{w}}_t
\end{aligned}$$

and the identity that $-2\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}+\mathbf{b}\|^2$.

It follows that

$$\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 \\
&\leq (1-\alpha)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|_{\mathbf{H}^{-1}}^2 + \frac{1-\alpha}{\alpha}\delta\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 \\
&+ (\frac{\alpha}{\mu} - \frac{1-\alpha}{\alpha}\delta)\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 + 2\delta^2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N})\|(\overline{\mathbf{u}}_t - \mathbf{w}^*)\|_H^2
\end{aligned}$$

$$+ 2\delta^2 \| \sum_k p_k \tilde{\mathbf{H}}_t^k (\overline{\mathbf{u}}_t - \mathbf{u}_t^k) \|^2_{\mathbf{H}^{-1}}$$
$$- 2\delta \langle \sum_k p_k \tilde{\mathbf{H}}_t^k (\mathbf{u}_t^k - \overline{\mathbf{u}}_t), (1-\alpha)\overline{\mathbf{v}}_t + \alpha \overline{\mathbf{u}}_t - \mathbf{w}^* \rangle_{\mathbf{H}^{-1}}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \mathbb{E}\|\overline{\mathbf{u}}_t - \mathbf{w}^* - \eta \sum_k p_k \tilde{\mathbf{H}}_t^k (\overline{\mathbf{u}}_t - \mathbf{w}^*)\|^2 \\
&= \mathbb{E}\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 - 2\eta \|\overline{\mathbf{u}}_t - \mathbf{w}^*\|_H^2 + \eta^2 \| \sum_k p_k \tilde{\mathbf{H}}_t^k (\overline{\mathbf{u}}_t - \mathbf{w}^*)\|^2 \\
&\leq \mathbb{E}\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 - 2\eta \|\overline{\mathbf{u}}_t - \mathbf{w}^*\|_H^2 + \eta^2 (\nu_{\max} \frac{1}{N} \ell + L \frac{N - \nu_{\min}}{N}) \|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2
\end{aligned}$$

where we use the following bound:

$$\begin{aligned}
&\mathbb{E}\left(\sum_k p_k \tilde{\mathbf{H}}_t^k\right)\left(\sum_k p_k \tilde{\mathbf{H}}_t^k\right) \\
&= \mathbb{E}\sum_k p_k^2 \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \sum_{k \neq j} p_k p_j \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^j \\
&\preceq \nu_{\max} \frac{1}{N} \mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \sum_{k \neq j} p_k p_j \mathbf{H}^k \mathbf{H}^j \\
&= \nu_{\max} \frac{1}{N} \mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \sum_{k,j} p_k p_j \mathbf{H}^k \mathbf{H}^j - \sum_k p_k^2 \mathbf{H}^k \mathbf{H}^k \\
&\preceq \nu_{\max} \frac{1}{N} \mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \mathbf{H}^2 - \frac{1}{N} \nu_{\min} \sum_k p_k \mathbf{H}^k \mathbf{H}^k \\
&\preceq \nu_{\max} \frac{1}{N} \mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \mathbf{H}^2 - \frac{1}{N} \nu_{\min} (\sum_k p_k \mathbf{H}^k)(\sum_k p_k \mathbf{H}^k) \\
&= \nu_{\max} \frac{1}{N} \mathbb{E}\sum_k p_k \tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k + \frac{N - \nu_{\min}}{N} \mathbf{H}^2 \\
&\preceq \nu_{\max} \frac{1}{N} l \mathbf{H} + L \frac{N - \nu_{\min}}{N} \mathbf{H}
\end{aligned}$$

again using that $W \to W^2$ is operator convex and that $\mathbb{E}\tilde{\mathbf{H}}_t^k \tilde{\mathbf{H}}_t^k \preceq l \mathbf{H}^k$ by definition of l.

Combining the bounds for $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$ and $\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}}$,

$$\begin{aligned}
&\mathbb{E}\frac{\delta}{\alpha}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 + \|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}} \\
&\leq (1-\alpha)\|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}} + \frac{1-\alpha}{\alpha}\delta\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + (\frac{\alpha}{\mu} - \delta)\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 \\
&+ (2\delta^2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N - \nu_{\min}}{N}) - 2\eta\delta/\alpha + \eta^2\delta(\nu_{\max}\frac{1}{N} l + L\frac{N - \nu_{\min}}{N})/\alpha)\|\overline{\mathbf{u}}_t - \mathbf{w}^*\|^2 \\
&+ 2\delta^2 \| \sum_k p_k \tilde{\mathbf{H}}_t^k (\overline{\mathbf{u}}_t - \mathbf{u}_t^k) \|^2_{\mathbf{H}^{-1}} \\
&+ \delta L \sum_k p_k \| (\overline{\mathbf{u}}_t - \mathbf{u}_t^k) \|^2_{\mathbf{H}^{-1}}
\end{aligned}$$

