Performative Federated Learning: A Solution to Model-Dependent and Heterogeneous Distribution Shifts

Kun Jin[*1], **Tongxin Yin**[*1], **Zhongzhu Chen**[*1],
Zeyu Sun [1], **Xueru Zhang** [2], **Yang Liu** [3], **Mingyan Liu** [1]

[1] University of Michigan
[2] The Ohio State University
[3] University of California, Santa Cruz

Abstract

We consider a federated learning (FL) system consisting of multiple clients and a server, where the clients aim to collaboratively learn a common decision model from their distributed data. Unlike the conventional FL framework that assumes the client's data is static, we consider scenarios where the clients' data distributions may be reshaped by the deployed decision model. In this work, we leverage the idea of distribution shift mappings in *performative prediction* to formalize this model-dependent data distribution shift and propose a performative FL framework. We first introduce necessary and sufficient conditions for the existence of a unique performative stable solution and characterize its distance to the performative optimal solution. Then 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 particular, we use novel proof techniques and show how the clients' heterogeneity influences the convergence. Numerical results validate our analysis and provide valuable insights into real-world applications.

1 Introduction

Traditional learning problems often assume static data distributions, which holds true for applications like face recognition. However, in many other domains, this assumption is invalid. In some cases, there is a natural evolution or shift in the data distribution, requiring periodic acquisition of new data and re-training of the algorithm. Additionally, distribution shifts can occur as a result of user responses to algorithmic decisions or attempts to manipulate the system. These changes directly impact the features and labels used by the algorithm for decision-making. Such shifts are considered *model-dependent*. A typical example is banks' loan issuance decisions. The deployed decision model at this bank will influence the data distribution of all its corresponding applicants. For example, if increasing the number of credit cards significantly decreases the default rate prediction, then the applicants will try to get more credit cards to get the loan.

Adapting algorithms to these evolving distributions is crucial for maintaining effective learning performance. The model-dependent distribution shifts, where the deployed model itself can trigger changes in the data distribution and influence the objective, is said to be *performative*. Performing prediction in the presence of such distribution shift is called *performative prediction* (PP) (Perdomo et al. 2020). The *strategic learning* problem (Hardt et al. 2016; Dong et al. 2018; Milli et al. 2019; Hu, Immorlica, and Vaughan 2019; Braverman and Garg 2020; Chen, Wang, and Liu 2020; Miller, Milli, and Hardt 2020; Shavit, Edelman, and Axelrod 2020; Haghtalab et al. 2020; Kleinberg and Raghavan 2020; Zrnic et al. 2021) is a typical scenario of PP. In these problems, the users can "game the algorithm" through honest or dishonest means to attempt to improve critical features so as to obtain a favorable decision by the algorithm (e.g., in loan approvals or job applications). Such user actions directly lead to the distributional change in features and label that the algorithm relies on for decision making.

Performative prediction has been primarily studied in a centralized setting, with fruitful literature including the convergence analysis (Mendler-Dünner et al. 2020; Drusvyatskiy and Xiao 2020; Brown, Hod, and Kalemaj 2020; Li and Wai 2022; Wood, Bianchin, and Dall'Anese 2022) and algorithm development (Izzo, Ying, and Zou 2021; Izzo, Zou, and Ying 2022; Miller, Perdomo, and Zrnic 2021; Ray et al. 2022).

In modern large-scale machine learning, distributed learning offers greater privacy protection and better avoids the computational resource bottlenecks compared to centralized learning, and federated learning (FL) is a very popular example. Suppose multiple banks use FL to jointly train a model to predict applicants' default rate. The aggregated model on the central server is influenced by the strategic manipulation of each bank's applicants, and this model will be later deployed to every bank, which influences that bank's applicants' strategic manipulation in the next round. 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 shifts, such as (Guo, Lin, and Tang 2021; Casado et al. 2022; Rizk, Vlaski, and Sayed 2020; Hosseinalipour et al. 2022; Zhu et al. 2021; Eichner et al. 2019; Ding et al. 2020), typically do not consider shifts in local distributions at the client end induced by the model. In this work, we propose the *performative FL* framework to study and handle such data shifts in FL.

[*]These authors contributed equally.

Extending the current results in PP to the decentralized FL has a number of challenges. To highlight a few: 1) *Data heterogeneity:* As already one of the major difficulties in FL, tackling data heterogeneity faces additional challenges when considering the disparity of client distribution shift. 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.

