Federated Multi-Task Learning under a Mixture of Distributions

Othmane Marfoq [1 2] **Giovanni Neglia** [1] **Aurélien Bellet** [3] **Laetitia Kameni** [2] **Richard Vidal** [2]

Abstract

The increasing size of data generated by smartphones and IoT devices motivated the development of *Federated Learning* (FL), a framework for on-device collaborative training of machine learning models. First efforts in FL focused on learning a single global model with good average performance across clients, but the global model may be arbitrarily bad for a given client, due to the inherent heterogeneity of local data distributions. Federated *multi-task learning* (MTL) approaches can learn *personalized models* by formulating an opportune penalized optimization problem. The penalization term can capture complex relations among these models, but eschews clear statistical assumptions about local data distributions.

In this work, we propose to study federated MTL under the flexible assumption that each local data distribution is a *mixture of unknown underlying distributions*. This assumption encompasses most of the existing personalized FL approaches and leads to federated EM (*expectation maximization*) like algorithms for both client-server and fully decentralized settings. Moreover, it provides a principled way to serve personalized models to clients not seen at training time. The algorithms' convergence is analyzed through a novel federated surrogate optimization framework, which can be of general interest. Experimental results on FL benchmarks show that our approach provides models with higher accuracy and fairness than state-of-the-art methods.

[1]Inria, Université Côte d'Azur, Sophia Antipolis, France [2]Accenture Labs, Sophia Antipolis, France [3]Inria, Lille, France. Correspondence to: Othmane Marfoq <othmane.marfoq@inria.fr>.

This work was presented at the International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2021 (FL-ICML'21). This workshop does not have official proceedings and this paper is non-archival.

1. Introduction

Federated Learning (FL) (Kairouz et al., 2019) allows a set of clients to collaboratively train models without sharing their local data. Standard FL approaches train a unique model for all clients (McMahan et al., 2017; Konečny et al., 2016; Sahu et al., 2018; Karimireddy et al., 2020; Mohri et al., 2019). However, as discussed in (Sattler et al., 2020), the existence of such a global model suited for all clients is at odds with the statistical heterogeneity observed across different clients (Li et al., 2020; Kairouz et al., 2019). Indeed, clients can have non-iid data and *varying preferences*. Consider for example a language modeling task: given the sequence of tokens "*I love eating*," the next word can be arbitrarily different from one client to another. Thus, having personalized models for each client is a necessity in many FL applications.

Previous work on personalized FL. A naive approach for FL personalization is learning first a global model and then fine-tuning its parameters at each client via a few iterations of stochastic gradient descent (Sim et al., 2019). In this case, the global model plays the role of a meta-model to be used as initialization for few-shot adaptation at each client. In particular, the connection between FL and Model Agnostic Meta Learning (MAML) (Jiang et al., 2019) has been studied in (Fallah et al., 2020) and (Khodak et al., 2019) in order to build a more suitable meta-model for local personalization. Unfortunately, these methods can fail to build a model with low generalization error (as exemplified by LEAF synthetic dataset (Caldas et al., 2018, App. 1)). An alternative approach is to jointly train a global model and one local model per client and then let each client build a personalized model by interpolating them (Deng et al., 2020; Corinzia & Buhmann, 2019; Mansour et al., 2020). However, if local distributions are far from the average distribution, a relevant global model does not exist and this approach boils down to every client learning only on its own local data. This issue is formally captured by the generalization bound in (Deng et al., 2020, Theorem 1).

Clustered FL (Sattler et al., 2020; Ghosh et al., 2020b; Mansour et al., 2020) addresses the potential lack of a global model by assuming that clients can be partitioned into several clusters. Clients belonging to the same cluster share the same optimal model, but those models can be arbitrarily

different across clusters (see (Sattler et al., 2020, Assumption 2) for a rigorous formulation). During training, clients learn the cluster to which they belong as well as the cluster model. The clustered FL assumption is also quite limiting, as no knowledge transfer is possible across clusters. In the extreme case where each client has its own optimal local model (recall the example on language modeling), the number of clusters coincides with the number of clients and no federated learning is possible.