Following Liu & Belkin (2020) if we choose step sizes so that

$$\frac{\alpha}{\mu} - \delta \leq 0$$
$$2\delta^2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N - \nu_{\min}}{N}) - 2\eta\delta/\alpha + \eta^2\delta(\nu_{\max}\frac{1}{N} l + L\frac{N - \nu_{\min}}{N})/\alpha \leq 0$$

or equivalently

$$\alpha/\delta \leq \mu$$

$$2\alpha\delta(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N}) + \eta(\eta(\nu_{\max}\frac{1}{N}l + L\frac{N-\nu_{\min}}{N}) - 2) \leq 0$$

the second and third terms are negative. To optimize the step sizes, note that the two inequalities imply

$$\alpha^2 \leq \eta(2 - \eta(\nu_{\max}\frac{1}{N}l + L\frac{N-\nu_{\min}}{N}))\mu/2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N})$$

and maximizing the right hand side with respect to η, which is quadratic, we see that $\eta \equiv 1/(\nu_{\max}\frac{1}{N}l + L\frac{N-\nu_{\min}}{N})$ maximizes the right hand side, with

$$\alpha \equiv \frac{1}{\sqrt{2(\nu_{\max}\frac{1}{N}\kappa_1 + \kappa\frac{N-\nu_{\min}}{N})(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N})}}$$

$$\delta \equiv \frac{\alpha}{\mu} = \frac{\eta}{\alpha(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N})}$$

Note that $\alpha = \frac{1}{\sqrt{2(\nu_{\max}\frac{1}{N}\kappa_1+\kappa\frac{N-\nu_{\min}}{N})(\nu_{\max}\frac{1}{N}\tilde{\kappa}+\frac{N-\nu_{\min}}{N})}} = O(\frac{N}{\sqrt{\kappa_1\tilde{\kappa}}})$ when $N = O(\min\{\tilde{\kappa}, \kappa_1/\kappa\})$.

Finally, to deal with the terms $2\delta^2\|\sum_k p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{u}_t^k)\|^2_{\mathbf{H}^{-1}} + \delta L\sum_k p_k\|(\overline{\mathbf{u}}_t - \mathbf{u}_t^k)\|^2_{\mathbf{H}^{-1}}$, we can use Jensen

$$\begin{aligned}
&2\delta^2\|\sum_k p_k\tilde{\mathbf{H}}_t^k(\overline{\mathbf{u}}_t - \mathbf{u}_t^k)\|^2_{\mathbf{H}^{-1}} + \delta L\sum_k p_k\|(\overline{\mathbf{u}}_t - \mathbf{u}_t^k)\|^2_{\mathbf{H}^{-1}} \\
&\leq (2\delta^2 l^2 + \delta L)\sum_k p_k\|\overline{\mathbf{u}}_t - \mathbf{u}_t^k\|^2_{\mathbf{H}^{-1}} \\
&= (2\delta^2 l^2 + \delta L)\sum_k p_k\|\frac{\alpha}{1+\alpha}\overline{\mathbf{v}}_t + \frac{1}{1+\alpha}\overline{\mathbf{w}}_t - (\frac{\alpha}{1+\alpha}v_t^k + \frac{1}{1+\alpha}w_t^k)\|^2_{\mathbf{H}^{-1}} \\
&\leq (2\delta^2 l^2 + \delta L)(2(\frac{\alpha}{1+\alpha})^2\delta^2 + 2(\frac{1}{1+\alpha})^2\eta^2)\sum_k p_k\|\tilde{\mathbf{H}}_{t-1}^k(\overline{\mathbf{u}}_{t-1} - \mathbf{w}^*)\|^2 \\
&\leq (2\delta^2 l^2 + \delta L)(2(\frac{\alpha}{1+\alpha})^2\delta^2 + 2(\frac{1}{1+\alpha})^2\eta^2)l^2\|(\overline{\mathbf{u}}_{t-1} - \mathbf{w}^*)\|^2
\end{aligned}$$

which can be combined with the terms with $\|(\overline{\mathbf{u}}_{t-1} - \mathbf{w}^*)\|^2$ in the recursive expansion of $\mathbb{E}\frac{\delta}{\alpha}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}}$:

