PERFORMATIVE FEDERATED LEARNING

Kun Jin *,1 **Tongxin Yin** *,1 **Zhongzhu Chen** *,1 **Zeyu Sun** 1 **Xueru Zhang** 2
Yang Liu 3,4 **Mingyan Liu** 1
1 University of Michigan, Ann Arbor, 2 OSU, 3 UC Santa Cruz 4 ByteDance AI Lab

ABSTRACT

We consider a federated learning (FL) system comprising multiple clients and a server, wherein the clients collaborate to learn a common decision model from their distributed data. Unlike the conventional FL framework, we consider the scenario where the clients' data distributions change with the deployed decision model. In this work, we propose a performative federated learning framework that formalizes model-dependent distribution shift by leveraging the concept of distribution shift mappings in the performative prediction literature. We introduce necessary and sufficient conditions for the existence of a unique performative stable solution and characterize its distance to the performative optimal solution. Under such conditions, we propose the performative `FedAvg` algorithm and show that it converges to the performative stable solution at a rate of $\mathcal{O}(1/T)$ under both full and partial participation schemes. In addition, we show how the clients' heterogeneity influences the convergence both theoretically and using numerical results.

1 INTRODUCTION

Traditional learning problems typically assume the data distribution are static or unaffected by the learned model itself. However, in many real-world applications, the data distribution can shift as the result of the very learning outcome, when individuals respond to the algorithmic decisions they are subjected to. For instance, users with certain accents may stop using a speech recognition software when they experience excessive errors, directly impacting the diversity of speech samples collected by the software used for improving the product. Another example is "gaming the algorithm", where users may attempt to manipulate critical features, either honestly or dishonestly, to obtain a favorable decision from the algorithm (e.g., in loan approvals or job applications). This again can directly lead to the distributional change in features and labels that the algorithm relies on for decision making.

When the deployed model itself can trigger changes in the data distribution and influence the objective, the prediction problems are defined as *performative predictions* (PP) Perdomo et al. (2020). Typical scenarios of PP include *strategic learning* Hardt et al. (2016); Dong et al. (2018); Milli et al. (2019); Hu et al. (2019); Braverman & Garg (2020); Chen et al. (2020); Miller et al. (2020); Shavit et al. (2020); Haghtalab et al. (2020); Kleinberg & Raghavan (2020); Zrnic et al. (2021). PP has been primarily studied in a centralized setting, with fruitful literature including the convergence analysis Mendler-Dünner et al. (2020); Drusvyatskiy & Xiao (2020); Brown et al. (2020); Li & Wai (2022); Wood et al. (2022) and algorithm development Izzo et al. (2021; 2022); Miller et al. (2021); Ray et al. (2022).

In modern large-scale machine learning, distributed learning offers better privacy protection and avoids the computational resource bottlenecks compared to centralized learning, and federated learning (FL) is one of the most popular examples. Here the issue of distribution shift is further compounded due to data heterogeneity in a distributed setting. Specifically, the distributed data sources can be heterogeneous in nature, and their respective distribution shifts can also be different. Prior works in FL systems that address data distribution shiftstypically do not consider shifts in local distributions at the client end induced by the model (Guo et al., 2021; Casado et al., 2022; Rizk et al., 2020; Hosseinalipour et al., 2022; Zhu et al., 2021a; Eichner et al., 2019; Ding et al., 2020). In this work, we propose the *performative federated learning* framework to study and handle such

[*] the first three authors contributed equally

distribution shifts in FL. Extending the current results in PP to the decentralized FL has a number of challenges: 1) *Data heterogeneity:* As already one of the major difficulties in FL, tackling data heterogeneity faces additional challenges when taking the disparity of client distribution shift into consideration. 2) *Central* $\rightleftarrows$ *Local:* During training, clients receive the aggregated model at certain steps and train from it. While fitting better as an entity, such aggregation may fail to fit well on each client, which may lead to more severe shifting issues. 3) *heterogeneity in shift:* some clients may be more sensitive to the deployed decisions and have more drastic data shifts than other clients, e.g., due to different manipulation costs in strategic learning.

Toward this end, we formally introduce the *performative FedAvg* algorithm, or P-FedAvg, and establish its convergence. Our main findings are as follows. First of all, we prove the uniqueness of the performative stable (PS) solution reached by the algorithm, and show that it is a provable approximation to the performative optimal (PO) solution under mild conditions. Both solutions will be formally defined in Section 2. Secondly, we show in Section 3 that the P-FedAvg algorithm converges to the performative stable solution and has a $\mathcal{O}(1/T)$ convergence rate with both the full and partial participation schemes under mild assumptions similar to those in prior works. Finally, in doing so we also introduce some novel proof techniques: we prove convergence without a bounded gradient assumption and use a more relaxed Assumption 2.6. This technique can be directly applied to conventional FL, which is a special case of the performative setting.

Our work is closely related to federated learning Li et al. (2020a); Karimireddy et al. (2020); Wang et al. (2020); Haddadpour et al. (2021); Zhu et al. (2021b); Li & Wang (2019); Lin et al. (2020); Guo et al. (2021); Casado et al. (2022); Rizk et al. (2020); Hosseinalipour et al. (2022); Zhu et al. (2021a); Eichner et al. (2019); Ding et al. (2020), multi-agent PP Li et al. (2022); Narang et al. (2022); Raab & Liu (2021), and strategic classification and regression Hardt et al. (2016); Kleinberg & Raghavan (2020); Shavit et al. (2020); Haghtalab et al. (2020), where we provide detailed descriptions of the related works in Appendix A.

We discuss the system design hyperparameters and the solution concepts in Appendix C. We also perform numerical experiments on synthetic and real-world datasets under the performative setting, where we adopt different distribution shifts $\mathcal{D}_i(\boldsymbol{\theta})$. Our results demonstrate that P-FedAvg converges in both performative classification and regression problems. Additionally, we evaluated the impact of various system design hyperparameters. Please find the results in Appendix D.

2 Problem Formulation

To help with the understanding of performative federated learning, we first recall the performative prediction problem in Perdomo et al. (2020). Consider a typical loss minimization problem where the data distribution experiences a shift induced by the model parameter, expressed as a mapping $\mathcal{D}(\boldsymbol{\theta})$. This mapping from model to distribution is a key concept in performative prediction. It indicates the distribution is *not static* and the shift is *model-dependent*. The objective function is thus given by $f(\boldsymbol{\theta}) := \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)]$, where ℓ denotes the loss function, $Z = (X, Y)$. Then the performative optimal (PO) solution is $\boldsymbol{\theta}^{PO} := \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$. Nonetheless, obtaining a PO solution proves challenging due to the distribution's reliance on $\boldsymbol{\theta}$, which undermines traditional risk minimization. Moreover, the map $\mathcal{D}_i(\cdot)$ is generally unknown to the decision maker. Therefore, Perdomo et al. (2020) also introduces a second, decoupled objective function, also called the performatively stable (PS) model, which separates decision parameters ($\boldsymbol{\theta}$) from deployed parameters ($\tilde{\boldsymbol{\theta}}$): $f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \mathbb{E}_{Z\sim\mathcal{D}(\tilde{\boldsymbol{\theta}})}[\ell(\boldsymbol{\theta}; Z)]$. Minimizing this objective achieves minimal risk for the distribution induced by the deployed parameters, eliminating the need for retraining, which makes it more practical. The PS solution is defined as $\boldsymbol{\theta}^{PS} := \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{\theta}^{PS})$. Perdomo et al. (2020) showed that $\boldsymbol{\theta}^{PS} \neq \boldsymbol{\theta}^{PO}$ in general. We next consider a distributed setting and introduce performative federated learning.

Consider a system with N clients and a server, where client i's data distribution is $\mathcal{D}_i(\boldsymbol{\theta})$, supported on $Z = (X, Y) \subseteq \mathbb{R}^M$. $\boldsymbol{\theta} \in \mathbb{R}^m$ denotes the decision (model) parameters deployed on the i-th client. We consider the general case where clients can have heterogeneous distributions $\mathcal{D}_i(\boldsymbol{\theta}) \neq \mathcal{D}_j(\boldsymbol{\theta})$, and each client represents a $p_i > 0$ fraction of the total data population, $\sum_{i=1}^N p_i = 1$. The system aims to minimize the weighted average loss across all agents, which is given by the performative optimal objective $\boldsymbol{\theta}^{PO} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \sum_{i=1}^N p_i \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z_i)]$. This objective can typically model the strategic learning problem with different sub-populations in the system, where each client corresponds to a sub-population. Each sub-population may differ in some attributes so that they respond to the decision parameters differently, e.g., due to different action costs Milli et al. (2019); Hu et al. (2019); Braverman & Garg (2020); Zhang et al. (2022); Jin et al. (2022). The decision maker uses a common decision rule for the entire population and aims to minimize the expected loss, and p_i represents the population fraction of each sub-population. Correspondingly, the decoupled/performative stable objective is $f_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \mathbb{E}_{Z_i \sim \mathcal{D}_i(\tilde{\boldsymbol{\theta}})}[\ell(\boldsymbol{\theta}; Z_i)]$, $f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \sum_{i=1}^{N} p_i f_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$, where the first argument denotes the client's decision parameter, and the second argument is the deployed parameters, which determine the distribution of the samples together with $\mathcal{D}_i(\cdot)$. Similar to Perdomo et al. (2020), we then introduce the performative stable solution

$$\boldsymbol{\theta}^{PS} := \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} p_i \mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta}^{PS})}[\ell(\boldsymbol{\theta}; Z_i)] = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{\theta}^{PS}).$$

Note that this is a fixed point equation with $\boldsymbol{\theta}^{PS}$ as a fixed point. We establish the uniqueness of the PS solution under certain assumptions in Proposition 2.7. Moreover, in Proposition 2.8, we show that the distance between $\boldsymbol{\theta}^{PS}$ and $\boldsymbol{\theta}^{PO}$ is bounded. We then make some key assumptions.

Assumption 2.1 (Strong Convexity)**.** Given any $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^m$, $f(\cdot, \tilde{\boldsymbol{\theta}})$ is μ-strongly convex in $\boldsymbol{\theta}$, i.e., $f(\boldsymbol{\theta}'; \tilde{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \langle \nabla f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{\mu}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2, \forall \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^K$.

Assumption 2.2 (Smoothness)**.** The loss function $\ell(\boldsymbol{\theta}; z)$ is L-smooth, i.e., $\|\nabla \ell(\boldsymbol{\theta}; \boldsymbol{z}) - \nabla \ell(\boldsymbol{\theta}'; \boldsymbol{z}')\|_2 \leq L(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 + \|\boldsymbol{z} - \boldsymbol{z}'\|_2)$.

Assumption 2.3 (Distribution Mapping Sensitivity)**.** For any $i = 1, \ldots, n$ there exists $\epsilon_i > 0$ such that $\mathcal{W}_1(\mathcal{D}_i(\boldsymbol{\theta}), \mathcal{D}_i(\boldsymbol{\theta}')) \leq \epsilon_i \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \ \forall \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^m$, where $\mathcal{W}_1(\mathcal{D}, \mathcal{D}'))$ is the 1-Wasserstein distance under L_2 norm between the distributions $\mathcal{D}, \mathcal{D}'$.

Lemma 2.4 (Continuity of ∇f_i)**.** *Under Assumption 2.2 and 2.3, for any* $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \mathbb{R}^m$, $\|\nabla f_i(\boldsymbol{\theta}_0; \boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}_1; \hat{\boldsymbol{\theta}})\|_2 \leq L\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1\|_2 + L\epsilon_i\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2$.

Assumption 2.5 (Stochastic Gradient Variance Bound)**.** For any $i = 1, \ldots, N$ and $\boldsymbol{\theta} \in \mathbb{R}^m$, there exists $\sigma \geq 0$ such that $\mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta})}\|\nabla \ell(\boldsymbol{\theta}; Z_i) - \nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|_2^2 \leq \sigma^2(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^{PS}\|_2^2)$.

Assumption 2.6 (Local Gradient Variance Bound)**.** For any $i = 1, \ldots, N$ and $\boldsymbol{\theta} \in \mathbb{R}^m$, there exists $\varsigma \geq 0$ such that $\|\nabla f(\boldsymbol{\theta}; \boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|_2^2 \leq \varsigma^2(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^{PS}\|_2^2)$.

Assumption 2.1, 2.2 and 2.3 are presented in Perdomo et al. (2020) and have been extended to the decentralized case here. In Assumption 2.1, we do not require strong convexity for every single f_i but only the weighted average f. Assumption 2.2 and 2.3 together induce the smoothness of $f_i(\cdot, \cdot)$, which is a result of Lemma 2.1 in Drusvyatskiy & Xiao (2022) and will be used in the later proofs. Assumption 2.5, 2.6 are made in decentralized performative predictions Li et al. (2022). An elaborate on Assumption 2.6 can be found in Appendix B. Instead of the commonly used assumption of bounded gradient expectation, Assumption B.1 Li et al. (2020b), we use Assumption 2.6, which is a weaker assumption. Assumption 2.6 better characterizes the system heterogeneity, as we show how the heterogeneity impacts convergence (more details are in Theorem 3.1, 3.2, and 3.3).

Properties of the PS Solution Define the *average sensitivity* as $\bar{\epsilon} := \sum_{i=1}^{N} p_i \epsilon_i$, and the mapping $\Phi(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}' \in \mathbb{R}^m} f(\boldsymbol{\theta}', \boldsymbol{\theta})$. We can establish the existence and uniqueness of the PS solution. The proofs are in Appendix E.

Proposition 2.7 (Uniqueness of $\boldsymbol{\theta}^{PS}$)**.** *Under Assumptions 2.1, 2.2 and 2.3, if $\bar{\epsilon} < \mu/L$, then $\Phi(\cdot)$ is a contraction mapping with the unique fixed point $\boldsymbol{\theta}^{PS} = \Phi(\boldsymbol{\theta}^{PS})$; if $\bar{\epsilon} \geq \mu/L$, then there is an instance where any sequence generated by $\Phi(\cdot)$ will diverge.*

Proposition 2.7 establishes a sufficient and necessary condition for the existence of $\boldsymbol{\theta}^{PS}$, similar to Li et al. (2022). This condition only depends on the average sensitivity $\bar{\epsilon}$, which implies that we may still have a unique performative stable solution $\boldsymbol{\theta}^{PS}$ for the whole system even if certain clients do not. The following proposition further validates the quality of $\boldsymbol{\theta}^{PS}$ in terms of its distance to $\boldsymbol{\theta}^{PO}$.

Proposition 2.8 (Distance $\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2$ Bound)**.** *Under Assumption 2.1 and 2.3, suppose that the loss $\ell(\boldsymbol{\theta}; Z)$ is L_z-Lipschitz in Z, then for every performative stable solution $\boldsymbol{\theta}^{PS}$ and every performative optimal solution $\boldsymbol{\theta}^{PO}$, we have $\|\boldsymbol{\theta}^{PS} - \boldsymbol{\theta}^{PO}\|_2 \leq \big(2L_z\bar{\epsilon}\big)/\mu$.*

The `P-FedAvg` Algorithm We now introduce the proposed `P-FedAvg` algorithm. In `P-FedAvg`, the clients communicate with the server every E local updates. Denote $\mathcal{I}_E := \{nE | n = 1, 2, \dots\}$ as the set of aggregation steps.