Multi-Task Learning (MTL) has recently emerged as an alternative approach to learn personalized models in the federated setting and allows for more nuanced relations among clients' models (Smith et al., 2017; Vanhaesebrouck et al., 2017; Zantedeschi et al., 2020; Hanzely & Richtárik, 2020; Dinh et al., 2020). Smith et al. (2017) and Vanhaesebrouck et al. (2017) were the first to frame FL personalization as a MTL problem. In particular, they defined federated MTL as a penalized optimization problem, where the penalization term models relationships among tasks (clients). Smith et al. (2017) proposed the MOCHA algorithm for the client-server scenario, while Vanhaesebrouck et al. (2017) and Zantedeschi et al. (2020) presented decentralized algorithms for the same problem. Unfortunately, these algorithms can only learn simple models (linear models or linear combination of pre-trained models), because of the complex penalization term. Other MTL-based approaches (Hanzely & Richtárik, 2020; Hanzely et al., 2020; Dinh et al., 2020) are able to train more general models at the cost of considering simpler penalization terms (e.g., the distance to the average model), thereby losing the capability to capture complex relations among tasks. Moreover, a general limitation of this line of work is that the penalization term is justified qualitatively and not on the basis of clear statistical assumptions on local data distributions.

Overall, although current personalization approaches can lead to superior empirical performance in comparison to a shared global model or individually trained local models, it is still not well understood whether and under which conditions clients are guaranteed to benefit from collaboration.

Our contributions. In this work, we first show that federated learning is impossible without assumptions on local data distributions. Motivated by this negative result, we formulate a general and flexible assumption: *the data distribution of each client is a mixture of M underlying distributions.* The proposed formulation has the advantage that each client can benefit from knowledge distilled from all other clients' datasets (even if any two clients can be arbitrarily different from each other). We also show that this assumption encompasses most of the personalized FL approaches proposed in the literature.

In our framework, a personalized model is a linear combination of M shared component models. All clients jointly learn the M components, while each client learns its personalized mixture weights. We show that federated EM-like algorithms can be used for training. In particular, we propose `FedEM` and `D-FedEM` for the client-server and the fully decentralized settings, respectively, and we prove convergence guarantees. Our approach also provides a principled and efficient way to infer personalized models for clients unseen at training time. Our algorithms can easily be adapted to solve more general problems in a novel framework, which can be seen as a federated extension of the centralized surrogate optimization approach in (Mairal, 2013). To the best of our knowledge, our paper is the first work to propose federated surrogate optimization algorithms with convergence guarantees.

Through extensive experiments on FL benchmark datasets, we show that our approach generally yields models that 1) are on average more accurate, 2) are fairer across clients, and 3) generalize better to unseen clients than state-of-the-art personalized and non-personalized FL approaches.

Paper outline. In Sec. 2 we provide our impossibility result, introduce our main assumptions, and show that several popular personalization approaches can be obtained as special cases of our framework. Section 3 describes `FedEM` and states its convergence results. Finally, we provide experimental results in Sec. 4 before concluding in Sec. 5.

2. Problem Formulation

We consider a (countable) set $\mathcal{T}$ of classification (or regression) tasks which represent the set of possible clients. We will use the terms task and client interchangeably. Data at client $t \in \mathcal{T}$ is generated according to a local distribution $\mathcal{D}_t$ over $\mathcal{X} \times \mathcal{Y}$. Local data distributions $\{\mathcal{D}_t\}_{t\in\mathcal{T}}$ are in general different, thus it is natural to fit a separate model (hypothesis) $h_t \in \mathcal{H}$ to each data distribution $\mathcal{D}_t$. The goal is thus to solve (in parallel) the following optimization problems

$$\forall t \in \mathcal{T}, \quad \underset{h_t\in\mathcal{H}}{\text{minimize}}\, \mathcal{L}_{\mathcal{D}_t}(h_t), \tag{1}$$

where $h_t : \mathcal{X} \mapsto \Delta^{|\mathcal{Y}|}$ (Δ^D denoting the unitary simplex of dimension D), $l : \Delta^{|\mathcal{Y}|} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a loss function,[1] and $\mathcal{L}_{\mathcal{D}_t}(h_t) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[l(h_t(\mathbf{x}), y)\right]$ is the true risk of a model h_t under data distribution $\mathcal{D}_t$. For $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, we will denote the joint distribution density associated to $\mathcal{D}_t$ by $p_t(\mathbf{x}, y)$, and the marginal densities by $p_t(\mathbf{x})$ and $p_t(y)$.