$$\begin{aligned}
&\mathbb{E}\frac{\delta}{\alpha}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \|\overline{\mathbf{v}}_t - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}} \\
&\leq (1-\alpha)\|\overline{\mathbf{v}}_{t-1} - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}} + \frac{1-\alpha}{\alpha}\delta\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2 + (\frac{\alpha}{\mu} - \delta)\|\overline{\mathbf{u}}_{t-1} - \mathbf{w}^*\|^2 \\
&+ (2\delta^2(\nu_{\max}\frac{1}{N}\tilde{\kappa} + \frac{N-\nu_{\min}}{N}) - 2\eta\delta/\alpha + \eta^2\delta(\nu_{\max}\frac{1}{N}l + L\frac{N-\nu_{\min}}{N})/\alpha)\|\overline{\mathbf{u}}_{t-1} - \mathbf{w}^*\|^2
\end{aligned}$$

and the step sizes can be chosen so that the resulting coefficients are negative. Therefore, we have shown that

$$\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq (1-\alpha)^2\|\overline{\mathbf{w}}_{t-1} - \mathbf{w}^*\|^2$$

where $\alpha = \frac{1}{\sqrt{2(\nu_{\max}\frac{1}{N}\kappa_1+\kappa\frac{N-\nu_{\min}}{N})(\nu_{\max}\frac{1}{N}\tilde{\kappa}+\frac{N-\nu_{\min}}{N})}} = O(\frac{N}{\nu_{\max}\sqrt{\kappa_1\tilde{\kappa}}+N-\nu_{\min}})$ when $N = O(\min\{\tilde{\kappa}, \kappa_1/\kappa\})$.

For general $E > 1$, choosing $\eta = c/E(\nu_{\max}\frac{1}{N}l + L\frac{N-\nu_{\min}}{N})$ for some small constant c results in $\alpha = O(\frac{1}{E\sqrt{(\nu_{\max}\frac{1}{N}\kappa_1+\kappa\frac{N-\nu_{\min}}{N})(\nu_{\max}\frac{1}{N}\tilde{\kappa}+\frac{N-\nu_{\min}}{N})}})$ and this guarantees that

$$\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 \leq (1-\alpha)^t\|\mathbf{w}_0 - \mathbf{w}^*\|^2$$

for all t.

$\square$

I DETAILS ON EXPERIMENTS AND ADDITIONAL RESULTS

We describe the precise procedure to reproduce the results in this paper. As we mentioned in Section 5, we empirically verified the linear speed up on various convex settings for both FedAvg and its accelerated variants. For all the results, we set random seeds as $0, 1, 2$ and report the best convergence rate across the three folds. For each run, we initialize $\mathbf{w}_0 = \mathbf{0}$ and measure the number of iteration to reach the target accuracy ϵ. We use the small-scale dataset w8a Platt (1998), which consists of $n = 49749$ samples with feature dimension $d = 300$. The label is either positive one or negative one. The dataset has sparse binary features in $\{0, 1\}$. Each sample has 11.15 non-zero feature values out of 300 features on average. We set the batch size equal to four across all experiments. In the next following subsections, we introduce parameter searching in each objective separately.

I.1 STRONGLY CONVEX OBJECTIVES

We first consider the strongly convex objective function, where we use a regularized binary logistic regression with regularization $\lambda = 1/n \approx 2e - 5$. We evenly distributed on $1, 2, 4, 8, 16, 32$ devices and report the number of iterations/rounds needed to converge to $\epsilon-$accuracy, where $\epsilon = 0.005$. The optimal objective function value f^* is set as $f^* = 0.126433176216545$. This is determined numerically and we follow the setting in Stich (2019). The learning rate is decayed as the $\eta_t = \min(\eta_0, \frac{nc}{1+t})$, where we extensively search the best learning rate $c \in \{2^{-1}c_0, 2^{-2}c_0, c_0, 2c_0, 2^2c_0\}$. In this case, we search the initial learning rate $\eta_0 \in \{1, 32\}$ and $c_0 = 1/8$.

I.2 CONVEX SMOOTH OBJECTIVES

We also use binary logistic regression without regularization. The setting is almost same as its regularized counter part. We also evenly distributed all the samples on $1, 2, 4, 8, 16, 32$ devices. The figure shows the number of iterations needed to converge to $\epsilon-$accuracy, where $\epsilon = 0.02$. The optiaml objective function value is set as $f^* = 0.11379089057514849$, determined numerically. The learning rate is decayed as the $\eta_t = \min(\eta_0, \frac{nc}{1+t})$, where we extensively search the best learning rate $c \in \{2^{-1}c_0, 2^{-2}c_0, c_0, 2c_0, 2^2c_0\}$. In this case, we search the initial learning rate $\eta_0 \in \{1, 32\}$ and $c_0 = 1/8$.