Full client participation. All clients communicate with the server at every aggregation step and update the local models $\boldsymbol{\theta}_i^{t+1}$ based on the following: let $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$, then

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{\theta}_i^t - \eta_t \nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}); \quad \boldsymbol{\theta}_i^{t+1} = \begin{cases} \sum_{j=1}^N p_j \boldsymbol{w}_j^{t+1} & \text{if } t+1 \in \mathcal{I}_E \\ \boldsymbol{w}_i^{t+1} & \text{o.w.} \end{cases}$$

Partial client participation. At each aggregation step, $K (< N)$ clients are sampled to communicate with the server and update the local models $\boldsymbol{\theta}_i^{t+1}$ based on the following: let $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$, denote the chosen clients in t-th step as a size-K set $\mathcal{S}_t := \{i_1, \dots, i_K\} \in [N]$, then

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{\theta}_i^t - \eta_t \nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}); \quad \boldsymbol{\theta}_i^{t+1} = \begin{cases} \text{samples } \mathcal{S}_{t+1}, \text{ average } \left\{\boldsymbol{w}_{t+1}^k\right\}_{k \in \mathcal{S}_{t+1}} & \text{if } t+1 \in \mathcal{I}_E \\ \boldsymbol{w}_i^{t+1} & \text{o.w.} \end{cases}$$

We consider two schemes of partial participation:

1. (**Scheme I** Sahu et al. (2018)) The server establishes $\mathcal{S}_{t+1}$ by ***i.i.d. with replacement*** sampling an index $k \in \{1, \cdots, N\}$ with probabilities $p_1, \cdots, p_N$ for K times, and averages the parameters by $\boldsymbol{\theta}_i^{t+1} = \frac{1}{K} \sum_{k \in \mathcal{S}_{t+1}} \boldsymbol{w}_k^{t+1}$.
2. (**Scheme II**) The server samples $\mathcal{S}_{t+1}$ uniformly ***without replacement*** and averages the parameters by $\boldsymbol{\theta}_i^{t+1} = \sum_{k \in \mathcal{S}_{t+1}} p_k \frac{N}{K} \boldsymbol{w}_k^{t+1}$. Note that when the probabilities $\{p_k\}$ are not the same, one cannot ensure $\sum_{k \in \mathcal{S}_{t+1}} p_k \frac{N}{K} = 1$ Li et al. (2020b).

The `P-FedAvg` requires two rounds of communications, aggregation, and broadcast for every E iterations. So at time step T, the system completes $2\lfloor T/E \rfloor$ communications. We follow the setting in Li et al. (2020b) where the server aggregates based on the chosen scheme and broadcasts the aggregated parameters to all clients.

3 Convergence Analysis

In this section, we show that the `P-FedAvg` converges to the unique $\boldsymbol{\theta}^{PS}$ at a rate of $\mathcal{O}(1/T)$ under the assumptions made in Section 2, which holds for all above-introduced schemes. The definition of the constants can be found in Appendix F and G.

Theorem 3.1 (Full Participation). *Consider `P-FedAvg` with full participation and diminishing step size $\eta_t = \frac{2}{\tilde{\mu}(t+\gamma)}$, where $\gamma = \max\left\{\frac{2}{\tilde{\mu}\hat{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2 + 2E)c_3}\right\}$. Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, it holds $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma+t}$, $\forall t$ where $\upsilon = \max\left\{\frac{4B}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.*

Theorem 3.2 (Partial Participation, Scheme I). *Consider `P-FedAvg` with partial participation (scheme I) and a diminishing step size $\eta_t = \frac{2}{\tilde{\mu}(t+\gamma)}$, where $\gamma = \max\left\{\frac{2}{\tilde{\mu}\tilde{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2 + 10E + 6)c_3}\right\}$. Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, it holds $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma+t}$, $\forall t$ where $\upsilon = \max\left\{\frac{4B_1}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.*

Theorem 3.3 (Partial Participation, Scheme II). *Consider `P-FedAvg` with partial participation (scheme II) and a diminishing step size $\frac{2}{\tilde{\mu}(t+\gamma)}$, where $\gamma = \max\left\{\frac{2}{\tilde{\mu}\tilde{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2 + 10E + 6)c_5}\right\}$. Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, it holds that $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma+t}$, $\forall t$ where $\upsilon = \max\left\{\frac{4B_2}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.*

Scheme II requires $p_i = \frac{1}{N}, \forall i$, which violates the unbalanced nature of FL. One solution in Li et al. (2020b) is scaling the local objectives to $g_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = p_i N f_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$, and then the global objective is a simple average of the scaled local objectives $f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \sum_{i=1}^N p_i f_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{i=1}^N g_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$. We need to be careful with the Assumptions in Section 2 since scaling the objective will change those properties. The convergence theorems still hold if we replace $L, \mu, \sigma, \varsigma$ with $L' := q_{max} L, \mu' := q_{min} \mu, \sigma' := \sqrt{q_{max}}\sigma, \varsigma' := \sqrt{q_{max}}\varsigma$, where $q_{max} := N \cdot \max_i p_i, q_{min} := N \cdot \min_i p_i$.

4 Numerical Experiments

Weighted Gaussian mean performative prediction. As a numerical simulation, we perform P-FedAvg to estimate the mean of heterogeneous Gaussian data under performative effects and examine the impact of the hyperparameters, the sampling schemes, and client heterogeneity. We consider $N = 25$ clients, with the i-th client minimizing the loss function $\ell(\theta; Z_i) := (\theta - Z)^2/2$, $\theta, Z \in \mathbb{R}$ on data $Z_i \sim \mathcal{D}_i(\theta) := \mathcal{N}(m_i + \epsilon_i\theta, \sigma^2)$. For this loss function, we have $\mu = 1$, $L = 1$. For $\bar{\epsilon} \in [0, 1)$, the PS solution is $\theta^{PS} = \frac{\sum_{i=1}^{N} p_i m_i}{1-\bar{\epsilon}}$; while θ^{PS} does not exist when $\bar{\epsilon} \geq 1$. Denote the weighted average of m_i as $\overline{m} = \sum_{i=1}^{N} p_i m_i$ and the variance as $\text{Var}(m) = \sum_{i=1}^{N} p_i(m_i - \overline{m})^2$. In experiment, we set $\bar{\epsilon} = 0.9$, $\overline{m} = 10$.

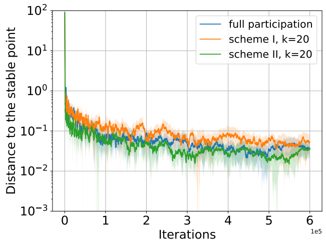
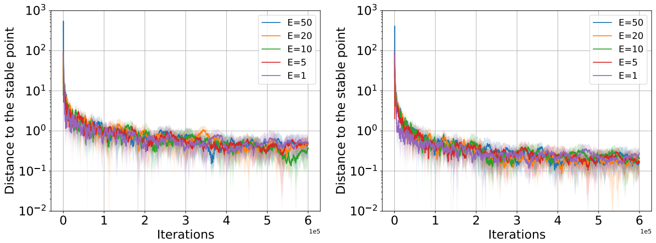


Figure 1: Distance to the performative stable solution vs. the number of iterations for full participation, Scheme I, and Scheme II.



(a) Scheme I

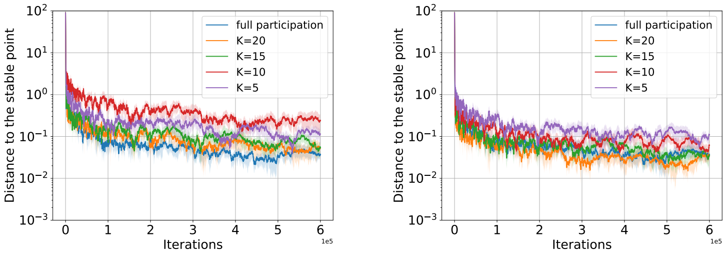
(b) Scheme II

Figure 2: Impact of E on Scheme I and Scheme II. $K = 20$, $\text{Var}(m) = 0.6$, $\text{Var}(\epsilon) = 0.1$ for both (a) and (b). For (b), $p_i = \frac{1}{25}$.

Figure 1 shows P-FedAvg converges to the performative stable solution in all three communication settings: full participation and the two schemes for the partial participation. Interestingly, partial participation with scheme II converges the fastest in this experiment. Despite the full participation scheme having the lowest upper bound on the number of iterations sufficient to convergence, our experimental results show that the actual convergence behaviors of all three schemes are very similar and weakly depend on K, especially when $p_i = \frac{1}{N}$.

Impact of E. We conduct an experiment to compare the performance of our algorithm with a variety of E values, under a homogeneous system. Figure 2 shows the result on both sampling schemes, with $K = 20$. A slightly larger E leads to faster convergence. However, an extremely large $E(E = 50$ in the experiment) can also cause slower convergence. Since at this case, the clients deviate too much at each aggregation, which causes low efficiency issues. In real world scenarios, as the communication cost changes, E should be carefully chosen.

Impact of K. Figure 3 shows the convergence of FedAvg under different k values, For scheme I, larger k leads to faster convergence. While for scheme II, as k increasing, the convergence rate will first increase and then decrease.



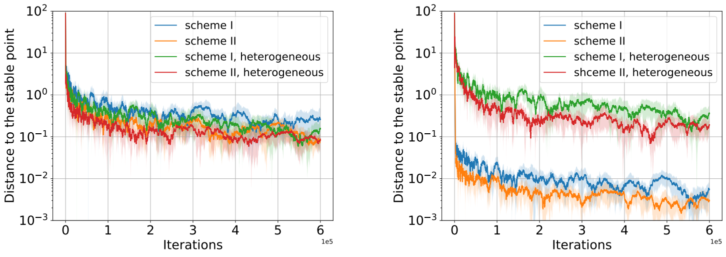
(a) Scheme I

(b) Scheme II

Figure 3: Impact of K on Scheme I and Scheme II . $E = 5$, $\text{Var}(m) = 0.6$, $\text{Var}(\epsilon) = 0.1$ for both (a) and (b). For (b), $p_i = \frac{1}{25}$.

Impact of sampling schemes. Figure 1 also compares different schemes. We can see if the clients' data are uniformly sampled ($p_i = \frac{1}{N}$), then scheme II achieves a better convergence rate, which conforms to our theoretical result because $B_1 > B_2$.

Data heterogeneity and shifting heterogeneity. In Figure 4 we test our algorithm under data heterogeneity. Specifically, we set m and ϵ to have large variances, respectively. In this example, m mainly captures the data heterogeneity and ϵ capture the shifting heterogeneity. This experiment shows our algorithm still converges under a certain amount of heterogeneity. Comparing the performance of our algorithm on both figures, we can see shifting heterogeneity is the main factor in performative federated learning.



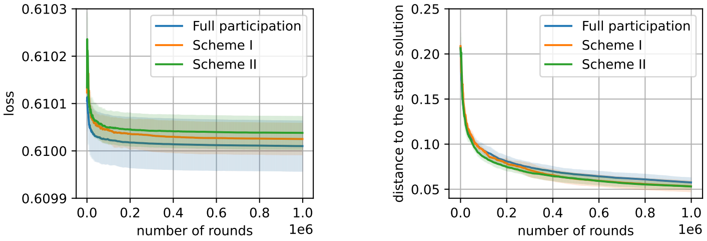
(a) Heterogeneity in m_i

(b) Heterogeneity in ϵ_i

Figure 4: Impact of heterogeneity on the two schemes of partial participation. The impact of m_i is shown in (a) and the impact of ϵ_i is shown in (b). $K = 20, p_i = \frac{1}{25}$ and. In (a), Var$(m) = 6$ for hetergeneous case and 0 for homogeneous case, Var$(\epsilon) = 0.1$. In (b), Var$(\epsilon) = 0.6$ for hetergeneous case and 0.1 for homogeneous case, Var$(m) = 0.6$.

Credit score strategic classification. To show the performance of `P-FedAvg` on a real world dataset, we follow Perdomo et al. (2020) and use the same Kaggle GiveMeSomeCredit dataset, where a bank predicts whether loan applicants are creditworthy. The features consist of the information about an individual, and the target is 1 if the individual defaulted on a loan, and 0 otherwise. We use the same strategic setting as in Perdomo et al. (2020) where the applicants can manipulate their features in (1) revolving utilization of unsecured lines, (2) number of open credit lines and loans, and (3) number real estate loans or lines. The strength of manipulation for the i-th population is controlled by ϵ_i. We equally partition the training set into 10 subsets and distributed it to 10 clients, and thus $p_i = 0.1, \forall i$. The sensitivities ϵ_i for the 10 clients are independently and uniformly sampled from $[0.9, 1.1]$. We set $K = 5$ in partial participation. We train a logistic regression binary classifier. In each round of `P-FedAvg`, we perform $E = 5$ gradient descent steps on a random minibatch of size 4. A discussion on the effect of the batch size can be found in Appendix D.3.

Figure 5 shows the loss function and the distance to the PS solution as the number of deployment rounds increases. The mean and 1 standard deviation error bar are generated from 5 experiments with different random seeds. Similar to the numerical simulation, the actual convergence behaviors of all three schemes are very similar.



(a) Training losses

(b) Distances to $\boldsymbol{\theta}^{PS}$

Figure 5: The losses (a) and the distances to the PS solution (b) for the full participation, Scheme I and Scheme II.

References

Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. *CoRR*, abs/2011.03885, 2020. URL `https://arxiv.org/abs/2011.03885`.

Fernando E Casado, Dylan Lema, Marcos F Criado, Roberto Iglesias, Carlos V Regueiro, and Senén Barro. Concept drift detection and adaptation for federated and continual learning. *Multimedia Tools and Applications*, 81(3):3397–3419, 2022.

Yatong Chen, Jialu Wang, and Yang Liu. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 2020.

Yucheng Ding, Chaoyue Niu, Yikai Yan, Zhenzhe Zheng, Fan Wu, Guihai Chen, Shaojie Tang, and Rongfei Jia. Distributed optimization over block-cyclic data. *arXiv preprint arXiv:2002.07454*, 2020.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions, 2020. URL `https://arxiv.org/abs/2011.11173`.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Hubert Eichner, Tomer Koren, Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1764–1773. PMLR, 2019.

Yongxin Guo, Tao Lin, and Xiaoying Tang. Towards federated learning on time-evolving heterogeneous data. *arXiv preprint arXiv:2112.13246*, 2021.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Wang. Maximizing welfare with incentive-aware evaluation mechanisms. pp. 160–166, 07 2020. doi: 10.24963/ijcai.2020/23.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. pp. 111–122, 01 2016. doi: 10.1145/2840728.2840730.

Seyyedali Hosseinalipour, Su Wang, Nicolo Michelusi, Vaneet Aggarwal, Christopher G Brinton, David J Love, and Mung Chiang. Parallel successive learning for dynamic distributed model training over heterogeneous wireless networks. *arXiv preprint arXiv:2202.02947*, 2022.

Lily Hu, Nicole Immorlica, and Jennifer Vaughan. The disparate effects of strategic manipulation. pp. 259–268, 01 2019. doi: 10.1145/3287560.3287597.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4641–4650. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/izzo21a.html`.

Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3998–4035. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/izzo22a.html`.

Kun Jin, Xueru Zhang, Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Incentive mechanisms for strategic classification and regression problems. In David M. Pennock, Ilya Segal, and Sven Seuken (eds.), *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pp. 760–790. ACM, 2022. doi: 10.1145/3490486.3538300. URL https://doi.org/10.1145/3490486.3538300.

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

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation*, 8:1–23, 11 2020. doi: 10.1145/3417742.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3164–3186. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/li22c.html.

Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents, 2022. URL https://arxiv.org/abs/2209.03811.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=HJxNAnVtDS.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/33e75ff09dd601bbe69f351039152189-Paper.pdf.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6917–6926. PMLR, 13–18 Jul 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7710–7720. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/miller21a.html.

Smitha Milli, John Miller, Anca Dragan, and Moritz Hardt. The social cost of strategic classification. pp. 230–239, 01 2019. doi: 10.1145/3287560.3287576.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *CoRR*, abs/2201.03398, 2022. URL https://arxiv.org/abs/2201.03398.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/perdomo20a.html`.

Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34:26053–26065, 2021.

Mitas Ray, Lillian J. Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 8081–8088. AAAI Press, 2022. URL `https://ojs.aaai.org/index.php/AAAI/article/view/20780`.

Elsa Rizk, Stefan Vlaski, and Ali H. Sayed. Dynamic federated learning. *CoRR*, abs/2002.08782, 2020. URL `https://arxiv.org/abs/2002.08782`.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *CoRR*, abs/1812.06127, 2018. URL `http://arxiv.org/abs/1812.06127`.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression, 2020.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6: 1646–1651, 2022. doi: 10.1109/LCSYS.2021.3124187.

Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. Fairness interventions as (Dis)Incentives for strategic manipulation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26239–26264. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/zhang22l.html`.