A set of T clients $[T] \triangleq \{1, 2, \dots, T\} \subseteq \mathcal{T}$ participate to the initial training phase; other clients may join the system in a later stage. We denote by $\mathcal{S}_t = \{s_t^{(i)} \triangleq (\mathbf{x}_t^{(i)}, y_t^{(i)})\}_{i=1}^{n_t}$ the dataset at client $t \in [T]$ drawn i.i.d. from $\mathcal{D}_t$, and by $n = \sum_{t=1}^{T} n_t$ the total dataset size.

[1] In the case of (multi-output) regression, we have $h_t : \mathcal{X} \mapsto \mathbb{R}^d$ for some $d \geq 1$ and $l : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^+$.

The idea of federated learning is to enable each client to benefit from data samples available at other clients in order to get a better estimation of $\mathcal{L}_{\mathcal{D}_t}$, and therefore get a model with a better generalization ability to unseen examples.

2.1. Learning under a Mixture Model

As shown in Appendix A, without any assumptions on the local data distributions $p_t(\mathbf{x}, y)$, a client cannot provably benefit from larger amounts of data available at other clients. Motivated by this impossibility result, in this work we propose to consider that each local data distribution $\mathcal{D}_t$ is a mixture of M underlying distributions $\tilde{\mathcal{D}}_m$, $1 \leq m \leq M$, as formalized below.

Assumption 1. *There exist M underlying (independent) distributions $\tilde{\mathcal{D}}_m$, $1 \leq m \leq M$, such that for $t \in \mathcal{T}$, $\mathcal{D}_t$ is mixture of the distributions $\{\tilde{\mathcal{D}}_m\}_{m=1}^M$ with weights $\pi_t^* = [\pi_{t1}^*, \dots, \pi_{tM}^*] \in \Delta^M$, i.e.*

$$z_t \sim \mathcal{M}(\pi_t^*), \quad ((\mathbf{x}_t, y_t) \,|\, z_t = m) \sim \tilde{\mathcal{D}}_m, \quad \forall t \in \mathcal{T}, \quad (2)$$

where $\mathcal{M}(\pi)$ is a multinomial (categorical) distribution with parameters π.

Similarly to what was done above, we use $p_m(\mathbf{x}, y)$, $p_m(\mathbf{x})$, and $p_m(y)$ to denote the probability distribution densities associated to $\tilde{\mathcal{D}}_m$. We further assume that marginals over $\mathcal{X}$ are identical.

Assumption 2. *For all $m \in [M]$, we have $p_m(\mathbf{x}) = p(\mathbf{x})$.*

Assumption 2 is not strictly required for our analysis to hold, but, in the most general case, solving Problem (1) requires to learn generative models. Instead, under Assumption 2 we can restrict our attention to discriminative models (e.g., neural networks). More specifically, we consider a parameterized set of models $\tilde{\mathcal{H}}$ with the following properties.

Assumption 3. *$\tilde{\mathcal{H}} = \{h_\theta\}_{\theta \in \mathbb{R}^d}$ is a set of hypotheses parameterized by $\theta \in \mathbb{R}^d$, whose convex hull is in $\mathcal{H}$. For each distribution $\tilde{\mathcal{D}}_m$ with $m \in [M]$, there exists a hypothesis $h_{\theta_m^*}$, such that*

$$l\left(h_{\theta_m^*}(\mathbf{x}), y\right) = -\log p_m(y|\mathbf{x}) + c, \quad (3)$$

where $c \in \mathbb{R}$ is a normalization constant. $l(\cdot, \cdot)$ is then the log loss associated to $p_m(y|\mathbf{x})$.