Recently, Raab and Liu (2021); Li, Yau, and Wai (2022); Narang et al. (2022) generalizes the PP beyond the centralized setting, and formalize the multi-agent/player PP problem to address the data and shifts heterogeneity challenges mentioned above. In Li, Yau, and Wai (2022), 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. Narang et al. (2022) propose a decentralized multi-player PP framework where the players react to competing institutions' actions. Raab and Liu (2021) proposes a replicator dynamics model with label shift and Yin et al. (2023) proposes an reinforcement learning method that works on this dynamic. The multi-agent PP framework provides inspiration for our formulation of the performative FL framework.

However, these works are missing two key properties to fit for FL, *multi-step aggregations* and *partial participation*, which makes the FL system practical and efficient.

In this paper, we formally introduce the *performative FedAvg* algorithm, or `P-FedAvg`, and establish its convergence. `P-FedAvg` can be viewed as a substantial algorithmic extension to multi-agent PP algorithms since it supports unbalanced data, much less frequent synchronizations (multi-step aggregation), and partial device participation. Our main findings are as follows.

- We prove the uniqueness of the performative stable (PS) solution in the performative FL problem, and show that it is a provable approximation to the performative optimal (PO) solution under mild conditions. Both the PS and PO solutions will be formally defined in Section 2.1.
- We show in Section 3.3 that the `P-FedAvg` algorithm converges to the PS 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.
- In doing so we also introduce some novel proof techniques: we prove convergence of `P-FedAvg` without a bounded gradient assumption, and instead use a relaxed assumption that characterizes the clients' heterogeneity. The new proof techniques illustrate how the heterogeneity influences the convergence and they also work on conventional FL problems with static data distributions.
- To our best knowledge, we are the first to define and study the performative predictions in computer vision tasks and show interesting empirical convergence results.

Our work is closely related to the works in FL (Li et al. 2020a; Karimireddy et al. 2020; Wang et al. 2020; Haddadpour et al. 2021; Zhu, Hong, and Zhou 2021; Li and Wang 2019; Lin et al. 2020; Guo, Lin, and Tang 2021; Casado et al. 2022; Rizk, Vlaski, and Sayed 2020; Hosseinalipour et al. 2022; Zhu et al. 2021; Eichner et al. 2019; Ding et al. 2020), strategic learning, PP, and multi-agent PP, where we highlight the key properties of our work and demonstrate its differences with previous works in Table 1. And we refer the reader to the definition of each property in Section 2. Specifically, compared to the multi-agent PP, the performative FL framework utilize the *multi-step aggregation* and *partial participation* schemes to significantly improve the system's efficiency. The technical challenges are explained in Appendix B. Please also see more on related works in Appendix A.

2 Problem Formulation

In this section, we formulate the performative FL problem, define the learning objective, and introduce our performative FL algorithm to optimize the objective function.

To help with the understanding of performative FL, we first recall the performative prediction problem (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 $\mathcal{D}(\boldsymbol{\theta})$. Such a distribution shift is model-dependent, and is called **Performative Shift**. 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. Then the performative optimal (PO) solution is $\boldsymbol{\theta}^{PO} := \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$. (Perdomo et al. 2020) also introduces a 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. Naturally, Perdomo et al. (2020) also showed algorithms that ignore Performative Shifts will not find the PS or PO solution in general. We next consider a distributed setting and introduce performative FL.

2.1 System Settings and Objectives

Consider a system with N clients and a server, where the clients have feature distributions as $\mathcal{D}_i(\boldsymbol{\theta})$, supported on $\mathcal{Z} \subseteq \mathbb{R}^M$, and $\boldsymbol{\theta} \in \mathbb{R}^m$ denotes the decision (model) parameters deployed on the i-th client. We consider the 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, $\sum_{i=1}^N p_i = 1$. We would like to emphasize that in contrast to static federated learning, the dynamic setting not only encompasses **data heterogeneity**, which refers to the varying initial distribution of data among clients, but also **shift heterogeneity**, wherein the shift mapping $\mathcal{D}_i(\boldsymbol{\theta})$ differs

Related Works	Performative Shifts (Sec 2)	Data Heterogeneity (Sec 2.1)	Multi-step Aggregation (Sec 2.4)	Partial Participation (Sec 2.4)
Strategic Learning & PP	✓			
Federated Learning		✓	✓	✓
Multi-agent PP (Li, Yau, and Wai 2022)	✓	✓		
Our Work	✓	✓	✓	✓

Table 1: A Summary of Key Properties of Our Works and Related Fields.

across clients. Additionally, when $\mathcal{D}_i(\boldsymbol{\theta}) = \mathcal{D}_i$ are fixed, the problem returns to conventional FL.