Chen Zhu, Zheng Xu, Mingqing Chen, Jakub Konečnỳ, Andrew Hard, and Tom Goldstein. Diurnal or nocturnal? federated learning of multi-branch networks from periodically shifting distributions. In *International Conference on Learning Representations*, 2021a.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 12878–12889. PMLR, 2021b.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15257–15269. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/812214fb8e7066bfa6e32c626c2c688b-Paper.pdf`.

A Related Works

Federated Learning. Our work is strongly related to the literature on federated learning (FL). Although many studies have tried to address client heterogeneity in FL through constrained gradient optimization and knowledge distillation Li et al. (2020a); Karimireddy et al. (2020); Wang et al. (2020); Haddadpour et al. (2021); Zhu et al. (2021b); Li & Wang (2019); Lin et al. (2020), most of them still assume the data is static without considering the distribution shifts. To the best of our knowledge, only a few recent works consider distribution shifts in FL Guo et al. (2021); Casado et al. (2022); Rizk et al. (2020); Hosseinalipour et al. (2022); Zhu et al. (2021a); Eichner et al. (2019); Ding et al. (2020). For example, Guo et al. (2021) considered FL with time-evolving clients where the time-drift of each client is modeled as a *time-independent* additive noise with *zero-mean* and *bounded variance*. Casado et al. (2022) proposed an FL algorithm adaptable to distribution drifts; it monitors the confidence scores of the model prediction throughout the learning process and assumes the drift happens whenever there is a substantial drop in confidence scores. Rizk et al. (2020) also studied dynamic FL and assumed the true model under time-evolving data follows a *random walk*. Hosseinalipour et al. (2022) considered FL with dynamic clients and modeled the drift using the variation in *local loss* over two consecutive time steps. Zhu et al. (2021a); Eichner et al. (2019); Ding et al. (2020) considered the *periodical* distribution shift of client population in FL; they assume the *block-cyclic* structure where the clients from two different time zones alternately participate in training.

Performative Prediction. In addition to the one we discussed in the introduction that focuses on the centralized setting for performative prediction, more recently, Li et al. (2022) formalize the multi-agent/player performative predictions where agents try to learn a common decision rule but have heterogeneous distribution shifts (responses) to the model, and study the convergence of decentralized algorithms to the PS solution. The decentralized performative predictions capture the heterogeneity in agents'/clients' responses to the decision model and avoid centralized data collection for training. This work provides inspiration for our formulation of the performative federated learning framework, and our proposed `P-FedAvg` can be viewed as a substantial algorithmic extension that supports unbalanced data, much less frequent synchronizations, and partial device participation. Narang et al. (2022) propose a decentralized multi-player performative prediction framework where the players react to competing institutions' actions. Raab & Liu (2021) proposes a replicator dynamics model with label shift.

Strategic Classification and Regression. As discussed in Perdomo et al. (2020), performative prediction can be used to solve repeated strategic classification and regression problems. We can use Stackelberg games to model these problems, where the decision maker moves in the first stage by designing, publishing, and committing to a decision rule, then the agents move in the second stage, best responding to the decision rule by manipulating their features to get more desirable decision outcomes, and such manipulation can be modeled by the distribution shift mappings. Conventional strategic learning literature focus on finding the Stackelberg equilibrium Hardt et al. (2016); Kleinberg & Raghavan (2020); Shavit et al. (2020); Haghtalab et al. (2020), i.e., the PO solution where the decision maker and the agents know each others' utilities, whereas performative prediction can find the PS solution in repeated strategic learning problems regardless of the knowledge on the utilities.

B Elaboration on Assumption 2.6

In this section we elaborate on Assumption 2.6, and explain reasons for using it over another commonly used assumption in federated learning Li et al. (2020b), which is

Assumption B.1.

$$\mathbb{E}_{Z_i \sim \mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla \ell(\boldsymbol{\theta}; Z_i)\|_2^2] \leq G^2. \tag{1}$$

First, it can be shown that equation B.1 implies Assumption 2.6, thus Assumption 2.6 is weaker than equation B.1. To see this: when equation B.1 holds, let $\varsigma^2 = 4G^2$, then $\|\nabla f(\boldsymbol{\theta};\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2 \leq 2\left\|\nabla f(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2 + 2\|\nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\right\|_2^2 \leq 4G^2 = \varsigma^2$.

We further give a concrete example where equation B.1 does not hold but Assumption 2.6 holds.

Example B.2. *Suppose we have a two-client Gaussian mean estimation problem* $\ell(\theta, Z) = \frac{1}{2}(\theta - Z)^2$ *where* $\theta, Z \in \mathbb{R}$, $\mathcal{D}_1(\theta) = \mathcal{N}(\frac{1}{2}\theta, \sigma^2)$, $\mathcal{D}_2(\theta) = \mathcal{N}(-\frac{1}{2}\theta, \sigma^2)$, *and* $p_1 = p_2 = \frac{1}{2}$. *Then*

$\mathbb{E}_{Z_1\sim\mathcal{D}_1(\theta)}[\|\nabla\ell(\theta;Z_1)\|_2^2] = \mathbb{E}_{Z_1\sim\mathcal{D}_1(\theta)}[(\theta - Z_1)^2] = \sigma^2 + (\mathbb{E}_{Z_1\sim\mathcal{D}_1(\theta)}[\theta - Z_1])^2 = \frac{1}{4}\theta^2 + \sigma^2$ *and* $\mathbb{E}_{Z_2\sim\mathcal{D}_2(\theta)}[\|\nabla\ell(\theta;Z_2)\|_2^2] = \frac{9}{4}\theta^2 + \sigma^2$ *which all go to infinity when* θ *goes to infinity. Thus equation B.1 does not hold. On the other hand,* $\nabla f_1(\theta;\theta) = \frac{1}{8}\theta$, $\nabla f_2(\theta;\theta) = \frac{9}{8}\theta$, *and* $\nabla f(\theta;\theta) = \frac{5}{8}\theta$, $\theta^{PS} = 0$, *then by taking* $\varsigma = \frac{1}{2}$, *we can verify Assumption 2.6 holds.*

Secondly, equation B.1 also implies Assumption 2.5: when equation B.1 holds, letting $\sigma^2 = G^2$ leads to $\mathbb{E}[\|\nabla l(\boldsymbol{\theta};Z_i) - \nabla f_i(\boldsymbol{\theta},\boldsymbol{\theta})\|_2^2] \leq \mathbb{E}[\|\nabla l(\boldsymbol{\theta};Z_i)\|_2^2] \leq G^2 = \sigma^2$. On the other hand, Assumption 2.6 does not imply Assumption 2.5.

It turns out that Assumption 2.6 better characterizes the system heterogeneity, as we show how the heterogeneity impacts convergence (more details are in Theorem 3.1, 3.2, and 3.3).

C Discussions on the Algorithm and Solution

We will only discuss with respect to the aggregation step in this sub-section for convenience, denoted as $T \in \mathcal{I}_E$, then we can simply use $\frac{T}{E}$ when dividing E. Note for a general step t, we only need to use $\lfloor \frac{t}{E} \rfloor$ to obtain an integer.

Choice of E. We are interested in the total time we need to achieve an ϵ accuracy, and how this total time changes with E. We use our results in Theorem 3.1, 3.2, and 3.3, and denote $T_\epsilon := \frac{v}{\epsilon} - \gamma$ as the number of computation steps that is sufficient to guarantee an ϵ-accuracy. To connect T_ϵ to the total time needed, suppose the expected time for each communication step is C times the expected time of each computation step, then the total time required for ϵ-accuracy is linear in $T_\epsilon + C \cdot \frac{T_\epsilon}{E}$. Below we separately analyze the influence of E on $\frac{T_\epsilon}{E}$ and T_ϵ, and then discuss how to choose the optimal E for different C values.

Let $B_0 := B$ in Theorem 3.1 for full participation and γ_i $(i = 0, 1, 2)$ denotes the γ in Theorem 3.1, 3.2, and 3.3 respectively. Then in Theorem 3.1, 3.2, and 3.3, T_ϵ is dominated by $\mathcal{O}\big(4B_i/\tilde{\mu}^2 + \gamma_i\mathbb{E}[\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\big)$ where $i = 0, 1, 2$. From the definition, we know that B_i $(i = 0, 1, 2)$ is almost a constant w.r.t. E and γ_i is of $\mathcal{O}(E^2 \log E)$. This means that when E grows, the total update steps to reach ϵ-accuracy, T_ϵ will grow, while the number of aggregation steps needed, $\frac{T_\epsilon}{E}$ will first grow and then decrease.

Now we consider $T_\epsilon + C \cdot \frac{T_\epsilon}{E}$, the total time needed to reach ϵ-accuracy. From the above analysis, we know it is of order $\mathcal{O}(E^2 \log E) + C \cdot \mathcal{O}(E \log E) + C \cdot \mathcal{O}(\log E/E)$. When communication is fast, i.e., C is small, $\mathcal{O}(E^2 \log E)$ is the dominating term, and we can focus more on the number of computation iterations T_ϵ, and smaller E values are preferable. However, when C is large, $C \cdot \mathcal{O}(E \log E) + C \cdot \mathcal{O}(\log E/E)$ becomes the dominating term, and we should focus more on the number of communication rounds $\frac{T_\epsilon}{E}$ and some middle E values are preferable.

Choice of K. Again T_ϵ is dominated by $\mathcal{O}\big(4B_i/\tilde{\mu}^2 + \gamma_i\mathbb{E}[\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\big)$ where $i = 1, 2$. Then by the formulae of B_i $(i = 1, 2)$, we know T_ϵ monotonically decreases with K, but the total communication time increases with K due to more severe stragglers' effect. Generally, as we show in Theorem 3.2 and 3.3, the convergence rate has a weak dependence on K. We have empirically observed this phenomenon in Figure 3(a). Therefore, we can set $\frac{K}{N}$ to an appropriate small value to reduce the straggler's effect while keeping the convergence rate.

Choice of sampling schemes. We formalize the two sampling schemes in Section 2 and show their convergence properties in Theorem 3.2 and 3.3. We note that Scheme I has a desirable property that it naturally supports unbalanced clients, so if the server has control over the sampling, Scheme I should be chosen.

But as discussed in Li et al. (2020b), sometimes the server may have no control over the sampling and simply use the first K received results for update. In this case, if the reception times from each client are IID random variables, we can treat this process as uniformly sampling K out of N at random without replacement. Theorem 3.3 showed the convergence, and the discussion on scaling the objectives provides instructions on how to make the system work with arbitrary initial $p_1, \ldots, p_N$ values. However, it's worth noting that when $p_1, \ldots, p_N$ are highly non-uniform, the corresponding L', σ', ς' values will be much larger than from L, σ, ς, and μ' will be much smaller than μ. Then by the formula of $\hat{\eta}_0$ and $\tilde{\eta}_0$, we have to use much smaller starting learning rates and thus slower convergence. However, such a small learning rate may cause the model to fail to train at all. We also empirically show this in Figure 7.

However, an interesting observation is that when $p_i = \frac{1}{N}$, we empirically show in Figure 4 and 5, Scheme II slightly outperforms Scheme I.

Learning rate decay. The learning rate decay is a necessity for stochastic gradient descent (SGD) to converge, even when clients have static, independent and individually distributed (IID) data. The decay is used in Li et al. (2022) in decentralized performative prediction and the necessity for such decay is proved in `FedAvg` with static, non-IID clients. We also empirically show that constant learning rates fail to converge in Figure 6.

$\boldsymbol{\theta}^{PS}$ and $\boldsymbol{\theta}^{PO}$. Here we discuss the relationship between the $\boldsymbol{\theta}^{PS}$ and $\boldsymbol{\theta}^{PO}$ solutions more in depth. In the strategic learning setting, Perdomo et al. (2020) showed that $\boldsymbol{\theta}^{PO}$ is the Stackelberg equilibrium. It's worth noting that $\boldsymbol{\theta}^{PS}$ is not merely an approximation to $\boldsymbol{\theta}^{PO}$, but a natural convergence point of the best response dynamics (BRD). More specifically, when the clients and the decision maker have no information about others' utilities, backward induction is unavailable, and playing the Stackelberg equilibrium is unrealistic. In this case, treating others' strategies in the previous time step as constants, and optimizing one's own strategy accordingly is a rational strategy. Such an optimization step is a best response, and in multi-round sequential strategic learning problems Zrnic et al. (2021), the best responses can form the BRD, and $\boldsymbol{\theta}^{PS}$ is **the convergence point of the BRD**. Although the decision maker's natural best response step is a risk minimization step, the gradient-based `P-FedAvg` can find the same $\boldsymbol{\theta}^{PS}$. Another interesting observation of $\boldsymbol{\theta}^{PS}$ is that if we remove the sequential decision nature, then $\{\boldsymbol{\theta}^{PS}, \mathcal{D}_1(\boldsymbol{\theta}^{PS}), \ldots, \mathcal{D}_N(\boldsymbol{\theta}^{PS})\}$ is a **Nash equilibrium** since no participant has an incentive to unilaterally deviate.

D More Numerical Simulations

D.1 Learning rate decay (Figure 6)

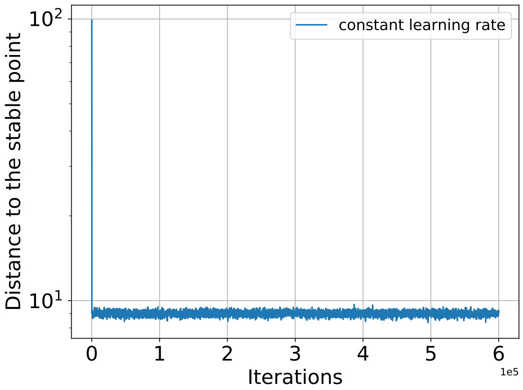


Figure 6: Constant learning rate on ExpGaussian. Full participation with $E = 10$. The learning rate is set to 0.02.

D.2 Scheme II with lower learning rate (Figure 7)

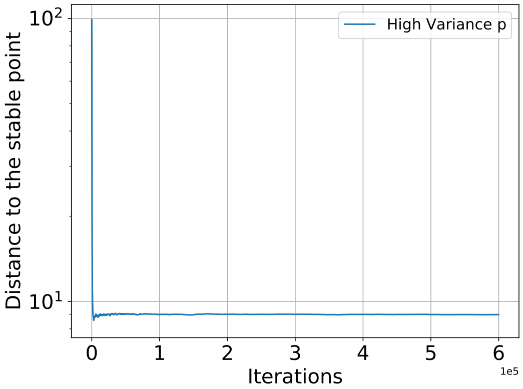


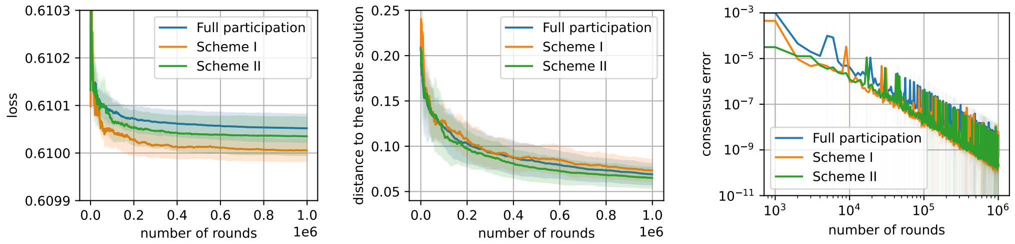
Figure 7: Sampling without replacement on ExpGaussian. The variance of p_i is set to 0.01. The learning rate is set to $\frac{5}{t+10000}$.

D.3 More experiments on the credit score strategic classification

To show the performance of P-FedAvg on a real world dataset, we follow Perdomo et al. (2020) and use the same Kaggle dataset [1], where a bank predicts whether loan applicants are creditworthy. The features consist of the information about an individual, and the target is 1 if the individual defaulted on a loan, and 0 otherwise. We use the same strategic setting as in Perdomo et al. (2020) where the applicants can manipulate their features in (1) revolving utilization of unsecured lines, (2) number of open credit lines and loans, and (3) number real estate loans or lines. The strength of manipulation for the i-th population is controlled by ϵ_i. We equally partition the training set into 10 subsets and distributed it to 10 clients, and thus $p_i = 0.1, \forall i$. The sensitivities ϵ_i for the 10 clients are independently and uniformly sampled from $[0.9, 1.1]$. We set $K = 5$ in partial participation. We train a logistic regression binary classifier. In each round of P-FedAvg, we perform $E = 5$ gradient descent steps on a random minibatch of size 4.