We refer to the hypotheses in $\tilde{\mathcal{H}}$ as *component models* or simply *components*. We denote by $\Theta^* \in \mathbb{R}^{M \times d}$ the matrix whose m-th row is θ_m^*, and by $\Pi^* \in \Delta^{T \times M}$ the matrix whose t-th row is $\pi_t^* \in \Delta^M$. Similarly, we will use Θ and Π to denote arbitrary parameters.

Remark 1. *Assumptions 2–3 are mainly technical and are not required for our approach to work in practice. Experiments in Sec. 4 show that our algorithms perform well on standard FL benchmark datasets, for which these assumptions do not hold in general.*

Note that, under the above assumptions, $p_t(\mathbf{x}, y)$ depends on Θ^* and π_t^*. Moreover, we can prove (see App. B) that the optimal local model $h_t^* \in \mathcal{H}$ for client t is a weighted average of models in $\tilde{\mathcal{H}}$.

Proposition 2.1. *Let $l(\cdot, \cdot)$ be the mean squared error loss, the logistic loss or the cross-entropy loss, and $\breve{\Theta}$ and $\breve{\Pi}$ be a solution of the following optimization problem:*

$$\underset{\Theta, \Pi}{\text{minimize}} \; \mathop{\mathbb{E}}_{t \sim D_{\mathcal{T}}} \mathop{\mathbb{E}}_{(\mathbf{x}, y) \sim \mathcal{D}_t} \left[-\log p_t(\mathbf{x}, y | \Theta, \pi_t) \right], \quad (4)$$

where $D_{\mathcal{T}}$ is any distribution with support $\mathcal{T}$. Under Assumptions 1, 2, and 3, the predictors

$$h_t^* = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}), \quad \forall t \in \mathcal{T} \quad (5)$$

solve Problem (1).

Proposition 2.1 suggests the following approach to solve Problem (1). First, we estimate the parameters $\breve{\Theta}$ and $\breve{\pi}_t$, $1 \leq t \leq T$, by minimizing the empirical version of Problem (4) on the training data:

$$f(\Theta, \Pi) \triangleq -\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \log p(s_t^{(i)} | \Theta, \pi_t), \quad (6)$$

which is the (negative) likelihood of the probabilistic model (2).[2] Second, we use (5) to get the client predictor for the T clients present at training time. Finally, to deal with a client $t_{\text{new}} \notin [T]$ not seen during training, we keep the mixture component models fixed and simply choose the weights $\pi_{t_{\text{new}}}$ that maximize the likelihood of the client data and build the client predictor via (5).

2.2. Generalizing Existing Frameworks

Before presenting our FL algorithms in Sec. 3, we show that the generative model in Assumption 1 extends some popular multi-task/personalized FL formulations in the literature.

Clustered Federated Learning (Sattler et al., 2020; Ghosh et al., 2020a) assumes that each client belongs to one among C clusters and proposes that all clients in the same cluster learn the same model. Our framework recovers this scenario considering $M = C$ and $\pi_{tc}^* = 1$ if task (client) t is in cluster c and $\pi_{tc}^* = 0$ otherwise.

Personalization via model interpolation (Mansour et al., 2020; Deng et al., 2020) relies on learning a global model h_{glob} and T local models $h_{\text{loc},t}$, and then using at each client the linear interpolation $h_t = \alpha_t h_{\text{loc},t} + (1 - \alpha_t) h_{\text{glob}}$. Each client model can thus be seen as a linear combination of $M = T + 1$ models $h_m = h_{\text{loc},m}$ for $m \in [T]$ and $h_0 =$

[2] As the distribution $\mathcal{D}_{\mathcal{T}}$ in Prop. 2.1 is arbitrary, any positively weighted sum of clients' empirical losses could be considered.

h_{glob} with specific weights $\pi^*_{tt} = \alpha_t$, $\pi^*_{t0} = 1 - \alpha_t$, and $\pi^*_{tt'} = 0$ for $t' \in [T] \setminus \{t\}$.