The system aims to minimize the weighted average loss across all agents, which is given by the performative optimal objective as follows

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

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, Immorlica, and Vaughan 2019; Braverman and 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 $\mathcal{D}_i(\cdot)$. The PS solution is

$$\begin{aligned} \boldsymbol{\theta}^{PS} &:= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^N p_i \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PS})}[\ell(\boldsymbol{\theta}; Z_i)] \\ &= \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{\theta}^{PS}). \end{aligned} \quad (2)$$

This is a fixed point equation with $\boldsymbol{\theta}^{PS}$ as a fixed point.

2.2 Key Assumptions

We make similar but weaker assumptions compared to (Li, Yau, and Wai 2022; Perdomo et al. 2020).

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., $f(\boldsymbol{\theta}'; \tilde{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \langle \nabla f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{\mu}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2, \forall \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^K$.

In Assumption 2.1, we do not require strong convexity for every f_i but only the f.

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 $i = 1, \ldots, n$, $\exists \epsilon_i > 0$ 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 $\mathcal{D}, \mathcal{D}'$.

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 and Xiao 2022) and will be used in the later proofs.

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

We introduce the following assumptions specifically made in decentralized PP (Li, Yau, and Wai 2022).

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

Here we elaborate on Assumption 2.6, and explain our contribution for using it over another commonly used assumption in FL (Li et al. 2020b), which is

$$\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla\ell(\boldsymbol{\theta}; Z_i)\|_2^2] \leq G^2. \quad (3)$$

First, we can show (3) is a stronger condition than Assumption 2.6. To see this: when (3) holds, let $\varsigma^2 = 4G^2$, then $\|\nabla f(\boldsymbol{\theta}; \boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|_2^2 \leq 2\left\|\nabla f(\boldsymbol{\theta}; \boldsymbol{\theta})\right\|_2^2 + 2\|\nabla f_i(\boldsymbol{\theta}; \boldsymbol{\theta})\|_2^2 \leq 4G^2 = \varsigma^2$.

We further give a concrete example where (3) does not hold but Assumption 2.6 holds.

Example 2.7. *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* (3) *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, (3) also implies Assumption 2.5: when (3) 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. Moreover, 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).

2.3 Properties of the PS Solution

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

Proposition 2.8 (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 2.8 establishes a sufficient and necessary condition for the existence of $\boldsymbol{\theta}^{PS}$, similar to (Li, Yau, and Wai 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 2.9 (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* $\|\boldsymbol{\theta}^{PS} - \boldsymbol{\theta}^{PO}\|_2 \leq \big(2L_z\bar{\epsilon}\big)/\mu$.

Please find all proofs in the Appendix.

2.4 The `P-FedAvg` Algorithm

In `P-FedAvg`, the clients communicate with the server every E local updates (**Multi-step Aggregation** for $E > 1$). Denote $\mathcal{I}_E := \{nE | n = 1, 2, \dots\}$ as the set of aggregation steps. Next, we formalize the full and partial participation schemes of the proposed **`P-FedAvg`**.

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 $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$,

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

Partial client participation. A more realistic setting that does not require the response of all clients' output at every aggregation step. In this case, the central server only collects the outputs of the first $K < N$ responded clients at the aggregation step. Denote the first $K < N$ responded clients in t-th step as a size-K set $\mathcal{S}_t := \{i_1, \dots, i_K\} \in [N]$. Let $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$, then

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

We further consider two schemes of partial participation:

1. (**Scheme I**) 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. Hence $\mathcal{S}_{t+1}$ is a multiset that allows an element to occur more than once. Then the server averages the parameters by $\boldsymbol{\theta}_i^{t+1} = \frac{1}{K}\sum_{k \in \mathcal{S}_{t+1}} \boldsymbol{w}_k^{t+1}$. This sampling scheme is first proposed in (Sahu et al. 2018) but the theoretical analysis was first done in (Li et al. 2020b).
2. (**Scheme II**) The server samples $\mathcal{S}_{t+1}$ uniformly ***without replacement***. Hence each element in $\mathcal{S}_{t+1}$ only occurs once. Then the server 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 we cannot ensure $\sum_{k \in \mathcal{S}_{t+1}} p_k \frac{N}{K} = 1$ unless $p_k = \frac{1}{N}, \forall k$ (Li et al. 2020b).