[1] www.kaggle.com/competitions/GiveMeSomeCredit/data

To study the impact of batch size, we change the batch size and plot the losses and distances to the PS solution for the full participation, Scheme I and Scheme II. The results are shown in Figure 8, 9 and 10 for batch size 1, 4, and 16, respectively. The scales of y axes are set equal for convenience of comparison. Using a larger batch size improves the convergence speed for all three schemes, especially for the two schemes of partial participation, both converging as fast as the full participation with batch size 16.

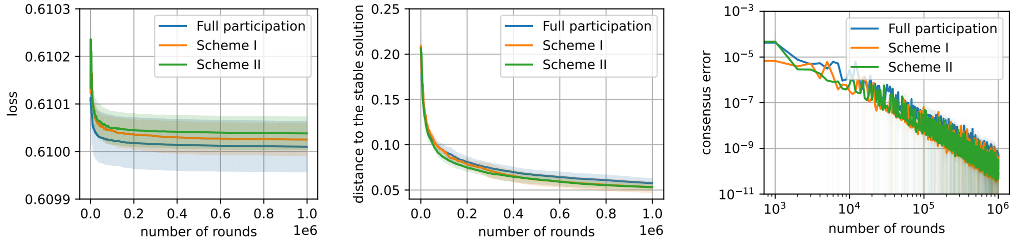


(a)

(b)

(c)

Figure 8: The losses (a) and distances (b) to the PS solution for the full participation, Scheme I and Scheme II using batch size 1 in client gradient descent.

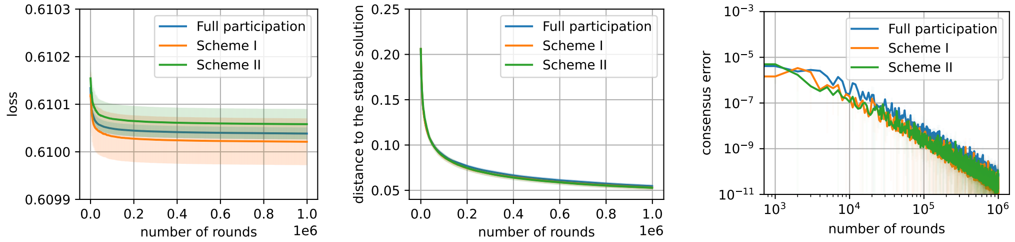


(a)

(b)

(c)

Figure 9: Same as Figure 8, but using batch size 4.



(a)

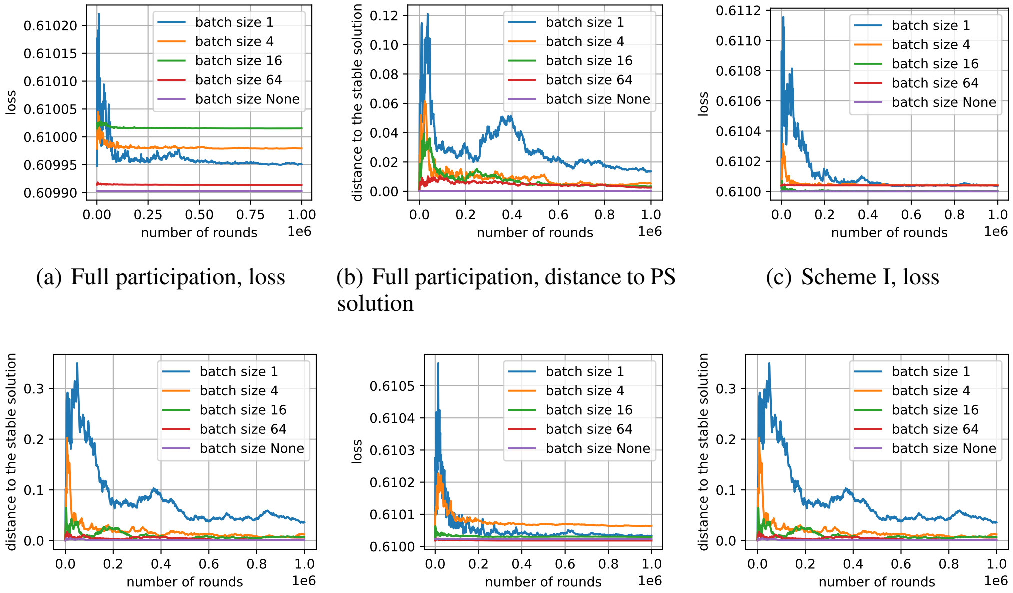
(b)

(c)

Figure 10: Same as Figure 8, but using batch size 16.

To study how batch batch size affect the convergence, we initialize `P-FedAvg` with $\boldsymbol{\theta}^{PS}$, the solution that minimizes the performative objective function. Due to the randomness of minibatch stochastic descent, we expect the parameter to deviate from $\boldsymbol{\theta}^{PS}$ and gradually stabilize back to $\boldsymbol{\theta}^{PS}$ as the algorithm proceeds with decaying step sizes. It can be seen from Figure 11 (b), (c) and (f) that it is indeed the case for batch sizes larger than 1. This motivates our choice of a batch size larger than 1.

Remark D.1. We did not compare `P-FedAvg` with algorithms that **do not have re-sample steps** after every parameter update steps (e.g., SCAFFOLD or conventional FedAvg) due to the performative nature of the prediction problems.



(a) Full participation, loss

(b) Full participation, distance to PS solution

(c) Scheme I, loss

(d) Scheme I, distance to PS solution

(e) Scheme II, loss

(f) Scheme II, distance to PS solution

Figure 11: The loss functions and the distance to $\boldsymbol{\theta}^{PS}$ of P-FedAvg initialized with $\boldsymbol{\theta}^{PS}$.

According to Perdomo et al. (2020), it is clear that such algorithms will converge to values that have constant errors from the PS or PO solutions since they do not account for the model-dependent distribution shifts.

E Proof of Proposition 2.7 and 2.8

Proposition 2.5. (Uniqueness of $\boldsymbol{\theta}^{PS}$) Under Assumptions 2.1, 2.2 and 2.3, define the map $\Phi : \mathbb{R}^m \mapsto \mathbb{R}^m$

$$\Phi(\boldsymbol{\theta}) := arg \min_{\boldsymbol{\theta}' \in \mathbb{R}^M} f(\boldsymbol{\theta}', \boldsymbol{\theta})$$

If $\overline{\epsilon} := \sum_{i=1}^N p_i \epsilon_i < \mu/L$, then $\Phi(\cdot)$ is a contraction mapping with the unique fixed point $\boldsymbol{\theta}^{PS} = \Phi(\boldsymbol{\theta}^{PS})$. On the contrary, if $\overline{\epsilon} \geq \mu/L$, then there is an instance where any sequence generated by $\Phi(\cdot)$ will diverge.

Proof. This proof simulates the proof of Proposition 1 in Li et al. (2022).

Fix $\boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^m$, the optimality condition implies that

$$\sum_{i=1}^N p_i \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) = \mathbf{0}, \quad \sum_{i=1}^N p_i \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') = \mathbf{0}$$

where the gradients are taken w.r.t the first argument in f_i. Then we have

$$\begin{aligned} 0 =& \langle \mathbf{0}, \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \\ =& \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle. \end{aligned}$$

Rearranging the above equation and adding $\sum_{i=1}^N p_i f_i(\Phi(\boldsymbol{\theta}), \boldsymbol{\theta}')$ to both hand sides leads to

$$\begin{aligned} & \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \\ =& \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle. \end{aligned}$$

By strong convexity in assumption 2.1, we have

$$f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') \geq f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') + \langle \nabla f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle + \frac{\mu}{2} \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2,$$

$$f(\Phi(\boldsymbol{\theta})'; \boldsymbol{\theta}') \geq f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') + \langle \nabla f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle + \frac{\mu}{2} \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2,$$

and thus

$$\langle \nabla f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \geq \mu \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2. \tag{2}$$

Applying Lemma 2.4, we have

$$\sum_{i=1}^N p_i \langle \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')) \rangle \leq \sum_{i=1}^N p_i L \epsilon_i \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \cdot \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\|_2. \tag{3}$$

Combine equation 2 and equation 3, we have

$$\|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2 \leq \frac{\sum_{i=1}^N p_i \epsilon_i L}{\mu} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 = \frac{\overline{\epsilon} L}{\mu} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2. \tag{4}$$

Therefore, if $\overline{\epsilon} < \frac{L}{\mu}$, $\Phi(\cdot)$ is a contraction mapping by Banach fixed point theorem and admits a unique fixed point $\boldsymbol{\theta}^{PS}$.

To show the divergence when $\overline{\epsilon} \geq \frac{L}{\mu}$, we consider the following example where $\theta \in \mathbb{R}$, $\frac{L}{\mu} = 1$, $\overline{\gamma} := \sum_{i=1}^N p_i \gamma_i \neq 0$, and

$$l(\theta; Z) = \frac{1}{2}(\theta - Z)^2, Z \sim \mathcal{D}_i(\theta) = \mathcal{N}(\gamma_i + \epsilon_i \theta, 1)$$

we observe

$$\begin{aligned}
f_i(\theta';\theta) =&\mathbb{E}_{Z\sim\mathcal{D}_i(\theta)}[\frac{1}{2}(\theta - Z)^2]\\
=&\mathbb{E}_{\tilde{Z}\sim\mathcal{N}(0,1)}[\frac{1}{2}(\theta' - \gamma_i - \epsilon_i\theta - \tilde{Z})^2]\\
=&\frac{1}{2}(\theta' - \gamma_i - \epsilon_i\theta)^2 + \frac{1}{2},\\
\Phi(\theta) = \operatorname{argmin}_{\theta\in\mathbb{R}}& \sum_{i=1}^{N} p_i(\theta' - \gamma_i - \epsilon_i\theta)^2 = \overline{\epsilon}\theta + \overline{\gamma},
\end{aligned}$$

so by applying $\Phi(\cdot)$ t times, we obtain

$$\Phi^t(\theta) = \overline{\epsilon}^t\theta + (1 + \overline{\epsilon} + \cdots + \overline{\epsilon}^{(t-1)})\overline{\gamma},$$

and since $\overline{\epsilon} \geq \frac{L}{\mu} = 1, \overline{\gamma} \neq 0$, we have $\lim_{t\to\infty} \|\Phi^t(\theta)\|_2 = \infty$. □

Proposition 2.6. Under Assumption 2.1 and 2.3, suppose that the loss $l(\boldsymbol{\theta}; Z)$ is L_z-Lipschitz in Z, let $\overline{\epsilon} := \sum_{i=1}^{N} p_i\epsilon_i$, we have for every performative stable solution and every performative optimal solution $\boldsymbol{\theta}^{PO}$ that

$$\|\boldsymbol{\theta}^{PS} - \boldsymbol{\theta}^{PO}\|_2 \leq \frac{2L_z\overline{\epsilon}}{\mu}.$$

Proof. This proof simulates the proof of Theorem 4.3 in Perdomo et al. (2020).

First by the optimality of $\boldsymbol{\theta}^{PO}$, we have $f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PO}) \leq f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})$. By strong convexity in Assumption 2.1, we have

$$f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PS}) \geq f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}) + \langle\nabla f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}),\ \boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\rangle + \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2 \geq \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2.$$

Further by Assmption 2.3, the the loss $l(\boldsymbol{\theta}; Z)$ is L_z-Lipschitz in Z, and Kantorovich-Rubinstein duality, we have

$$\begin{aligned}
&f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PO})\\
=&\sum_{i=1}^{N} p_i\Big(\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PS})}[l(\boldsymbol{\theta}^{PO};Z_i)] - \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PO})}[l(\boldsymbol{\theta}^{PO};Z_i)]\Big)\\
\leq&\sum_{i=1}^{N} p_iL_z\mathcal{W}_1(\mathcal{D}_i(\boldsymbol{\theta}^{PS}),\mathcal{D}_i(\boldsymbol{\theta}^{PO}))\\
=&\sum_{i=1}^{N} p_iL_z\epsilon_i\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 = L_z\overline{\epsilon}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2.
\end{aligned} \tag{5}$$

where the inequality is a well-know conclusion in optimal tranport theory. Equation 5, we have $L_z\overline{\epsilon}\epsilon_i\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 \geq f(\boldsymbol{\theta}^{PO},\boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PO}) \geq f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}) \geq \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2$, implying that $\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 \leq \frac{2L_z\overline{\epsilon}}{\mu}$.

□

F Proof of Theorem 3.1

F.1 Additional Notation

In our analysis, for the sake of convenience, we will define two additional sequences as $\overline{\boldsymbol{w}}^t := \sum_{i=1}^{N} p_i\boldsymbol{w}_i^t$ and $\overline{\boldsymbol{\theta}}^t := \sum_{i=1}^{N} p_i\boldsymbol{\theta}_i^t$, following that of Li et al. (2020b). We note that $\overline{\boldsymbol{w}}^t$ results from a single step of SGD from $\overline{\boldsymbol{\theta}}^t$. When $t+1 \notin \mathcal{I}_E$, both $\overline{\boldsymbol{w}}^t$ and $\overline{\boldsymbol{\theta}}^t$ are unaccessible. When $t+1 \in \mathcal{I}_E$, we can obtain $\overline{\boldsymbol{\theta}}^t$. In addition, we also define $\overline{\boldsymbol{g}}_t := \sum_{i=1}^{N} p_i\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)$, $\boldsymbol{g}_t := \sum_{i=1}^{N} p_i\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1})$ where $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$. It is clear that in full participation, $\overline{\boldsymbol{w}}^{t+1} = \overline{\boldsymbol{\theta}}^t - \eta_t\boldsymbol{g}_t$ and $\mathbb{E}\boldsymbol{g}_t = \overline{\boldsymbol{g}}_t$. Clearly we have $\overline{\boldsymbol{\theta}}^t = \overline{\boldsymbol{w}}^t$ for any t.

F.2 KEY LEMMAS

For clarity, we will present several lemmas for establishing our main theorem. In particular, we will present a descent lemma for $\mathbb{E}[\|\overline{\boldsymbol{w}}^t - \boldsymbol{\theta}^{PS}\|_2^2]$ and an upper bound for $\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$, which together gives a standard descent lemma for $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$ in SGD analysis and leads to $\mathcal{O}(\frac{1}{t})$ convergence.

In the following lemma, we aim to establish an upper bound for $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$. Because $\overline{\boldsymbol{\theta}}^{t+1} = \overline{\boldsymbol{w}}^{t+1}$ in full participation, this is equivalent to establishing an upper bound for $\mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$.

Lemma F.1. *(Descent Lemma) Under Assumptions 2.1, 2.2, 2.3, 2.5, in full participation*

$$\begin{aligned}\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] = \mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ +2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\end{aligned}$$

for any t, where $\epsilon_{max} := \max_i \epsilon_i, \bar{\epsilon} := \sum_{i=1}^N p_i\epsilon_i, c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\bar{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], \tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L$.