Federated MTL via task relationships. The authors of (Smith et al., 2017) proposed to learn personalized client models by solving the following optimization problem inspired from classic MTL formulations:

$$\min_{W,\Omega} \sum_{t=1}^{T} \sum_{i=1}^{n_t} l(h_{w_t}(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda \operatorname{tr}(W\Omega W^\intercal), \quad (7)$$

where h_{w_t} are linear predictors parameterized by the rows of matrix W and the matrix Ω captures task relationships (similarity). This formulation is motivated by the alternating structure optimization method (ASO) (Ando & Zhang, 2005; Zhou et al., 2011). In App. C, we show that, when predictors $h_{\theta^*_m}$ are linear and have bounded norm, our framework leads to the same ASO formulation that motivated Problem (7). Problem (7) can also be justified by probabilistic priors (Zhang & Yeung, 2010) or graphical models (Lauritzen, 1996) (see (Smith et al., 2017, App. B.1)). Similar considerations hold for our framework (see again App. C). Reference (Zantedeschi et al., 2020) extends the approach in (Smith et al., 2017) by letting each client learn a personalized model as a weighted combination of M *known* hypotheses. Our approach is more general and flexible as clients learn both the weights and the hypotheses. Finally, other personalized FL algorithms, like `pFedMe` (Dinh et al., 2020), `FedU` (Dinh et al., 2021), and those studied in (Hanzely & Richtárik, 2020) and in (Hanzely et al., 2020), can be framed as special cases of formulation (7). Their assumptions can thus also be seen as a particular case of our framework.

3. Federated Expectation-Maximization

3.1. Centralized Expectation-Maximization

Our goal is to estimate the optimal components' parameters $\Theta^* = (\theta^*_m)_{1 \le m \le M}$ and mixture weights $\Pi^* = (\pi^*_t)_{1 \le t \le T}$ by minimizing the negative log-likelihood $f(\Theta, \Pi)$ in (6). A natural approach to solve such non-convex problems is via the Expectation-Maximization algorithm (EM), which alternates between two steps. Expectation steps update the distribution (denoted by q_t) over the latent variables $z_t^{(i)}$ for every instance $s_t^{(i)}$, given the current estimates of the parameters $\{\Theta, \Pi\}$. Maximization steps update the parameters $\{\Theta, \Pi\}$ by maximizing the expected log-likelihood, where the expectation is computed according to the current latent variables' distributions. The following proposition provides the EM updates for our problem (proof in App. D).

Proposition 3.1. *Under Assumptions 1 and 2, at the k-th iteration the EM algorithm updates parameter estimates through the following steps:*

$$\textbf{\textit{E-step:}}\ q_t^{k+1}(z_t^{(i)} = m) \propto \pi_{tm}^k e^{-l(h_{\theta_m^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})}, \quad (8)$$

$$\textbf{\textit{M-step:}}\ \pi_{tm}^{k+1} = \frac{\sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m)}{n_t}, \quad (9)$$

$$\theta_m^{k+1} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m) \times l\big(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)}\big), \quad (10)$$

The EM updates in Proposition 3.1 have a natural interpretation. In the E-step, given current component models Θ^k and mixture weights Π^k, (8) updates the a-posteriori probability $q_t^{k+1}(z_t^{(i)} = m)$ that point $s_t^{(i)}$ of client t was drawn from the m-th distribution based on the current mixture weight π_{tm}^k and on how well the corresponding component θ_m^k classifies $s_t^{(i)}$. The M-step consists of two updates under fixed probabilities q_t^{k+1}. First, (9) updates the mixture weights π_t^{k+1} to reflect the prominence of each distribution $\tilde{\mathcal{D}}_m$ in $\mathcal{S}_t$ as given by q_t^{k+1}. Finally, (10) updates the components' parameters Θ^{k+1} by solving M independent, weighted empirical risk minimization problems with weights given by q_t^{k+1}. These weights aim to construct an unbiased estimate of the true risk over each underlying distribution $\tilde{\mathcal{D}}_m$ using only points sampled from the client mixtures, similarly to importance sampling strategies used to learn from data with sample selection bias (Sugiyama et al., 2008; Cortes et al., 2008; 2010; Vogel et al., 2020).

3.2. Federated Expectation-Maximization

Federated learning aims to train machine learning models directly on the clients, without exchanging raw data, and thus we should run EM while assuming that only client t has access to dataset $\mathcal{S}_t$. The E-step (8) and the Π update (9) in the M-step operate separately on each local dataset $\mathcal{S}_t$ and can thus be performed locally at each client t. On the contrary, the Θ update (10) requires interaction with other clients, since the computation spans all data samples $\mathcal{S}_{1:T}$.