Partial participation enhances real-world applicability. For instance, consider multiple banks collaborating to develop a loan approval model. Each bank communicates and syncs data only when a sufficient amount of new local data is gathered. The frequency of communication may vary significantly among banks due to local loan demand. Consequently, agents are expected to participate only partially over time.

Communication cost. 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. Our use of multi-step aggregation reduces communication costs compared to Multi-agent PP, especially in scenarios with high communication costs. In the same time period, `P-FedAvg` runs fewer communication steps but much more computation steps compared with Multi-agent PP, and potentially enables much faster convergence.

Next, we'll prove that `P-FedAvg` has $\mathcal{O}(1/T)$ convergence rate under the above assumption. As a supplement, we prove in Appendix F that `P-FedAvg` also has $\mathcal{O}(1/T)$ convergence rate if we replace Assumption 2.6 with (3).

3 Convergence Analysis

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 key observation is that for sufficiently small and decaying learning rates, the effect of E steps is similar to a one-step update with a larger learning rate in the static case, as stated in (Li et al. 2020a) without the performative setting. Therefore, given appropriate sampling and updating schemes that satisfy the above assumptions, the global update behaves similarly to the repeated performative SGD in (Perdomo et al. 2020). We

also show that partial device participation makes the averaged parameter sequence $\{\overline{\boldsymbol{\theta}}^t\}$ have the same mean as but a larger variance than the full participation, where the variance can be controlled with carefully chosen learning rates. It's worth noting that the heterogeneity of clients plays a key role in the convergence analysis, which we elaborate on below.

Quantifying the heterogeneity. The client heterogeneity can be quantified by the consensus error $\sum_{i=1}^N p_i \|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$, which is dynamic due to the nature of performative prediction. It depends on both the shift mappings $\mathcal{D}_i$ and the decision parameters. After every broadcast, the heterogeneity leads to heterogeneous distribution shifts, causing heterogeneous local updates, resulting in the consensus error. The ς value in Assumption 2.6 is also a good indicator for heterogeneity.

Next, we will first present the convergence analysis of the full participation scheme and later extend the analysis to partial participation schemes. Due to the complexity of analysis in the performative setting, we define the constants in Table 2 for ease of analysis and clarity of presentation.

3.1 Convergence of Full Participation

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

The key to the proof is that the expected distance $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2$ and the expected consensus error $\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2$ all depend on expected distance $\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t-1} - \boldsymbol{\theta}^{PS}\|_2^2$ and expected consensus error $\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t-1} - \overline{\boldsymbol{\theta}}^{t-1}\|_2^2$ in the previous step. While we can establish a descent lemma for expected distance including the expected consensus error, it is impossible to establish one for expected consensus error, which makes it impossible to establish a joint descent lemma for expected distance and expected consensus error as in (Li, Yau, and Wai 2022). Fortunately, consensus error will become zero at every aggregation step, which enables us to control expected consensus error at every step within a constant with a novel double-iteration technique under small enough step sizes, and establish a standard descent lemma in SGD analysis for expected distance.

3.2 Convergence of Partial Participation

As mentioned in Section 2, the partial participation scheme is more realistic in FL (Li et al. 2020b) and is of more interest since it reduces the stragglers' effect.

We first present the convergence result of Scheme I.

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, we have* $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma+t}$, $\forall t$ *where* $\upsilon = \max\left\{\frac{4B_1}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.

Then we present the convergence result of Scheme II. As discussed in Section 2, we need probabilities $p_i = \frac{1}{N}, \forall i$ to ensure $\sum_{i\in\mathcal{S}_t} p_k \frac{N}{K} = 1$.

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, we have* $\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma+t}$, $\forall t$ *where* $\upsilon = \max\left\{\frac{4B_2}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.

Besides the technical difficulty as that of Theorem 3.1, we also need to bound the variance of $\overline{\boldsymbol{\theta}}^t$ at the aggregation step. Fortunately, it can be bounded by the consensus error. Similar to the proof of Theorem 3.1, we can establish a standard descent lemma in SGD analysis for the expected distance.

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 $q_{max} := N \cdot \max_i p_i, q_{min} := N \cdot \min_i p_i$.