Proof. This proof follows from Lemma 3 in Li et al. (2022). We first decompose $\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2$ as

$$\mathbb{E}\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2 = \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \eta_t\boldsymbol{g}_t - \boldsymbol{\theta}^{PS}\|_2^2 = \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - 2\eta_t\mathbb{E}\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\, \boldsymbol{g}_t\rangle + \eta_t^2\mathbb{E}\|\boldsymbol{g}_t\|_2^2. \tag{6}$$

Next we present an upper bound for $\mathbb{E}\|\boldsymbol{g}_t\|_2^2$. By the definition of $\boldsymbol{\theta}^{PS}$, we have $\sum_{i=1}^N p_i \nabla f_i(\boldsymbol{\theta}^{PS}; \boldsymbol{\theta}^{PS}) = \mathbf{0}$, and thus

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_t\|_2^2 &= \mathbb{E}\|\sum_{i=1}^N p_i[\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) + \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})]\|_2^2 \\ &\leq 2\mathbb{E}\|\sum_{i=1}^N \nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)\|_2^2 + 2\mathbb{E}\|\sum_{i=1}^N p_i[\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})]\|_2^2 \\ &\leq 2\sum_{i=1}^N p_i\mathbb{E}[\|\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)\|_2^2] + 2\sum_{i=1}^N p_i\mathbb{E}[\|\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})\|_2^2] \\ &\leq 2\sum_{i=1}^N p_i\sigma^2\left(1 + \mathbb{E}\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\right) + 2\sum_{i=1}^N p_i L^2(1+\epsilon_i)^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\end{aligned}$$

where the second inequality is due to the convexity of 2-norm and the last inequality is due to Assumption 2.5 and Lemma 2.4. Since $\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq 2\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 2\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2$ and $\epsilon_i \leq \epsilon_{max}$, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{g}_t\|_2^2] &\leq 2\sigma^2 + 4[\sigma^2 + L^2(1+\epsilon_{max})^2]\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 4[\sigma^2 + L^2(1+\epsilon_{max})^2]\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ &= 2\sigma^2 + c_2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + c_2\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2. \end{aligned} \tag{7}$$

Next, we focus on establishing a lower bound for $\mathbb{E}[\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle]$. By the law of total expectation and $\sum_{i=1}^N p_i \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}) = \mathbf{0}$, we have

$$\begin{aligned}
\mathbb{E}\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle &= \mathbb{E}\big[\mathbb{E}_t\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle\big] \\
&= \mathbb{E}[\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \overline{\boldsymbol{g}}_t\rangle] \\
&= \mathbb{E}\Big[\underbrace{\sum_{i=1}^N p_i \big\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t;\boldsymbol{\theta}^{PS})\big\rangle}_{A} \\
&\qquad + \underbrace{\sum_{i=1}^N p_i \big\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \nabla f_i(\overline{\boldsymbol{\theta}}^t;\boldsymbol{\theta}^{PS}) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})\big\rangle}_{B}\Big].
\end{aligned}$$

On the one hand, applying Cauchy-Schwarz inequality and Lemma 2.4, we have

$$\begin{aligned}
A &\geq -\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \sum_{i=1}^N p_i \big(L\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2 + L\epsilon_i\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2\big) \\
&\geq -\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \sum_{i=1}^N p_i \big(L(1+\epsilon_i)\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2 + L\epsilon_i\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2\big) \\
&\geq -L\overline{\epsilon}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - L(1+\epsilon_{max}) \sum_{i=1}^N p_i \|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2.
\end{aligned}$$

On the other hand, with the strong convexity in Assumption 2.1, we have $B \geq \mu\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2$. Therefore, for any $\alpha > 0$, using the lower bounds on A, B, and the Young's inequality shows that

$$\begin{aligned}
&\mathbb{E}[\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle] \\
&\geq (\mu - L\overline{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - L(1+\epsilon_{max}) \sum_{i=1}^N p_i \mathbb{E}\big[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2\big] \\
&\geq \big(\mu - L\overline{\epsilon} - \frac{\alpha}{2}L(1+\epsilon_{max})\big)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - \frac{L(1+\epsilon_{max})}{2\alpha} \sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
&\geq (\mu - (1+\delta)L\overline{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - \frac{L(1+\epsilon_{max})^2}{4\delta\overline{\epsilon}} \sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \qquad (8)
\end{aligned}$$

where we have set $\alpha := \frac{2\delta\overline{\epsilon}}{1+\epsilon_{max}}$ in the last line.

Recall that we denote

$$c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\overline{\epsilon}},\ \ c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2],\ \ \tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L.$$

Combining equation 6, equation 7, equation 8, we have

$$\begin{aligned}
&\mathbb{E}\big[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2\big] \\
\le\, &\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\
&- 2\eta_t\left[(\mu - (1+\delta)L\bar{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - \frac{L(1+\epsilon_{max})}{4\delta\bar{\epsilon}}\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\right] \\
&+ \eta_t^2\left[2\sigma^2 + c_2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + c_2\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\right] \\
=\, &(1 - 2\tilde{\mu}\eta_t + c_2\eta_t^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 2\sigma^2\eta_t^2 \\
\le\, &(1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 2\sigma^2\eta_t^2
\end{aligned}$$

where the last inequality is obtained by observing the condition $\eta_t \le \tilde{\mu}/c_2$. □

Now we are going to establish an upper bound for $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$. Note that if $t \in \mathcal{I}_E$, the synchronization step, we have $\boldsymbol{\theta}_i^t = \overline{\boldsymbol{\theta}}^t$ for any $i \in [N]$, which implies that $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 = 0$. If $t \notin \mathcal{I}_E$, the following lemma gives an upper bound for $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$.

Lemma F.2. *(Consensus Error) Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, if $\{\eta_t\}$ is non-increasing, $\eta_t \le 2\eta_{t+E}$, $t \notin \mathcal{I}_E$, $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, and*

$$\eta_0 \le \hat{\eta}_0 := \frac{-2\sigma^2 + \sqrt{4\sigma^4 + 4a_1 \cdot \tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}}{2a_1}$$

where

$$a_1 := (12c_1 + 2c_2(c_3)^{-1})(2E^2 - E)\log E\bigg((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2(8\sigma^2 + 12\varsigma^2)\bigg),$$

then in full participation, we have

$$\begin{aligned}
\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \le\, &4\eta_t^2(2E^2 - E)\log E(48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&+4\eta_t^2(2E^2 - E)\log E(24\sigma^2 + 36\varsigma^2).
\end{aligned}$$

where for any t, where $\epsilon_{max} := \max_i \epsilon_i, \bar{\epsilon} := \sum_{i=1}^N p_i\epsilon_i, c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\bar{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], \tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L$, $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$.

(One should note that $4\eta_t^2$, $(48\sigma^2 + 36\varsigma^2)$, and $(24\sigma^2 + 36\varsigma^2)$ comes from several times of applying $\eta_{t-1} \le 2\eta_t$ and the real constants could be much smaller by choosing stepsizes carefully.)

Proof. In this proof, for convenience, we will discuss with respect to $t+1$ where we assume $t+1 \notin \mathcal{I}_E$ and transfer back to t in the last. First by the update rule, we have

$$\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1} = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t).$$

Using Young's inequality, we have

$$\begin{aligned}
\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 &= \sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t)\|_2^2 \\
&\le (1+\alpha_t)\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \eta_t^2(1+\alpha_t^{-1})\underbrace{\sum_{i=1}^N p_i\mathbb{E}\|\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t\|_2^2}_{B}
\end{aligned} \tag{9}$$

where $\alpha_t > 0$ is a free chosen parameter. Next, we are going to establish an upper bound for B. Notice that

$$
\begin{aligned}
B =& \sum_{i=1}^{N} p_i \mathbb{E}\|\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t\|_2^2 \\
=& \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \sum_{j=1}^{N} p_j \nabla l(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
=& \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) + \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) \\
&+ \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \sum_{j=1}^{N} p_j \nabla l(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
\leq& 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&+ 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \sum_{j=1}^{N} p_j \nabla l(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
\leq& 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&+ 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \sum_{j=1}^{N} p_j \|\nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \nabla l(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
=& 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&+ 3\mathbb{E}\Big[\sum_{j=1}^{N} p_j \|\nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \nabla l(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
\leq& 6\sigma^2\left(1 + \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\Big]\right) \\
&+ 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big]
\end{aligned}
$$

where the last inequality is by Assumption 2.5. On the other hand, we have

$$
\begin{aligned}
& 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \big\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t)\big\|_2^2\Big] \\
=& 3\sum_{i=1}^{N} p_i \mathbb{E}\big\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) + \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \big(\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big)\big\|_2^2 \\
\leq& \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2}_{B_1} + \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2}_{B_2} \\
&+ \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\sum_{j=1}^{N} p_j \big(\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big)\|_2^2}_{B_3}.
\end{aligned}
$$

Using Lemma 2.4, we have

$$B_1 \leq 9\sum_{i=1}^{N} p_i L^2 (1+\epsilon_i)^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \leq 9\sum_{i=1}^{N} p_i L^2 (1+\epsilon_{max})^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

Using Assumption 2.6, we have

$$\begin{aligned} B_2 &= 9\sum_{i=1}^{N} p_i \mathbb{E}\big\|\nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \nabla f(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big\|_2^2 \\ &\leq 9\sum_{i=1}^{N} p_i \varsigma^2 (1 + \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2) \\ &= 9\varsigma^2 + 9\varsigma^2 \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2. \end{aligned}$$

Using Lemma 2.4, we have

$$B_3 \leq 9\sum_{i=1}^{N} p_i \sum_{j=1}^{N} p_j \mathbb{E}\|\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2 \leq 9\sum_{i=1}^{N} p_i L^2 (1+\epsilon_{max})^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

Therefore,

$$B_1 + B_2 + B_3 \leq 18\sum_{i=1}^{N} p_i L^2 (1+\epsilon_{max})^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 9\varsigma^2 + 9\varsigma^2 \sum_{i=1}^{N} p_i \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2,$$

which results in that

$$\begin{aligned} B &\leq 6\sigma^2\bigg(1 + \mathbb{E}\bigg[\sum_{i=1}^{N} p_i \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\bigg]\bigg) + 18\sum_{i=1}^{N} p_i L^2 (1+\epsilon_{max})^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \\ &\quad 9\varsigma^2 + 9\varsigma^2 \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ &\leq 6\sigma^2\bigg(1 + 2\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 18\sum_{i=1}^{N} p_i L^2 (1+\epsilon_{max})^2 \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \\ &\quad 9\varsigma^2 + 9\varsigma^2 \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2\bigg) \\ &= 6\sigma^2 + 9\varsigma^2 + \big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + (12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2. \end{aligned}$$

Inserting this formula into equation 9, we obtain

$$\begin{aligned} &\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\ &\leq (1+\alpha_t)\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ &\quad + \eta_t^2(1+\alpha_t^{-1})\bigg(6\sigma^2 + 9\varsigma^2 + \big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + (12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2\bigg) \\ &= \big(1 + \alpha_t + \eta_t^2(1+\alpha_t^{-1})\big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\big)\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ &\quad + \eta_t^2(1+\alpha_t^{-1})(12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + \eta_t^2(1+\alpha_t^{-1})(6\sigma^2 + 9\varsigma^2) \end{aligned}$$

where $\alpha_t > 0$ is a free chosen parameter. Let $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$ and $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$. Then we choose $\alpha_t = \frac{1}{2(t+1-t_0)}$, if we have

$$\begin{aligned} &\eta_t^2(1+\alpha_t^{-1})\big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big) = \eta_t^2(1+\alpha_t^{-1})c_3 \leq \frac{1}{2(t+1-t_0)} \\ \iff &\eta_t^2 \leq \frac{1}{2c_3(t+1-t_0)\big(1+2(t+1-t_0)\big)}, \end{aligned} \tag{10}$$

then note that $1+\alpha_t^{-1} = 1 + 2(t+1-t_0) \le 2E-1$,

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\
&\le \frac{t+2-t_0}{t+1-t_0} \sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t} - \overline{\boldsymbol{\theta}}^{t}\|_2^2 + \eta_t^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + \eta_t^2(2E-1)(6\sigma^2+9\varsigma^2).
\end{aligned}$$

Continuing the above expansion until t_0 and leveraging $\eta_s \le \eta_{t_0} \le 2\eta_{t_0+E} \le 2\eta_t$ gives us

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\
&\le \frac{t+2-t_0}{t_0+1-t_0} \sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t_0} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2 + \sum_{s=t_0}^{t} \frac{t+2-t_0}{s+2-t_0}\eta_s^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{s} - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \sum_{s=t_0}^{t} \frac{t+2-t_0}{s+2-t_0}\eta_s^2(2E-1)(6\sigma^2+9\varsigma^2) \\
&= \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_s^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{s} - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_s^2(2E-1)(6\sigma^2+9\varsigma^2) \\
&\le \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E-1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{s} - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E-1)(24\sigma^2+36\varsigma^2).
\end{aligned} \tag{11}$$

With the above formula and Lemma F.1, we now prove that if η_0 is sufficiently small, the for any t, we have $\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2$. We first derive the following inequality, which we will use later. Note that for any t where $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$, we have

$$\sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} = (t+2-t_0)(\frac{1}{2} + \ldots + \frac{1}{t-t_0+2}) \le (t+2-t_0)\log(t+2-t_0) \le E\log E. \tag{12}$$

We prove $\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2$ by induction.

First, this inequality clearly holds for $t=0$. Suppose it holds for $0 \le s \le t$ where $t \le E-1$. Then by Lemma F.1 and equation 11, we have

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&= \mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\le (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + (c_1\eta_t + c_2\eta_t^2)\Bigg(\sum_{s=0}^{t-t_0-1} \frac{t+1-t_0}{s+2}\eta_{t-1}^2(2E-1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \sum_{s=0}^{t-t_0-1} \frac{t+1-t_0}{s+2}\eta_{t-1}^2(2E-1)(24\sigma^2+36\varsigma^2)\Bigg) \\
&= (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + (c_1\eta_t + c_2\eta_t^2)\eta_{t-1}^2(2E^2-E)\log E\Bigg((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Bigg).
\end{aligned}$$

By equation 10, we need $\eta_0 \le \frac{1}{6c_3}$, together with $\eta_{t-1} \le 2\eta_t$ implies

$$(c_1\eta_t + c_2\eta_t^2)\eta_{t-1}^2 \le \eta_t^3(4c_1 + 2c_2(3c_3)^{-1}).$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\ \leq& (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\ &+ \eta_t^3(12c_1 + 2c_2(c_3)^{-1})(2E^2 - E)\log E\bigg((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2 + 12\varsigma^2)\bigg)\end{aligned}$$

whose right-hand side is no larger than $\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$ if

$$\begin{aligned}0 \geq& \eta_t^2 \cdot (12c_1 + 2c_2(c_3)^{-1})(2E^2 - E)\log E\bigg((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2(8\sigma^2 + 12\varsigma^2)\bigg) \\ &+ \eta_t \cdot 2\sigma^2 - \tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\end{aligned}$$

which is satisfied if

$$\eta_t \leq \eta_0 \leq \frac{-2\sigma^2 + \sqrt{4\sigma^4 + 4a_1 \cdot \tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}}{2a_1} = \hat{\eta}_0. \tag{13}$$

where

$$a_1 := (12c_1 + 2c_2(c_3)^{-1})(2E^2 - E)\log E\bigg((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2(8\sigma^2 + 12\varsigma^2)\bigg).$$

Thus we have proved that for any $0 \leq t \leq E$, if equation 13 holds, then $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$. The same proof technique can be extended to any $nE \leq t \leq (n+1)E$ where $n \in \mathbb{N}_+$ and thus for any t, if equation 13 holds, then $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$.

Therefore, under equation 10, equation 13 and $\eta_{t-1} \leq 2\eta_t$, by equation 9 and equation 11, if $t \notin \mathcal{I}_E$, we have

$$\begin{aligned}\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \leq& \eta_{t-1}^2(2E^2 - E)\log E(48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\ &+ \eta_{t-1}^2(2E^2 - E)\log E(24\sigma^2 + 36\varsigma^2) \\ \leq& 4\eta_t^2(2E^2 - E)\log E(48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\ &+ 4\eta_t^2(2E^2 - E)\log E(24\sigma^2 + 36\varsigma^2).\end{aligned}$$

□

The following lemma gives us a standard descent lemma in SGD analysis under technical conditions for establishing the $\mathcal{O}(\frac{1}{T})$ convergence in Theorem 3.1.

Lemma F.3. *Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, if $\{\eta_t\}$ is non-increasing, $\eta_t \leq 2\eta_{t+E}$, $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, and*

$$\eta_0 \leq \hat{\eta}_0 := \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2 - E)\log E\big((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2 + 12\varsigma^2)\big)},$$

then in full participation, we have

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B\eta_t^2$$

where for any t, where $\epsilon_{max} := \max_i \epsilon_i, \overline{\epsilon} := \sum_{i=1}^N p_i\epsilon_i, c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\overline{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], \tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L$, $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$, and $B := 2\sigma^2 + (4c_1\hat{\eta}_0 + 4c_2\hat{\eta}_0^2)(2E^2 - E)\log E\big((48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)\big)$.