In this section, we consider a client-server setting, in which each client t can communicate only with a centralized server (the orchestrator) and wants to learn components' parameters $\Theta^* = (\theta^*_m)_{1 \le m \le M}$ and its own mixture weights π^*_t.

We propose the algorithm `FedEM` for *Federated Expectation-Maximization* (Alg. 1). `FedEM` proceeds through communication rounds similarly to most FL algorithms including `FedAvg` (McMahan et al., 2017), `FedProx` (Sahu et al., 2018), `SCAFFOLD` (Karimireddy et al., 2020), and `pFedMe` (Dinh et al., 2020). At each round, 1) the central server broadcasts the (shared) component models to the clients, 2) each client locally updates components and its personalized mixture weights, and 3) sends the updated components back to the server, 4) the server aggregates the updates. The local update performed at client t consists in performing the steps in (8) and (9) and updating the local estimates of θ_m through a

Algorithm 1 `FedEM` (see also the more detailed Alg. 7 in App. I.1)

1: **Input:** data $\mathcal{S}_{1:T}$; number of mixture distributions M; number of communication rounds K
2: **for** iterations $k = 1, \dots, K$ **do**
3: server broadcast $\theta_m^{k-1},\ 1 \leq m \leq M$ to the T clients
4: **for** tasks $t = 1, \dots, T$ in parallel over T clients **do**
5: **for** component $m = 1, \dots, M$ **do**
6: update $q_t^k(z_t^{(i)} = m)$ as in (8), $\forall i \in \{1, \dots, n_t\}$
7: update π_{tm}^k as in (9)
8: $\theta_{m,t}^k \leftarrow \texttt{LocalSolver}(m,\ \theta_m^{k-1}, q_t^k, \mathcal{S}_t)$
9: **end for**
10: **end for**
11: client t sends $\theta_{m,t}^k,\ 1 \leq m \leq M$, to the server
12: **for** component $m = 1, \dots, M$ **do**
13: $\theta_m^k \leftarrow \sum_{t=1}^T \frac{n_t}{n} \times \theta_{m,t}^k$
14: **end for**
15: **end for**

solver which approximates the exact minimization in (10) using only the local dataset $\mathcal{S}_t$ (see line 8). `FedEM` can operate with different local solvers—even different across clients—as far as they satisfy some local improvement guarantees (see the discussion in App. K). In what follows, we restrict our focus on the practically important case where the local solver performs multiple stochastic gradient descent updates (local SGD (Stich, 2018)).

Under standard mild assumptions (detailed in Appendix E), `FedEM` converges to a stationary point of f.

Theorem 3.2. *Under Assumptions 1–7, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, `FedEM`'s iterates satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left\|\nabla_\Theta f\left(\Theta^k, \Pi^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \tag{11}$$

$$\frac{1}{K}\sum_{k=1}^{K} \Delta_\Pi f(\Theta^k, \Pi^k) \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right), \tag{12}$$

where the expectation is over the random batches samples, and $\Delta_\Pi f(\Theta^k, \Pi^k) \triangleq f\left(\Theta^k, \Pi^k\right) - f\left(\Theta^k, \Pi^{k+1}\right) \geq 0$.

Theorem 3.2 (proof in App. J.1) expresses the convergence of both sets of parameters (Θ and Π) to a stationary point of f. Indeed, the gradient of f with respect to Θ becomes arbitrarily small (inequality (11)) and the update in Eq. (9) leads to arbitrarily small improvements of f (inequality (12)).

`FedEM` allows an *unseen client*, i.e., a client $t_{\text{new}} \notin [T]$ arriving after the distributed training procedure, to learn its personalized model. The client simply retrieves the learned components' parameters Θ^K and computes its personalized weights $\pi_{t_{\text{new}}}$ through one E-step (8) (e.g., starting from a uniform initialization) and the first update in the M-step (9).