3.3 Discussions on the `P-FedAvg` Design

For conciseness, we focus on the aggregation step denoted as $T \in \mathcal{I}_E$, then $\frac{T}{E}$ denotes the corresponding number of communication rounds.

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 $T_\epsilon := \frac{\upsilon}{\epsilon} - \gamma$ as the number of computation steps that is sufficient to guarantee an ϵ-accuracy. 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, B_i 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, $C \cdot \mathcal{O}(E \log E) + C \cdot \mathcal{O}(\log E/E)$

System independent constants		System dependent constants	
$\bar{\epsilon} :=$	$\sum_{i=1}^N p_i \epsilon_i$	$c_6 :=$	$(2E^2+3E+1)\log(E+1)$
$\epsilon_{max} :=$	$\max_i \epsilon_i$	$\tilde{\eta}_0 :=$	$\tilde{\mu}/\big(2\sigma^2+(c_1c_3+c_2/6)c_4c_6\big)$
$\tilde{\mu} :=$	$\mu-(1+\delta)\bar{\epsilon}L$	$\hat{\eta}_0 :=$	$\tilde{\mu}/\big(2\sigma^2+(c_1c_3+c_2/6)c_4(2E^2-E)\log E\big)$
$c_1 :=$	$\big(L(1+\epsilon_{max})^2\big)/(2\delta\bar{\epsilon})$	$B :=$	$2\sigma^2+(4c_1\hat{\eta}_0+4c_2\hat{\eta}_0^2)c_5(2E^2-E)\log E$
$c_2 :=$	$4\big[\sigma^2+L^2(1+\epsilon_{max})^2\big]$	$B_1 :=$	$2\sigma^2+(4c_1\tilde{\eta}_0+4c_2\tilde{\eta}_0^2+1/K)c_5c_6$
$c_3 :=$	$6\big[2\sigma^2+3L^2(1+\epsilon_{\max})^2\big]$	$B_2 :=$	$2\sigma^2+\big(4c_1\tilde{\eta}_0+4c_2\tilde{\eta}_0^2+\frac{N-K}{K(N-1)}\big)c_5c_6$
$c_4 :=$	$16\sigma^2+12\varsigma^2+(8\sigma^2+12\varsigma^2)/\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2$		
$c_5 :=$	$(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)$		

Table 2: Constants for Convergence Analysis

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 $i=1,2$. By the formulation 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. In general, we show in Theorem 3.2 and 3.3, the convergence rate has a weak dependence on K, which is also empirically observed in Figure 6(a). Therefore, we can set $\frac{K}{N}$ to an appropriate small value to reduce the straggler's effect while keeping the convergence rate.

We note that our discussions on the choice of sampling schemes and the learning rate decay are similar to Li et al. (2020b), and please refer to Appendix G for more details.

4 Numerical Experiments

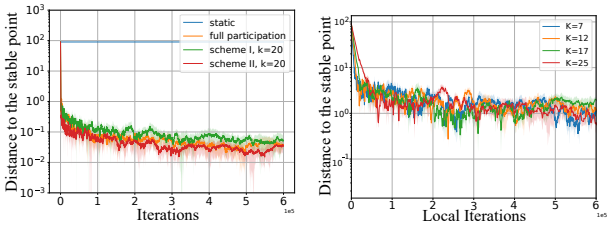
Our experiment is under a decentralized setting and we can use static data FL and multi-agent PP algorithms as baselines. For clarity of presentation, we only show one static data FL algorithm, `FedAvg`, since all static data FL algorithms will converge to a static solution with a constant bias to the PS solution due to performative shift. When we choose $E=1$ and full participation, `P-FedAvg` is equivalent to multi-agent PP Li, Yau, and Wai (2022) with a fully connected communication graph. The comparison of `P-FedAvg` with multi-agent PP is mainly around the convergence time. This is comparison not straightforward (Section 3.3) since we need extra settings to model the communication graph, the communication time, and the straggler's effect. Our code is publicly accessible[1]. Please see Appendix H for more details.

4.1 Performative Gaussian Mean Estimation

We perform `P-FedAvg` to estimate the mean of heterogeneous Gaussian distribution under performative effects and examine the impact of the hyperparameters, the sampling schemes, and client heterogeneity. This experiment is also used in PP (Perdomo et al. 2020) and multi-agent PP (Li, Yau, and Wai 2022) literature to clearly illustrate the impact of different system design parameters on the convergence. 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 $\mathrm{Var}(m)=\sum_{i=1}^N p_i(m_i-\overline{m})^2$. In this experiment, we set $\bar{\epsilon}=0.9$, $\overline{m}=10$.