Proof. We discuss in two cases, $t \in \mathcal{I}_E$ and $t \notin \mathcal{I}_E$. If $t \in \mathcal{I}_E$, then we have $\boldsymbol{\theta}_i^t = \overline{\boldsymbol{\theta}}^t$, and by Lemma F.1,

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
=\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B\eta_t^2.
\end{aligned}$$

If $t \notin \mathcal{I}_E$, combining Lemma F.1 and Lemma F.2, we have

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&+ (4c_1\eta_t + 4c_2\eta_t^2)\Bigg(\eta_t^2(2E^2-E)\log E(48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&+ \eta_t^2(2E^2-E)\log E(24\sigma^2 + 36\varsigma^2)\Bigg) \\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2B\eta_t^2.
\end{aligned}$$

□

F.3 Completing the Proof of Theorem 3.1

We restate the definitions of all the constants here:

Constants independent of system design.
$\epsilon_{max} := \max_i \epsilon_i$,

$\overline{\epsilon} := \sum_{i=1}^N p_i \epsilon_i$,

$\tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L$,

$c_1 := \big(L(1+\epsilon_{max})^2\big)/(2\delta\overline{\epsilon})$,

$c_2 := 4\big[\sigma^2 + L^2(1+\epsilon_{max})^2\big]$,

$c_3 := 6\big[2\sigma^2 + 3L^2(1+\epsilon_{\max})^2\big]$,

$c_4 := 16\sigma^2 + 12\varsigma^2 + (8\sigma^2 + 12\varsigma^2)/\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$,

$c_5 := (48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)$.

Constants related to system design (e.g., E, K).
$\hat{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4(2E^2 - E)\log E\big)$,

$B := 2\sigma^2 + (4c_1\hat{\eta}_0 + 4c_2\hat{\eta}_0^2)c_5(2E^2-E)\log E$,

$c_6 := (2E^2 + 3E + 1)\log(E+1)$,

$\tilde{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4c_6\big)$,

$B_1 := 2\sigma^2 + (4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + 1/K)c_5c_6$,

$B_2 := 2\sigma^2 + \big(4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + \frac{N-K}{KN(N-1)}\big)c_5c_6$.

Instead of proving Theorem 3.1 directly, we prove a more general version of convergence results suppose that some conditions about the stepsize are satisfied. Then we will show that the stepsizes given in Theorem 3.1 satisfy the conditions.

Theorem F.4. *Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, for a diminishing stepsize* $\eta_t = \frac{\beta}{t+\gamma}$ *where* $\beta > \frac{1}{\tilde{\mu}}$, $\gamma > 0$ *such that* $\eta_0 \le \hat{\eta}_0$, $\eta_t \le 2\eta_{t+E}$, *and* $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, *then in full participation, we have for any* t

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma + t}$$

where $\upsilon = \max\left\{\frac{4B}{\tilde{\mu}^2}, \gamma\mathbb{E}[\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\right\}$.

Proof. Let $\Delta_t := \mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2]$, then from Lemma F.3, we have

$$\Delta_{t+1} \le (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B\eta_t^2.$$

For a diminishing stepsize $\eta_t = \frac{\beta}{t+\gamma}$ where $\beta > \frac{1}{\tilde{\mu}}$, $\gamma > 0$ such that $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, $\eta_0 \le \hat{\eta}_0$, and $\eta_t \le 2\eta_{t+E}$, we will prove that $\Delta_t \le \frac{\upsilon}{\gamma+t}$ where $\upsilon = \max\left\{\frac{\beta^2 B}{\beta\tilde{\mu}-1}, \gamma\Delta_0\right\} = \max\left\{\frac{4B}{\tilde{\mu}^2}, \gamma\Delta_0\right\}$ by induction.

Firstly, $\Delta_0 \le \frac{\upsilon}{\gamma}$ by the definition of υ. Assume that for some $0 \le t$, $\Delta_t \le \frac{\upsilon}{\gamma+t}$, then

$$\begin{aligned}\Delta_{t+1} &\le (1-\eta_t\tilde{\mu})\,\Delta_t + \eta_t^2 B\\ &\le \left(1 - \frac{\beta\tilde{\mu}}{t+\gamma}\right)\frac{\upsilon}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2}\\ &= \frac{t+\gamma-1}{(t+\gamma)^2}\upsilon + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\tilde{\mu}-1}{(t+\gamma)^2}\upsilon\right]\\ &\le \frac{\upsilon}{t+\gamma+1}.\end{aligned}$$

Specifically, if we choose $\beta = \frac{2}{\tilde{\mu}}$, $\gamma = \max\{\frac{2}{\tilde{\mu}\hat{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2)}\}$, then we have

$$\eta_0 = \frac{\beta}{\gamma} \le \frac{2}{\tilde{\mu}\frac{2}{\tilde{\mu}\hat{\eta}_0}} = \hat{\eta}_0$$

and

$$\eta_t - 2\eta_{t+E} = \frac{\beta}{\gamma+t} - \frac{2\beta}{\gamma+t+E} = \frac{\beta(E-\gamma-t)}{(\gamma+t)(\gamma+t+E)} \le \frac{\beta(E-\gamma)}{(\gamma+t)(\gamma+t+E)} \le 0.$$

To prove that $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$ for any t, it suffices to prove that for $0 \le t \le E-1$ because $\{\eta_t\}$, i.e., $t_0 = 0$, is non-increasing and $t+1-t_0$ is periodic with period E. When $t_0 = 0$, we need to prove $\eta_t^2 \le 1/\big(2c_3(t+1)(1+2(t+1))\big)$ for $0 \le t \le E-1$, which is satisfied if

$$\begin{aligned}&\max_{0\le t\le E-1}\eta_t \le \min_{0\le t\le E-1}\sqrt{1/\big(2c_3(t+1)(1+2(t+1))\big)}\\ \Longleftrightarrow\ &\eta_0 \le \sqrt{\frac{1}{2E(2E+1)c_3}}\\ \Longleftrightarrow\ &\gamma \ge \beta\sqrt{2E(2E+1)c_3} = \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)c_3} = \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}.\end{aligned}$$

□

G Proof of Theorem 3.2 and Theorem 3.3

G.1 Additional Notations

Similar to Appendix F, in our analysis, for the sake of convenience, we will define two additional sequences as $\overline{\boldsymbol{w}}^t := \sum_{i=1}^N p_i \boldsymbol{w}_i^t$ and $\overline{\boldsymbol{\theta}}^t := \sum_{i=1}^N p_i \boldsymbol{\theta}_i^t$, following that of Li et al. (2020b). We note that $\overline{\boldsymbol{w}}^t$ results from a single step of SGD from $\overline{\boldsymbol{\theta}}^t$. When $t+1 \notin \mathcal{I}_E$, both $\overline{\boldsymbol{w}}^t$ and $\overline{\boldsymbol{\theta}}^t$ are unaccessible. When $t+1 \in \mathcal{I}_E$, we can obtain $\overline{\boldsymbol{\theta}}^t$. In addition, we also define $\overline{\boldsymbol{g}}_t := \sum_{i=1}^N p_i \nabla f_i(\boldsymbol{\theta}_i^t; \boldsymbol{\theta}_i^t)$, $\boldsymbol{g}_t := \sum_{i=1}^N p_i \nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1})$ where $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$. It is clear that in full participation, $\overline{\boldsymbol{w}}^{t+1} = \overline{\boldsymbol{w}}^t - \eta_t \boldsymbol{g}_t$ and $\mathbb{E}\boldsymbol{g}_t = \overline{\boldsymbol{g}}_t$. Notice now we do not have $\overline{\boldsymbol{\theta}}^t = \overline{\boldsymbol{w}}^t$ for any t. But we will show later that they are equal with expectation to the choice of $\mathcal{S}_t$.

In particular, in our analysis, there would be two types of randomness, one from the stochastic gradients and one from the random sampling of the devices. All analysis in Appendix F only involves the former. To make a distinguishment, we use $\mathbb{E}_{\mathcal{S}_t}$ to denote the latter.

G.2 Key Lemmas

We first show that the sampling schemes I & II are unbiased.

Lemma G.1. *Li et al. (2020b) (Unbiased sampling scheme). If $t+1 \in \mathcal{I}_E$, for Scheme I and Scheme II, we have*

$$\mathbb{E}_{\mathcal{S}_t}\left[\overline{\boldsymbol{\theta}}^{t+1}\right] = \overline{\boldsymbol{w}}^{t+1}.$$

Proof. Let $\{x_i\}_{i=1}^N$ denote any fixed deterministic sequence. We sample a multiset $\mathcal{S}_t$ with $|\mathcal{S}_t| = K$ by the procedure where each sampling time, we sample x_k with probability q_k for each time. Note that two samples are not necessarily independent. We only require each sampling distribution is identical. Let $\mathcal{S}_t = \{i_1, \ldots, i_K\} \subset [N]$ (some i_k's may have the same value if sampling with replacement). Then

$$\mathbb{E}_{\mathcal{S}_t} \sum_{k \in \mathcal{S}_t} x_k = \mathbb{E}_{\mathcal{S}_t} \sum_{k=1}^K x_{i_k} = K \mathbb{E}_{\mathcal{S}_t} x_{i_1} = K \sum_{k=1}^K q_k x_k.$$

For Scheme I, $q_k = p_k$ and for Scheme II, $q_k = \frac{1}{N}$, replacing the values into the above proves the lemma. □

Similar to Lemma F.1, we are going to establish an upper bound for $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$. When $t+1 \notin \mathcal{I}_E$, we have $\overline{\boldsymbol{\theta}}^{t+1} = \overline{\boldsymbol{w}}^{t+1}$ for both schemes, and therefore this is equivalent to establishing an upper bound for $\mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$. However, when $t+1 \in \mathcal{I}_E$, we only have $\mathbb{E}_{\mathcal{S}_t}\left[\overline{\boldsymbol{\theta}}^{t+1}\right] = \overline{\boldsymbol{w}}^{t+1}$ and we need other upper-bounding strategies.

Lemma G.2. *Under Assumptions 2.1, 2.2, 2.3, 2.5, for scheme I & II:*

1. *if $t+1 \notin \mathcal{I}_E$,*

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] = \mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

2. *if $t+1 \in \mathcal{I}_E$: for scheme I,*

$$\begin{aligned} &\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\ &\leq \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ &\quad + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2, \end{aligned}$$

while for scheme II,

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\ &\le \frac{1}{K(N-1)}\left(1-\frac{K}{N}\right)\sum_{k=1}^{N}p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2\\ &\quad+2\sigma^2\eta_t^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2.\end{aligned}$$

where $\epsilon_{max}:=\max_i\epsilon_i, \bar{\epsilon}:=\sum_{i=1}^N p_i\epsilon_i, c_1:=\frac{L(1+\epsilon_{max})^2}{2\delta\bar{\epsilon}}, c_2:=4[\sigma^2+L^2(1+\epsilon_{max})^2], \tilde{\mu}:=\mu-(1+\delta)\bar{\epsilon}L.$

Proof. When $t+1\notin\mathcal{I}_E$, because $\overline{\boldsymbol{\theta}}^{t+1}=\overline{\boldsymbol{w}}^{t+1}$ for both schemes, by Lemma F.1, we got the conclusion. When $t+1\in\mathcal{I}_E$, we have

$$\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2=\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+\mathbb{E}\|\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2+2\mathbb{E}\langle\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1},\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\rangle.$$

By Lemma G.1 and the law of total expectation, we have

$$\mathbb{E}\langle\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1},\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\rangle=\mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\langle\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1},\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\rangle\big]=0.$$

Next we focus on upper bounding $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2]$ under two sampling schemes.

Denote $\mathcal{S}_{t+1}=\{i_1,\ldots,i_K\}$, then for scheme I, $\overline{\boldsymbol{\theta}}^{t+1}=\frac{1}{K}\sum_{l=1}^K\boldsymbol{w}_{i_l}^{t+1}$. Thus by the law of total expectation, we have

$$\begin{aligned}\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2&=\mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\big]\\ &=\mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\|\frac{1}{K}\sum_{l=1}^K\boldsymbol{w}_{i_l}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\big]\\ &\le\mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\frac{1}{K^2}\sum_{l=1}^K\|\boldsymbol{w}_{i_l}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\big]\\ &=\frac{1}{K}\sum_{k=1}^N p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2.\end{aligned}$$

Again with $\overline{\boldsymbol{\theta}}^{t+1}=\frac{1}{K}\sum_{l=1}^K\boldsymbol{w}_{i_l}^{t+1}$, for scheme II, by the law of total expectation, we have

$$\begin{aligned}&\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\\ &=\mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\big]\\ &=\mathbb{E}\left[\mathbb{E}_{\mathcal{S}_{t+1}}\left[\|\frac{1}{K}\sum_{l=1}^K\boldsymbol{w}_{i_l}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\right]\right]\\ &=\mathbb{E}\left[\frac{1}{K^2}\mathbb{E}_{\mathcal{S}_{t+1}}\left[\|\sum_{i=1}^N\mathbf{1}\{i\in\mathcal{S}_{t+1}\}(\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1})\|_2^2\right]\right]\\ &\le\frac{1}{K^2}\mathbb{E}\left[\sum_{i=1}^N\mathbb{P}(i\in\mathcal{S}_{t+1})\|\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+\sum_{j\neq i}\mathbb{P}(i,j\in\mathcal{S}_{t+1})\left\langle\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1}-\overline{\boldsymbol{w}}^{t+1}\right\rangle\right]\\ &=\frac{1}{KN}\sum_{i=1}^N\mathbb{E}\|\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+\frac{K-1}{KN(N-1)}\sum_{i\neq j}\mathbb{E}\left\langle\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1}-\overline{\boldsymbol{w}}^{t+1}\right\rangle\\ &=\frac{1}{K(N-1)}\left(1-\frac{K}{N}\right)\sum_{i=1}^N\mathbb{E}\|\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\end{aligned}$$

where we use the following equalities: (1) $\mathbb{P}\left(i \in S_{t+1}\right) = \frac{K}{N}$ and $\mathbb{P}\left(i, j \in S_{t+1}\right) = \frac{K(K-1)}{N(N-1)}$ for all $i \neq j$ and (2) $\sum_{i=1}^{N} \left\|\boldsymbol{w}_i^t - \overline{\boldsymbol{w}}^t\right\|^2 + \sum_{i \neq j} \langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}, \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1} \rangle = 0$.

The conclusion follows from the above discussion.

□

To really give a descent lemma as in SGD analysis, we have to bound $\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$ for $t \notin \mathcal{I}_E$ and $\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{w}_i^t - \overline{\boldsymbol{w}}^t\|_2^2$ for $t \in \mathcal{I}_E$, given by the following lemma.

Lemma G.3. *Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, if $\{\eta_t\}$ is non-increasing, $\eta_t \leq 2\eta_{t+E}$, $t \notin \mathcal{I}_E$, $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, and*

$$\eta_0 \leq \tilde{\eta}_0 := \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2 + 12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2 + 12\varsigma^2)\big)},$$

then

1. *for scheme I,*

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + (4c_1\eta_t + 4c_2\eta_t^2 + K^{-1})(2E^2+3E+1)\log(E+1)\eta_t^2\Bigg(\\
&\quad (48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)\Bigg),
\end{aligned}$$

2. *for scheme II,*

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + \left(4c_1\eta_t + 4c_2\eta_t^2 + \frac{N-K}{KN(N-1)}\right)(2E^2+3E+1)\log(E+1) \\
&\quad \eta_t^2\Bigg((48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)\Bigg).
\end{aligned}$$

where for any t, where $\epsilon_{max} := \max_i \epsilon_i, \bar{\epsilon} := \sum_{i=1}^{N} p_i\epsilon_i, c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\bar{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], \tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L$, $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$.

(One should note that $4c_1\eta_t + 4c_2\eta_t^2$, $(48\sigma^2 + 36\varsigma^2)$, and $(24\sigma^2 + 36\varsigma^2)$ comes from several times of applying $\eta_{t-1} \leq 2\eta_t$ and the real constants could be much smaller by choosing stepsizes carefully.)