I.3 LINEAR REGRESSION

For linear regression, we use the same feature vectors from w8a dataset and generate ground truth $[\mathbf{w}^*, b^*]$ from a multivariate normal distribution with zero mean and standard deviation one. Then we generate label based on $y_i = \mathbf{x}_i^t \mathbf{w}^* + b^*$. This procedure will ensure we satisfy the over-parameterized setting as required in our theorems. We also evenly distributed all the samples on $1, 2, 4, 8, 16, 32$ devices. The figure shows the number of iterations needed to converge to $\epsilon-$accuracy, where $\epsilon = 0.02$. The optiaml objective function value is $f^* = 0$. The learning rate is decayed as the $\eta_t = \min(\eta_0, \frac{nc}{1+t})$, where we extensively search the best learning rate $c \in \{2^{-1}c_0, 2^{-2}c_0, c_0, 2c_0, 2^2c_0\}$. In this case, we search the initial learning rate $\eta_0 \in \{0.1, 0.12\}$ and $c_0 = 1/256$.

I.4 PARTIAL PARTICIPATION

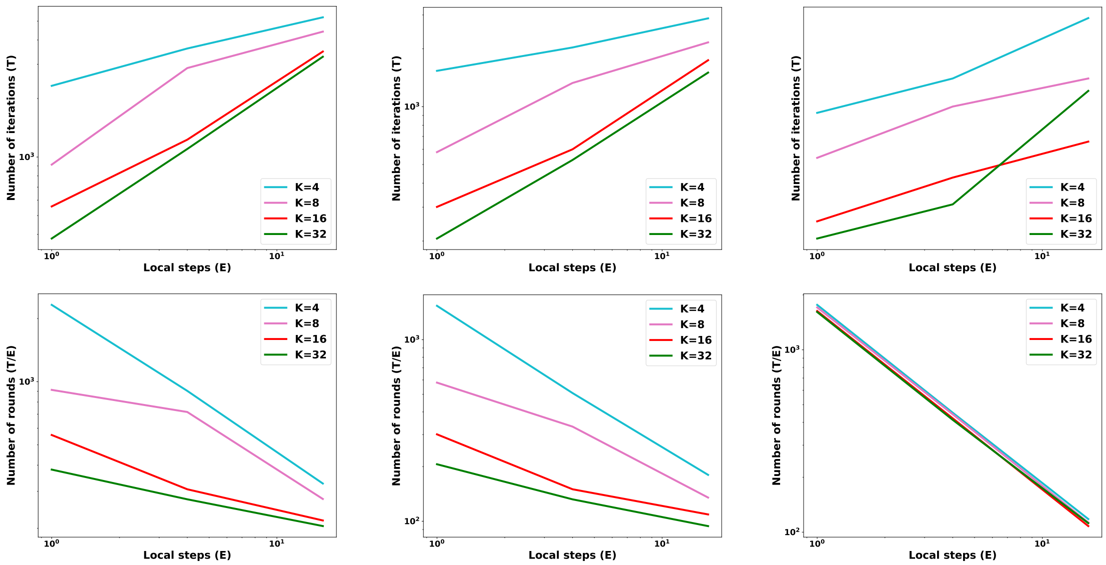
To examine the linear speedup of FedAvg in partial participation setting, we evenly distributed data on $4, 8, 16, 32, 64, 128$ devices and uniformly sample 50% devices without replacement. All other hyperparameters are the same as previous sections.

I.5 NESTEROV ACCELERATED FEDAVG

The experiments of Nesterov accelerated FedAvg (the update formula is given as follows) uses the same setting as previous three sections for vanilia FedAvg.

$$\mathbf{y}_{t+1}^k = \mathbf{w}_t^k - \alpha_t \mathbf{g}_{t,k}$$

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{y}_{t+1}^k + \beta_t(\mathbf{y}_{t+1}^k - \mathbf{y}_t^k) & \text{if } t+1 \notin \mathcal{I}_E \\ \sum_{k \in \mathcal{S}_{t+1}} \left(\mathbf{y}_{t+1}^k + \beta_t(\mathbf{y}_{t+1}^k - \mathbf{y}_t^k)\right) & \text{if } t+1 \in \mathcal{I}_E \end{cases}$$



(a) Strongly convex objective (b) Convex smooth objective (c) Linear regression

Figure 2: The convergence of FedAvg w.r.t the number of local steps E.

We set $\beta_t = 0.1$ and search α_t in the same way as η_t in FedAvg.

I.6 THE IMPACT OF E.

In this subsection, we further examine how does the number of local steps (E) affect convergence. As shown in Figure 2, the number of iterations increases as E increase, which slow down the convergence in terms of gradient computation. However, it can save communication costs as the number of rounds decreased when the E increases. This showcase that we need a proper choice of E to trade-off the communication cost and convergence speed.