Figure 1 (a) shows `P-FedAvg` converges to the PS solution in all three communication settings: full participation, Scheme I and Scheme II. Interestingly, 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 $p_i=\frac{1}{N}$. The static `FedAvg` converges to the static solution, which has a constant bias to the PS solution, was shown in the centralized setting Perdomo et al. (2020).

Impact of E. We conduct an experiment to compare the performance of `P-FedAvg` with different E values, in systems with high and low communication costs respectively. We let $C=20$ (resp. $C=5$) in the high (resp. low) cost system, and let $K=N$ to solely show the influence of E in Figure 2. More figures can be found in Appendix H.



(a) Distance to the PS solution (b) Impact of K

Figure 1: (a) Distance to the PS solution vs. the number of iterations for full and partial participation. (b) Impact of K, comparison with the multi-agent PP ($K=25$), $N=25$.

Impact of K and sampling schemes. Figure 1(b) show the impact of K on the convergence time. We fix $E=1$, thus $K=25$ represent Multi-agent PP. As discussed in Section

[1]https://github.com/tsy19/PerformativeFedAvg

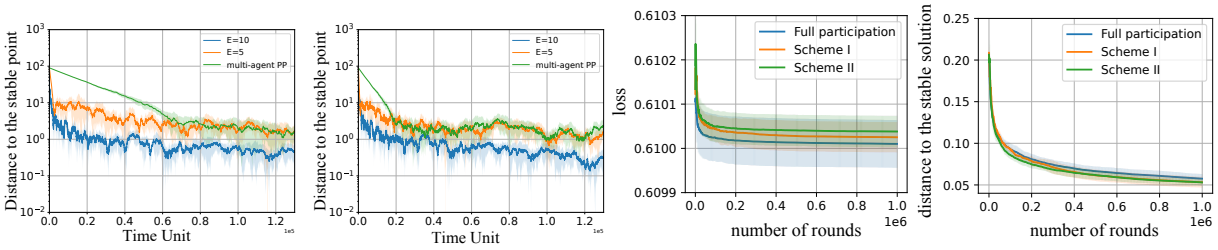


Figure 2: (a)(b) compares the impact of E with multi-agent PP in high ($C = 20$) and low ($C = 5$) communication cost systems. The efficiency of `P-FedAvg` surpasses multi-agent PP due to multi-step aggregation, with the advantage growing as communication cost increases. (c)(d) show losses and distances to the PS solution for full participation, Scheme I, and Scheme II, initialized with the empirical risk minimization solution. Mean and 1 std error bars are based on 5 individual runs.

3.3, a moderate number of K can control the balance between gradient variance and the straggler's effect. The details of how we model the straggler's effect can be found in Appendix H. Figure 1(a) also compares different schemes. If the clients' data are uniformly sampled ($p_i = \frac{1}{N}$), scheme II achieves a better convergence rate, which conforms to our theoretical result because $B_1 > B_2$. We also show the impact of Data heterogeneity and shifting heterogeneity in Appendix H.

4.2 Credit Score Strategic Classification

Demonstrating the efficacy of `P-FedAvg` on the Kaggle dataset[2] as per Perdomo et al. (2020). The dataset involves a bank predicting the creditworthiness of loan applicants, where features pertain to individual information, and the target is binary (1 for loan default, 0 otherwise). Following the performative shift framework of Perdomo et al. (2020), applicants can manipulate certain features related to credit lines and loans. Manipulation strength, denoted by ϵ_i, is independently and uniformly sampled from $[0.9, 1.1]$ for 10 clients, each receiving a 10% subset of the training data. Employing partial participation with $K = 5$, we train a logistic regression binary classifier using `P-FedAvg`, involving 5 gradient descent steps per round on a minibatch of size 4. Further details are discussed in Appendix H.4. Figure 2 (c)(d) shows the loss function and the distance to the PS solution as the number of deployment rounds increases. Similar to the numerical simulation, the actual convergence behaviors of all three schemes are very similar.