Proof. In this proof, for convenience, we will discuss with respect to $t+1$ where we assume $t+1 \notin \mathcal{I}_E$ and transfer back to t in the last. First by the update rule, we have when $t+1 \notin \mathcal{I}_E$

$$\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1} = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t)$$

and when $t+1 \in \mathcal{I}_E$,

$$\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1} = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla l(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t).$$

Then with the same method in Lemma F.2, let $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$ and $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$, if $\eta_t^2 \leq \frac{1}{2c_3(t+1-t_0)\big(1+2(t+1-t_0)\big)}$, we will have: if $t+1 \notin \mathcal{I}_E$,

$$\begin{aligned}
&\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\
&\leq \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E+1)(48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E+1)(24\sigma^2 + 36\varsigma^2)
\end{aligned}$$

and if $t+1 \in \mathcal{I}_E$,

$$
\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \\
\le& \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0}\frac{t+2-t_0}{s+2}\eta_t^2(2E+1)(24\sigma^2+36\varsigma^2).
\end{aligned}
$$

With the above formula and Lemma G.2, we now prove that if η_0 is sufficiently small, then for any t, we have $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$. We first derive the following inequality, which we will use later. Note that for any t where $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$, we have

$$
\sum_{s=0}^{t-t_0}\frac{t+2-t_0}{s+2} = (t+2-t_0)(\frac{1}{2}+\ldots+\frac{1}{t-t_0+2}) \le (t+2-t_0)\log(t+2-t_0) \le (E+1)\log(E+1).
$$

Then again by the same induction method in Lemma F.2, we have if

$$
\begin{aligned}
&\eta_t \le \eta_0 \\
&\le \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2+12\varsigma^2)\big)} = \tilde{\eta}_0,
\end{aligned}
$$

then for any t, we have $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2$.

Under all these conditions, if $t \notin \mathcal{I}_E$, we have

$$
\begin{aligned}
\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 &\le \eta_{t-1}^2(2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \eta_{t-1}^2(2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2) \\
&\le 4\eta_t^2(2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + 4\eta_t^2(2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2),
\end{aligned}
$$

and if $t+1 \in \mathcal{I}_E$, we have

$$
\begin{aligned}
\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 &\le \eta_{t-1}^2(2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \eta_{t-1}^2(2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2).
\end{aligned}
$$

Note that in Lemma G.2, the inequality for $t \notin \mathcal{I}_E$ is looser than the inequality for $t \in \mathcal{I}_E$. Therefore, we can apply the inequality for $t \in \mathcal{I}_E$ for all t. Combining this inequality with the above formula gives us that:

1. for scheme I,

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
\le& \frac{1}{K}\sum_{k=1}^{N} p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\
&+ 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
\le& (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&+ (4c_1\eta_t + 4c_2\eta_t^2 + K^{-1})(2E^2+3E+1)\log(E+1)\eta_t^2\Bigg(\\
&(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Bigg),
\end{aligned}
$$

2. for scheme II,

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq \frac{N-K}{KN(N-1)} \sum_{k=1}^{N} p_k \mathbb{E}\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t}\|_2^2 \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + \left(4c_1\eta_t + 4c_2\eta_t^2 + \frac{N-K}{KN(N-1)}\right)(2E^2+3E+1)\log(E+1)\eta_t^2\Bigg(\\
&\quad (48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)\Bigg).
\end{aligned}
$$

□

Lemma G.4. *Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, if $\{\eta_t\}$ is non-increasing, $\eta_t \leq 2\eta_{t+E}$, $t \notin \mathcal{I}_E$, $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, and*

$$
\eta_0 \leq \tilde{\eta}_0 := \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2+12\varsigma^2)\big)},
$$

then

1. *for scheme I,*

$$
\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + B_1\eta_t^2,
$$

with

$$
\begin{aligned}
B_1 := 2\sigma^2 + (4c_1\eta_t + 4c_2\eta_t^2 + K^{-1})(2E^2+3E+1)\log(E+1)\Bigg(\\
(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Bigg),
\end{aligned}
$$

2. *for scheme II,*

$$
\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + B_2\eta_t^2
$$

with

$$
\begin{aligned}
B_2 := 2\sigma^2 + \left(4c_1\eta_t + 4c_2\eta_t^2 + \frac{N-K}{KN(N-1)}\right)(2E^2+3E+1)\log(E+1)\Bigg(\\
(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Bigg),
\end{aligned}
$$

where for any t, where $\epsilon_{max} := \max_i \epsilon_i, \overline{\epsilon} := \sum_{i=1}^{N} p_i\epsilon_i, c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\overline{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], \tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L$, $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$.

Proof. The conclusion follows directly from Lemma G.3. □

G.3 Completing the proof of Theorem 3.2 and 3.3

We restate the definitions of all the constants here:

Constants independent of system design.
$\epsilon_{max} := \max_i \epsilon_i$,

$\overline{\epsilon} := \sum_{i=1}^{N} p_i\epsilon_i$,

$\tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L,$

$c_1 := \big(L(1+\epsilon_{max})^2\big)/(2\delta\bar{\epsilon}),$

$c_2 := 4\big[\sigma^2 + L^2(1+\epsilon_{max})^2\big],$

$c_3 := 6\big[2\sigma^2 + 3L^2(1+\epsilon_{\max})^2\big],$

$c_4 := 16\sigma^2 + 12\varsigma^2 + (8\sigma^2 + 12\varsigma^2)/\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2,$

$c_5 := (48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2).$

Constants related to system design (e.g., E, K).
$\hat{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4(2E^2 - E)\log E\big),$

$B := 2\sigma^2 + (4c_1\hat{\eta}_0 + 4c_2\hat{\eta}_0^2)c_5(2E^2 - E)\log E\ ,$

$c_6 := (2E^2 + 3E + 1)\log(E+1),$

$\tilde{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4c_6\big),$

$B_1 := 2\sigma^2 + (4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + 1/K)c_5c_6,$

$B_2 := 2\sigma^2 + \big(4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + \frac{N-K}{KN(N-1)}\big)c_5c_6.$

Instead of proving Theorem 3.2 and 3.3 directly, we prove a more general version of convergence results suppose that some conditions about the stepsize are satisfied. Then we will show that the stepsizes given in Theorem 3.2 and 3.3 satisfy the conditions.

Theorem G.5. *Under Assumption 2.1, 2.2, 2.3, 2.5, 2.6, for a diminishing stepsize $\eta_t = \frac{\beta}{t+\gamma}$ where $\beta > \frac{1}{\tilde{\mu}}$, $\gamma > 0$ such that $\eta_0 \le \tilde{\eta}_0$, $\eta_t \le 2\eta_{t+E}$, and $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, then*

1. *for scheme I,*

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma + t},$$

where $\upsilon = \max\Big\{\frac{4B_1}{\tilde{\mu}^2}, \gamma\mathbb{E}[\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\Big\}$;

2. *for scheme II,*

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma + t}$$

where $\upsilon = \max\Big\{\frac{4B_2}{\tilde{\mu}^2}, \gamma\mathbb{E}[\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\Big\}$.

Proof. We give a proof for scheme I and the proof for scheme II follows exactly the same way.

Let $\Delta_t := \mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2]$, then from Lemma G.4, we have

$$\Delta_{t+1} \le (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B_1\eta_t^2.$$

For a diminishing stepsize $\eta_t = \frac{\beta}{t+\gamma}$ where $\beta > \frac{1}{\tilde{\mu}}$, $\gamma > 0$ such that $\eta_t^2 \le 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$, $\eta_0 \le \hat{\eta}_0$, and $\eta_t \le 2\eta_{t+E}$, we will prove that $\Delta_t \le \frac{\upsilon}{\gamma+t}$ where $\upsilon = \max\Big\{\frac{\beta^2 B_1}{\beta\tilde{\mu}-1}, \gamma\Delta_0\Big\} = \max\Big\{\frac{4B_1}{\tilde{\mu}^2}, \gamma\Delta_0\Big\}$ by induction.

Firstly, $\Delta_0 \leq \frac{v}{\gamma}$ by the definition of v. Assume that for some $0 \leq t$, $\Delta_t \leq \frac{v}{\gamma+t}$, then

$$\begin{aligned}
\Delta_{t+1} &\leq (1-\eta_t\tilde{\mu})\,\Delta_t + \eta_t^2 B_1 \\
&\leq \left(1-\frac{\beta\tilde{\mu}}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B_1}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{\beta^2 B_1}{(t+\gamma)^2} - \frac{\beta\tilde{\mu}-1}{(t+\gamma)^2}v\right] \\
&\leq \frac{v}{t+\gamma+1}.
\end{aligned}$$

Specifically, if we choose $\beta = \frac{2}{\tilde{\mu}}, \gamma = \max\{\frac{2}{\tilde{\mu}\tilde{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}\}$, then we have

$$\eta_0 = \frac{\beta}{\gamma} \leq \frac{2}{\tilde{\mu}\frac{2}{\tilde{\mu}\tilde{\eta}_0}} = \hat{\eta}_0$$

and

$$\eta_t - 2\eta_{t+E} = \frac{\beta}{\gamma+t} - \frac{2\beta}{\gamma+t+E} = \frac{\beta(E-\gamma-t)}{(\gamma+t)(\gamma+t+E)} \leq \frac{\beta(E-\gamma)}{(\gamma+t)(\gamma+t+E)} \leq 0.$$

To prove that $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$ for any t, it suffices to prove that for $0 \leq t \leq E$ because $\{\eta_t\}$, i.e., $t_0 = 0$, is non-increasing and $t+1-t_0$ is periodic with period E. When $t_0 = 0$, we need to prove $\eta_t^2 \leq 1/\big(2c_3(t+1)(1+2(t+1))\big)$ for $0 \leq t \leq E$, which is satisfied if

$$\begin{aligned}
&\max_{0\leq t\leq E}\eta_t \leq \min_{0\leq t\leq E}\sqrt{1/\big(2c_3(t+1)(1+2(t+1))\big)} \\
\iff &\eta_0 \leq \sqrt{\frac{1}{(4E^2+10E+6)c_3}} \\
\iff &\gamma \geq \beta\sqrt{(4E^2+10E+6)c_3} = \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)c_3} \\
&= \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}.
\end{aligned}$$

□

H PROOF OF CONVERGENCE UNDER THE ALTERNATIVE ASSUMPTION IN EQUATION B.1

Assumption H.1. Suppose the following hold

$$\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla l(\boldsymbol{\theta}; Z_i)\|_2^2] \leq G^2. \tag{14}$$

Lemma H.2. *(Bound on the divergence of parameters, i.e., consensus error bound)*
When $E > 1$, under Assumption 2.1, 2.2, 2.3, 2.5, and if η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ holds for all $t \geq 0$, we have

$$\mathbb{E}\left[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\right] \leq 4(E-1)^2\eta_t^2 G^2 \tag{15}$$

Proof. FedAvg requires a communication every E steps, so for any $t \geq 0$, there exists a $t_0 \leq t, t_0 \in \mathcal{I}_E$, such that $t - t_0 \leq E-1$ and $\boldsymbol{\theta}_i^{t_0} = \overline{\boldsymbol{\theta}}^{t_0}, \forall i$. Also, we use the fact that $\eta_{t_0} \leq 2\eta_t$ for all $t - t_0 \leq E-1$, then

$$\mathbb{E}\left[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^N p_i\|(\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}) - (\overline{\boldsymbol{\theta}}^t - \overline{\boldsymbol{\theta}}^{t_0})\|_2^2\right] \leq \mathbb{E}\left[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2\right], \tag{16}$$

since $\mathbb{E}\|X - \mathbb{E}X\|_2^2 \le \mathbb{E}\|X\|_2^2$ where $X = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}$. Using Jensen's inequality, we further have

$$\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2 = \Big\| \sum_{s=t_0}^{t-1} \eta_s \nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\Big\|_2^2 \le (t-t_0)\sum_{s=t_0}^{t-1}\eta_s^2 \big\|\nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\big\|_2^2, \tag{17}$$

$$\mathbb{E}\Big[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^N p_i \sum_{s=t_0}^{t-1}\eta_s^2 \nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\Big\|_2^2\Big] \le (t-t_0)\sum_{s=t_0}^{t-1}\eta_s^2 \sum_{i=1}^N p_i \mathbb{E}\Big[\big\|\nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\big\|_2^2\Big], \tag{18}$$

$$\mathbb{E}\Big[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}\|_2\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^N p_i \sum_{s=t_0}^{t-1}\eta_s \nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\Big\|_2\Big] \le \sum_{s=t_0}^{t-1}\eta_s \sum_{i=1}^N p_i \mathbb{E}\Big[\big\|\nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\big\|_2\Big], \tag{19}$$

where we used $\eta_s \le \eta_{t_0}$. Therefore, based on **A5**, we have

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^N p_i\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t}\|_2^2\Big] &\le \sum_{i=1}^N p_i \mathbb{E}\Big[\sum_{s=t_0}^{t-1}(E-1)\eta_s^2\|\nabla l(\boldsymbol{\theta}_i^s; Z_i^{s+1})\|_2^2\Big] \\
&\le \sum_{i=1}^N p_i \Big[\sum_{s=t_0}^{t-1}(E-1)\eta_s^2 G^2\Big] \\
&\le \sum_{i=1}^N p_i (E-1)^2\eta_{t_0}^2 G^2 \\
&\le 4(E-1)^2\eta_t^2 G^2
\end{aligned} \tag{20}$$

since $\eta_s \le \eta_{t_0} \le 2\eta_{t_0+E} \le 2\eta_t$ in the last two inequalities. □

Lemma H.3. *Li et al. (2022) Consider a sequence of non-negative, non-increasing step sizes $\{\eta_t\}_{t\ge 1}$. Let $a > 0, p \in \mathbb{Z}_+$ and $\eta_1 < 2/a$. If $\eta_t^p/\eta_{t+1}^p \le 1 + (a/2)\eta_{t+1}^p$ for any $t \ge 1$, then*

$$\sum_{j=1}^t \eta_j^{p+1}\prod_{\ell=j+1}^t(1-\eta_\ell a) \le \frac{2}{a}\eta_t^p, \quad \forall t \ge 1 \tag{21}$$

Lemma H.4. *Under Assumptions 2.1, 2.2, 2.3, H.1 and the condition that $\eta_t \le \tilde{\mu}/c_2$, , $\eta_t \le \eta_{t_0} \le 2\eta_t$ where $t_0 = \max_s\{s \in \mathbb{N}|Es \le t\}$, $\eta_{t+1} < \eta_t$ for any $t \ge 0$, $\eta_1 < \frac{2}{\tilde{\mu}}$ and $\eta_t^q/\eta_{t+1}^q \le 1 + (\tilde{\mu}/2)\eta_{t+1}^q$ for any $t \ge 0$ and $q = 1, 2, 3$.*

$$\mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \le \prod_{i=0}^t(1-\tilde{\mu}\eta_i)\big\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\big\|_2^2 + \frac{2c_2c_7}{\tilde{\mu}}\eta_t^3 + \frac{2c_1c_7}{\tilde{\mu}}\eta_t^2 + \frac{4\sigma^2}{\tilde{\mu}}\eta_t, \tag{22}$$

where $c_7 := 4(E-1)G^2$.

Proof. From Lemma F.1, we have

$$\begin{aligned}
\mathbb{E}\big[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2\big] &\le (1-\tilde{\mu}\eta_t)\mathbb{E}\Big[\big\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big\|_2^2\Big] + (c_1\eta_t + c_2\eta_t^2)\mathbb{E}\Big[\sum_{i=1}^N p_i\big\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\big\|_2^2\Big] + 2\sigma^2\eta_t^2 \\
&\le (1-\tilde{\mu}\eta_t)\mathbb{E}\Big[\big\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big\|_2^2\Big] + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + 2\sigma^2\eta_t^2 \\
&= \prod_{i=0}^t(1-\tilde{\mu}\eta_i)\big\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\big\|_2^2 + \sum_{s=1}^t\prod_{i=s+1}^t(1-\tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4 + c_1c_7\eta_s^3 + 2\sigma^2\eta_s^2\Big).
\end{aligned} \tag{23}$$

The second inequality holds because of Lemma H.2. Using Lemma H.3,

$$\sum_{s=1}^{t}\prod_{i=s+1}^{t}(1-\tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4+c_1c_7\eta_s^3+2\sigma^2\eta_s^2\Big)\le\frac{2c_2c_7}{\tilde{\mu}}\eta_t^3+\frac{2c_1c_7}{\tilde{\mu}}\eta_t^2+\frac{4\sigma^2}{\tilde{\mu}}\eta_t. \tag{24}$$

□

Theorem H.5. *(Full participation convergence theorem, alternative assumption)*
Under Assumption 2.1, 2.2, 2.3, 2.5, the full participation scheme has convergence rate $\mathcal{O}(\frac{1}{T})$, i.e., denote $\Delta_t := \mathbb{E}[\|\overline{\boldsymbol{\theta}}^t-\boldsymbol{\theta}^{PS}\|_2^2]$, then for some $\gamma>0$,

$$\Delta_t\le\frac{v}{\gamma+t}, \tag{25}$$

where $v:=\max\{\frac{c_2c_7\beta^4\gamma^{-2}+c_1c_7\beta^3\gamma^{-1}+2\sigma^2\beta^2}{\beta\mu-1},(\gamma+1)\Delta_1\}$.