4.3 Performative Image Classification

We show empirically that `P-FedAvg` also has good performance on FMNIST (Deng 2012) and Cifar-10. Consider K classes of images and N clients, where each client has an arbitrary number of images from each class. At each time step t, we assume each client aims to achieve a good yet balanced classification outcome. Client i aims to minimize the objective $\min_{\boldsymbol{\theta}} \max_k \ell_{i,k}^{(t)}(\boldsymbol{\theta}_i^{(t)})$. The clients will attempt to optimize this objective by selecting sampling weights from each class at the next step, following the rule: $w_{i,k}^{(t+1)} \propto \exp(\beta \cdot \ell_{i,k}^{(t)})$, where $\beta > 0$ is the chosen temperature. A higher value of β leads to more aggressive changes in sample weights, we let $\beta = 0.5$. For heterogeneous performativeness, we employ distinct image sets across various clients, causing them to experience varying class losses and resulting in heterogeneous sample weight updates. We normalize the sample weights according to $w_{i,k}^{(t+1)} = \frac{exp(\beta \cdot a_{i,k}^{(t)})}{\sum_{k'} exp(\beta \cdot a_{i,k'}^{(t)})}$, where a can be either loss or accuracy, then track the total variation (TV) distances of consecutive aggregation steps and shows $d_{TV}(\boldsymbol{w}_i^{(t)}, \boldsymbol{w}_i^{(t-E)}) := \sum_k \frac{1}{2}|w_{i,k}^{(t)} - w_{i,k}^{(t-E)}|, \ \forall t \in \mathcal{I}_E$. Figure 3 shows that the average TV distance between consecutive aggregation steps is diminishing, showing that `P-FedAvg` converges to a PS solution. Please refer to Appendix H for results on Cifar-10.

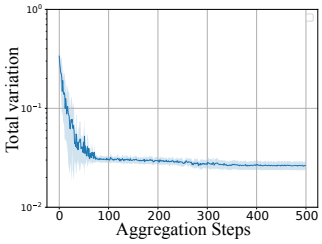


Figure 3: Convergence of the P-FedAvg on performative FMNIST classification.

5 Conclusions

In this work, we formulated the performative FL problem where data shifts of heterogeneous clients are model-dependent. We showed that a unique PS solution exists, and formalized the `P-FedAvg` algorithm where both the full participation and the partial participation schemes have $\mathcal{O}(1/T)$ convergence rate to the PS solution. We discussed the impact of key variables on the convergence, especially the aggregation interval size and the number of sampled devices in partial participation. Our numerical results validate our theory and provide valuable insights into the real-world applications of performative FL. To our best knowledge, we are the first to define performative shifts in computer vision tasks and show that `P-FedAvg` has good empirical convergence result.

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

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under grants IIS-2040800, IIS-2112471, IIS-2143895 and IIS-2202699; and a grant from the Ohio State University Translational Data Analytics Institute.

References

Braverman, M.; and Garg, S. 2020. The Role of Randomness and Noise in Strategic Classification. In *1st Symposium on Foundations of Responsible Computing*.

Brown, G.; Hod, S.; and Kalemaj, I. 2020. Performative Prediction in a Stateful World. *CoRR*, abs/2011.03885.

Casado, F. E.; Lema, D.; Criado, M. F.; Iglesias, R.; Regueiro, C. V.; and Barro, S. 2022. Concept drift detection and adaptation for federated and continual learning. *Multimedia Tools and Applications*, 81(3): 3397–3419.

Chen, Y.; Wang, J.; and Liu, Y. 2020. Strategic Recourse in Linear Classification. *arXiv preprint arXiv:2011.00355*.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Ding, Y.; Niu, C.; Yan, Y.; Zheng, Z.; Wu, F.; Chen, G.; Tang, S.; and Jia, R. 2020. Distributed optimization over block-cyclic data. *arXiv preprint arXiv:2002.07454*.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Drusvyatskiy, D.; and Xiao, L. 2020. Stochastic optimization with decision-dependent distributions.

Drusvyatskiy, D.; and Xiao, L. 2022. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*.

Eichner, H.; Koren, T.; McMahan, B.; Srebro, N.; and Talwar, K. 2019. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning*, 1764–1773. PMLR.

Guo, Y.; Lin, T.; and Tang, X. 2021. Towards federated learning on time-evolving heterogeneous data. *arXiv preprint arXiv:2112.13246*.

Haddadpour, F.; Kamani, M. M.; Mokhtari, A.; and Mahdavi, M. 2021. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2350–2358. PMLR.

Haghtalab, N.; Immorlica, N.; Lucier, B.; and Wang, J. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. 160–166.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic Classification. 111–122.

Hosseinalipour, S.; Wang, S.; Michelusi, N.; Aggarwal, V.; Brinton, C. G.; Love, D. J.; and Chiang, M. 2022. Parallel successive learning for dynamic distributed model training over heterogeneous wireless networks. *arXiv preprint arXiv:2202.02947*.

Hu, L.; Immorlica, N.; and Vaughan, J. 2019. The Disparate Effects of Strategic Manipulation. 259–268.

Izzo, Z.; Ying, L.; and Zou, J. 2021. How to Learn when Data Reacts to Your Model: Performative Gradient Descent. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4641–4650. PMLR.

Izzo, Z.; Zou, J.; and Ying, L. 2022. How to Learn when Data Gradually Reacts to Your Model. In Camps-Valls, G.; Ruiz, F. J. R.; and Valera, I., eds., *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 3998–4035. PMLR.

Jin, K.; Zhang, X.; Khalili, M. M.; Naghizadeh, P.; and Liu, M. 2022. Incentive Mechanisms for Strategic Classification and Regression Problems. In Pennock, D. M.; Segal, I.; and Seuken, S., eds., *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, 760–790. ACM.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Kleinberg, J.; and Raghavan, M. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Transactions on Economics and Computation*, 8: 1–23.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, Q.; and Wai, H.-T. 2022. State Dependent Performative Prediction with Stochastic Approximation. In Camps-Valls, G.; Ruiz, F. J. R.; and Valera, I., eds., *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 3164–3186. PMLR.

Li, Q.; Yau, C.-Y.; and Wai, H.-T. 2022. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35: 38449–38460.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020b. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363.

Mendler-Dünner, C.; Perdomo, J.; Zrnic, T.; and Hardt, M. 2020. Stochastic Optimization for Performative Prediction. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 4929–4939. Curran Associates, Inc.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6917–6926. PMLR.

Miller, J. P.; Perdomo, J. C.; and Zrnic, T. 2021. Outside the Echo Chamber: Optimizing the Performative Risk. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7710–7720. PMLR.

Milli, S.; Miller, J.; Dragan, A.; and Hardt, M. 2019. The Social Cost of Strategic Classification. 230–239.

Narang, A.; Faulkner, E.; Drusvyatskiy, D.; Fazel, M.; and Ratliff, L. J. 2022. Multiplayer Performative Prediction: Learning in Decision-Dependent Games. *CoRR*, abs/2201.03398.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative Prediction. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7599–7609. PMLR.

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

Ray, M.; Ratliff, L. J.; Drusvyatskiy, D.; and Fazel, M. 2022. 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*, 8081–8088. AAAI Press.

Rizk, E.; Vlaski, S.; and Sayed, A. H. 2020. Dynamic Federated Learning. *CoRR*, abs/2002.08782.

Sahu, A. K.; Li, T.; Sanjabi, M.; Zaheer, M.; Talwalkar, A.; and Smith, V. 2018. On the Convergence of Federated Optimization in Heterogeneous Networks. *CoRR*, abs/1812.06127.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Causal Strategic Linear Regression. arXiv:2002.10066.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623.

Wood, K.; Bianchin, G.; and Dall'Anese, E. 2022. Online Projected Gradient Descent for Stochastic Optimization With Decision-Dependent Distributions. *IEEE Control Systems Letters*, 6: 1646–1651.

Yin, T.; Raab, R.; Liu, M.; and Liu, Y. 2023. Long-Term Fairness with Unknown Dynamics. *arXiv preprint arXiv:2304.09362*.

Zhang, X.; Khalili, M. M.; Jin, K.; Naghizadeh, P.; and Liu, M. 2022. Fairness Interventions as (Dis)Incentives for Strategic Manipulation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 26239–26264. PMLR.

Zhu, C.; Xu, Z.; Chen, M.; Konečnỳ, J.; Hard, A.; and Goldstein, T. 2021. Diurnal or Nocturnal? Federated Learning of Multi-branch Networks from Periodically Shifting Distributions. In *International Conference on Learning Representations*.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, 12878–12889. PMLR.

Zrnic, T.; Mazumdar, E.; Sastry, S.; and Jordan, M. 2021. Who Leads and Who Follows in Strategic Classification? In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 15257–15269. Curran Associates, Inc.