Proof. We will show it on the partial participation algorithm, and the proof for the full participation is similar.

For a diminishing step size $\eta_t=\frac{\beta}{t+\gamma}$ for some $\beta>\frac{1}{\tilde{\mu}}$ and $\gamma>0$ such that $\eta_1\le\min\{\frac{1}{\tilde{\mu}},\frac{1}{4L}\}=\frac{1}{4L}$ and $\eta_t\le2\eta_{t+E}$. We will prove $\Delta_t:=\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t-\boldsymbol{\theta}^{PS}\|_2^2]\le\frac{v}{\gamma+t}$, where $v:=\max\{\frac{c_2c_7\beta^4\gamma^{-2}+c_1c_7\beta^3\gamma^{-1}+2\sigma^2\beta^2}{\beta\mu-1},(\gamma+1)\Delta_1\}$. We prove this by induction. Firstly, the definition if v ensures it holds for $t=1$. Assume it holds for some t, i.e., $\eta_t=\frac{\beta}{t+\gamma}$, then it follows from Lemma H.4 that

$$\begin{aligned}\Delta_{t+1}&\le(1-\eta_t\tilde{\mu})\Delta_t+c_2c_7\eta_t^4+c_1c_7\eta_t^3+2\sigma^2\eta_t^2\\&\le(1-\frac{\beta\tilde{\mu}}{t+\gamma})\frac{v}{t+\gamma}+\frac{c_2c_7\beta^4}{(t+\gamma)^4}+\frac{c_1c_7\beta^3}{(t+\gamma)^3}+\frac{2\sigma^2\beta^2}{(t+\gamma)^2}\\&=\frac{t+\gamma-1}{(t+\gamma)^2}v+\left[\frac{c_2c_7\beta^4}{(t+\gamma)^4}+\frac{c_1c_7\beta^3}{(t+\gamma)^3}+\frac{2\sigma^2\beta^2}{(t+\gamma)^2}-\frac{\beta\mu-1}{(t+\gamma)^2}v\right]\\&\le\frac{t+\gamma-1}{(t+\gamma)^2}v+\left[\frac{c_2c_7\beta^4}{(t+\gamma)^2\gamma^2}+\frac{c_1c_7\beta^3}{(t+\gamma)^2\gamma}+\frac{2\sigma^2\beta^2}{(t+\gamma)^2}-\frac{\beta\mu-1}{(t+\gamma)^2}v\right]\\&\le\frac{v}{t+\gamma+1}\end{aligned} \tag{26}$$

where $\tilde{\mu},c_1,c_2,c_3,c_7$ are defined the same as in earlier proofs, and thus the $\mathcal{O}(1/T)$ convergence rate is shown. □

Lemma H.6. *(Bounding the difference $\overline{\boldsymbol{w}}^{t+1}-\overline{\boldsymbol{\theta}}^{t+1}$ in partial participation)*
Suppose Assumption 2.1, 2.2, 2.3, and 2.5 hold. For $t+1\in\mathcal{I}_E$, assume that η_t is non-increasing and $\eta_t\le2\eta_{t+E}$ for all t, then we have the following results

1. *For Scheme I, the expected difference $\overline{\boldsymbol{w}}^{t+1}-\overline{\boldsymbol{\theta}}^{t+1}$ is bounded by*

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\le\frac{4}{K}\eta_t^2E^2G^2. \tag{27}$$

2. *For Scheme II, assuming $p_1=p_2=\cdots=p_N=\frac{1}{N}$, the expected difference $\overline{\boldsymbol{w}}^{t+1}-\overline{\boldsymbol{\theta}}^{t+1}$ is bounded by*

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\le\frac{4(N-K)}{K(N-1)}\eta_t^2E^2G^2 \tag{28}$$

Proof. We prove the bound for Scheme I as follows. Since $\overline{\boldsymbol{\theta}}^{t+1}=\frac{1}{K}\sum_{l=1}^{K}\boldsymbol{w}_{i_l}^t$, taking expectation over $\mathcal{S}_{t+1}$, we have

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2=\mathbb{E}_{\mathcal{S}_t}\frac{1}{K^2}\sum_{l=1}^{K}\|\boldsymbol{w}_{i_l}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2=\frac{1}{K}\sum_{k=1}^{N}p_k\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2 \tag{29}$$

We note that since $t+1 \in \mathcal{I}_E$, we know that the time $t_0 = t - E + 1 \in \mathcal{I}_E$ is the communication time, which implies $\{\boldsymbol{\theta}_{t_0}^k\}$ is identical. Then

$$\sum_{k=1}^{N} p_k \|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 = \sum_{i=1}^{N} p_k \|(\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}) - (\overline{\boldsymbol{w}}^{t+1} - \overline{\boldsymbol{\theta}}^{t_0})\|_2^2 \le \sum_{i=1}^{N} p_k \|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2 \tag{30}$$

Similar to Lemma H.2, the last inequality is due to $\mathbb{E}\|X - \mathbb{E}X\|_2^2 \le \mathbb{E}\|X\|_2^2$ where $X = \boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}$, and $\sum_{k=1}^N p_k(\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}) = \overline{\boldsymbol{w}}^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}$. Similarly, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}\big[\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] &\le \frac{1}{K}\sum_{k=1}^{N} p_k \mathbb{E}\big[\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2\big] \\
&= \frac{1}{K}\sum_{k=1}^{N} p_k \mathbb{E}\big[\|\boldsymbol{w}_k^{t+1} - \boldsymbol{\theta}_k^{t_0}\|_2^2\big] \\
&= \frac{1}{K}\sum_{k=1}^{N} p_k \mathbb{E}\Big[\big\|\sum_{s=t_0}^{t} \eta_s \nabla l(\boldsymbol{\theta}_k^s; Z_k^{s+1})\big\|_2^2\Big] \\
&\le \frac{1}{K}\sum_{k=1}^{N} p_k E \sum_{s=t_0}^{t} \mathbb{E}\big[\|\eta_s \nabla l(\boldsymbol{\theta}_k^s; Z_k^{s+1})\|_2^2\big] \\
&\le \frac{1}{K} E^2 \eta_{t_0}^2 G^2 \le \frac{4}{K}\eta_t^2 E^2 G^2.
\end{aligned} \tag{31}$$

Then we prove the bound for Scheme II. Since $\overline{\boldsymbol{\theta}}^{t+1} = \frac{1}{K}\sum_{l=1}^{K} \boldsymbol{w}_{i_l}^{t+1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\big[\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] \\
&= \mathbb{E}_{\mathcal{S}_t}\Big[\big\|\frac{1}{K}\sum_{l=1}^{K} \boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\big\|_2^2\Big] \\
&= \frac{1}{K^2}\mathbb{E}_{\mathcal{S}_t}\Big[\big\|\sum_{i=1}^{N} \mathbf{1}\{i \in \mathcal{S}_t\}(\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1})\big\|_2^2\Big] \\
&= \frac{1}{K^2}\Big[\sum_{i=1}^{N} \mathbb{P}(i \in \mathcal{S}_t)\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \sum_{j \neq i} \mathbb{P}(i, j \in \mathcal{S}_t)\left\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\right\rangle\Big] \\
&= \frac{1}{KN}\sum_{i=1}^{N} \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \frac{K-1}{KN(N-1)}\sum_{i \neq j}\left\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\right\rangle \\
&= \frac{1}{K(N-1)}(1 - \frac{K}{N})\sum_{i=1}^{N} \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2.
\end{aligned} \tag{32}$$

Note that the second last equality holds because $\mathbb{P}(i \in \mathcal{S}_t) = \frac{K}{N}$ and $\mathbb{P}(i, j \in \mathcal{S}_t) = \frac{K(K-1)}{N(N-1)}$; and the last equality holds because

$$\begin{aligned}
&\sum_{i=1}^{N} \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \sum_{i \neq j}\left\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\right\rangle \\
&= \sum_{i=1}^{N}\left\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \left(\sum_{j=1}^{N} \boldsymbol{w}_j^{t+1}\right) - N\overline{\boldsymbol{w}}^{t+1}\right\rangle = 0.
\end{aligned}$$

Recall that

$$\sum_{k=1}^{N} p_k \|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \le \sum_{i=1}^{N} p_k \|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2,$$

we get

$$\begin{aligned}\mathbb{E}\big[\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] &= \frac{1}{K(N-1)}(1-\frac{K}{N})\mathbb{E}\Big[\sum_{i=1}^{N}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\Big] \\ &\le \frac{N}{K(N-1)}(1-\frac{K}{N})\mathbb{E}\Big[\sum_{i=1}^{N}\frac{1}{N}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2\Big] \\ &\le \frac{N}{K(N-1)}(1-\frac{K}{N})4\eta_t^2E^2G^2 = \frac{4(N-K)}{K(N-1)}\eta_t^2E^2G^2 \qquad (33)\end{aligned}$$

where the last inequality can be found in equation 20 in the proof of Lemma H.2. □

Lemma H.7. *Under Under Assumption 2.1, 2.2, 2.3, 2.5, and the condition that* $\eta_t \le \tilde{\mu}/c_2$, , $\eta_t \le \eta_{t_0} \le 2\eta_t$ *where* $t_0 = \max_s\{s \in \mathbb{N}|Es \le t\}$, $\eta_{t+1} < \eta_t$ *for any* $t \ge 0$, $\eta_1 < \frac{2}{\tilde{\mu}}$ *and* $\eta_t^q/\eta_{t+1}^q \le 1 + (\tilde{\mu}/2)\eta_{t+1}^q$ *for any* $t \ge 0$ *and* $q = 1, 2, 3$, *we have*

$$\mathbb{E}_{\mathcal{S}_t}[||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2] \le \prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + \frac{2c_2c_3}{\tilde{\mu}}\eta_t^3 + \frac{2c_1c_3}{\tilde{\mu}}\eta_t^2 + \frac{2c_8}{\tilde{\mu}}\eta_t, \qquad (34)$$

(c_8 *for Scheme I, replace* c_8 *with* c_9 *in Scheme II) where we define* $c_8 := 2\sigma^2 + \frac{4}{K}E^2G^2$ *in Scheme I, and* $c_9 := 2\sigma^2 + \frac{4(N-K)}{K(N-1)}E^2G^2$ *in Scheme II.*

Proof. Note that

$$\begin{aligned} & ||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2 \\ =\ & ||\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1} + \overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2 \\ =\ & \underbrace{||\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}||_2^2}_{T_1} + \underbrace{||\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2}_{T_2} + \underbrace{2\langle\overline{\boldsymbol{w}}^{t+1} - \overline{\boldsymbol{\theta}}^{t+1},\ \overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\rangle}_{T_3} \qquad (35)\end{aligned}$$

When expectation is taken over $\mathcal{S}_{t+1}$, the last term T_3 vanishes due to Lemma G.1.

If $t+1 \notin \mathcal{I}_E$, T_1 vanishes since $\overline{\boldsymbol{\theta}}^{t+1} = \overline{\boldsymbol{w}}^{t+1}$ by definition when $t+1$ is not a communication step. For term T_2, it's not hard to see that we can use Lemma F.1 to derive one step bounds for it (and use equation 23 in Lemma H.4), and thus we have

$$\begin{aligned}\mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2\big] &= \mathbb{E}\big[||\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2\big] \\ &\le (1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + 2\sigma^2\eta_t^2, \qquad (36)\end{aligned}$$

and we recall that $c_1 := \frac{L^2(1+\epsilon_{max})^2}{2\delta\overline{\epsilon}}, c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2], c3 := 4(E-1)^2G^2, \tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L$.

If $t+1 \in \mathcal{I}_E$, then we have the following result from Lemma H.6,

$$\begin{aligned} & \mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2\big] \\ =\ & \mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}||_2^2\big] + \mathbb{E}\big[||\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2\big] \\ \le\ & (1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + c_8\eta_t^2, \qquad (37)\end{aligned}$$

where we recall $c_8 := 2\sigma^2 + \frac{4}{K}E^2G^2$ in Scheme I, and $c_9 := 2\sigma^2 + \frac{4(N-K)}{K(N-1)}E^2G^2$ in Scheme II.

The only difference between equation 24 and equation 37 is in $(c_8 - 2\sigma^2)\eta_t^2$. Therefore, we can use similar techniques to show the convergence,

$$\begin{aligned}
\mathbb{E}[||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2] &\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + c_8\eta_t^2 \\
&= \prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 \\
&\quad + \sum_{s=1}^{t}\prod_{i=s+1}^{t}(1-\tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4 + c_1c_7\eta_s^3 + c_8\eta_s^2\Big) \\
&\leq \prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + \frac{2c_2c_7}{\tilde{\mu}}\eta_t^3 + \frac{2c_1c_7}{\tilde{\mu}}\eta_t^2 + \frac{2c_8}{\tilde{\mu}}\eta_t
\end{aligned}$$

(c_8 for Scheme I, replace c_8 with c_9 in Scheme II). □

Theorem H.8. *(Full participation convergence theorem, alternative assumption)*
Under Assumption 2.1, 2.2, 2.3, 2.5, the full participation scheme has convergence rate $\mathcal{O}(\frac{1}{T})$, i.e., denote $\Delta_t := \mathbb{E}[||\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}||_2^2]$, then for some $\gamma > 0$,

$$\Delta_t \leq \frac{v}{\gamma + t}, \tag{38}$$

where $v := \max\{\frac{c_2c_7\beta^4\gamma^{-2} + c_1c_7\beta^3\gamma^{-1} + c_8\beta^2}{\beta\mu - 1}, (\gamma+1)\Delta_1\}$ (c_8 for Scheme I, replace c_8 with c_9 in Scheme II).

Proof. We will show it on the partial participation algorithm, and the proof for the full participation is similar.

For a diminishing step size $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\tilde{\mu}}$ and $\gamma > 0$ such that $\eta_1 \leq \min\{\frac{1}{\tilde{\mu}}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. We will prove $\triangle_t := \mathbb{E}[||\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}||_2^2] \leq \frac{v}{\gamma+t}$, where $v := \max\{\frac{c_2c_7\beta^4\gamma^{-2} + c_1c_7\beta^3\gamma^{-1} + c_8\beta^2}{\beta\mu - 1}, (\gamma+1)\triangle_1\}$. We prove this by induction. Firstly, the definition if v ensures it holds for $t = 1$. Assume it holds for some t, i.e., $\eta_t = \frac{\beta}{t+\gamma}$, then it follows that

$$\begin{aligned}
\triangle_{t+1} &\leq (1-\eta_t\tilde{\mu})\triangle_t + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + c_8\eta_t^2 \\
&\leq (1 - \frac{\beta\tilde{\mu}}{t+\gamma})\frac{v}{t+\gamma} + \frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{c_8\beta^2}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{c_8\beta^2}{(t+\gamma)^2} - \frac{\beta\mu - 1}{(t+\gamma)^2}v\right] \\
&\leq \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^2\gamma^2} + \frac{c_1c_7\beta^3}{(t+\gamma)^2\gamma} + \frac{c_8\beta^2}{(t+\gamma)^2} - \frac{\beta\mu - 1}{(t+\gamma)^2}v\right] \\
&\leq \frac{v}{t+\gamma+1}
\end{aligned} \tag{39}$$

(c_8 for Scheme I, replace c_8 with c_9 in Scheme II), where $\tilde{\mu}$, c_1 to c_9 are defined the same as in earlier proofs, and thus the $\mathcal{O}(1/T)$ convergence rate is shown. □