In some cases, clients may want to communicate directly in a peer-to-peer fashion instead of relying on the central server mediation (Kairouz et al., 2019, Sec. 2.1). In fact, fully decentralized schemes may provide stronger privacy guarantees (Cyffers & Bellet, 2021) and speed-up training as they better use communication resources (Lian et al., 2017; Marfoq et al., 2020) and reduce the effect of stragglers (Neglia et al., 2019). In App. I.2 we propose `D-FedEM` (Alg. 3), a *fully decentralized version* of our federated EM algorithm with similar convergence guarantees.

Both `FedEM` and `D-FedEM` can be seen as particular instances of a more general framework—of potential interest for other applications—that we call *federated surrogate optimization* and we describe in App. H.

4. Experiments

Datasets and models. We evaluated our method on five federated benchmark datasets spanning a wide range of machine learning tasks: image classification (CIFAR10 and CIFAR100 (Krizhevsky, 2009)), handwritten character recognition (EMNIST (Cohen et al., 2017) and FEMNIST (Caldas et al., 2018)), and language modeling (Shakespeare (Caldas et al., 2018; McMahan et al., 2017)). Shakespeare dataset (resp. FEMNIST) was naturally partitioned by assigning all lines from the same characters (resp. all images from the same writer) to the same client. We created federated versions of CIFAR10 and EMNIST by distributing samples with the same label across the clients according to a symmetric Dirichlet distribution with parameter 0.4, as in (Wang et al., 2020a). For CIFAR100, we exploited the availability of "coarse" and "fine" labels, using a two-stage Pachinko allocation method (Li & McCallum, 2006) to assign 600 sample to each of the 100 clients, as in (Reddi et al., 2021). We also evaluated our method on a synthetic dataset verifying Assumptions 1–3. For all tasks, we randomly split each local dataset into training (80%) and test (20%) sets. Details on datasets and models can be found in App. L.1, and additional experimental results are in App. M.

Other FL approaches. We compared our algorithms with global models trained with `FedAvg` (McMahan et al., 2017) and `FedProx` (Sahu et al., 2018) and different personalization approaches: a personalized model trained only on the local dataset, `FedAvg` with local tuning (`FedAvg+`) (Jiang et al., 2019), `clustered FL` (Sattler et al., 2020) and `pFedMe` (Dinh et al., 2020). For each method and each task, the learning rate and the other hyper-parameters were tuned via grid search (details in App. L.2). `FedAvg+` updated the local model through a single pass on the local

Table 1. Test accuracy: average across clients / bottom decile.

DATASET	LOCAL	FEDAVG	FEDPROX	FEDAVG+	CLUSTERED FL	PFEDME	FEDEM (OURS)
FEMNIST	71.0 / 57.5	78.6 / 63.9	78.9 / 64.0	75.3 / 53.0	73.5 / 55.1	74.9 / 57.6	**79.9** / **64.8**
EMNIST	71.9 / 64.3	82.6 / 75.0	83.0 / 75.4	83.1 / 75.8	82.7 / 75.0	83.3 / 76.4	**83.5** / **76.6**
CIFAR10	70.2 / 48.7	78.2 / 72.4	78.0 / 70.8	82.3 / 70.6	78.6 / 71.2	81.7 / 73.6	**84.3** / **78.1**
CIFAR100	31.5 / 19.9	40.9 / 33.2	41.0 / 33.2	39.0 / 28.3	41.5 / 34.1	41.8 / 32.5	**44.1** / **35.0**
SHAKESPEARE	32.0 / 16.6	**46.7** / 42.8	45.7 / 41.9	40.0 / 25.5	46.6 / 42.7	41.2 / 36.8	**46.7** / **43.0**
SYNTHETIC	65.7 / 58.4	68.2 / 58.9	68.2 / 59.0	68.9 / 60.2	69.1 / 59.0	69.2 / 61.2	**74.7** / **66.7**

dataset. Unless otherwise stated, the number of components considered by FedEM was $M = 3$, [3] training occurred over 80 communication rounds for Shakespeare and 200 rounds for all other datasets. At each round, clients train for one epoch. Results for D-FedEM are in Appendix M.1.

Average performance of personalized models. The performance of each personalized model (which is the same for all clients in the case of FedAvg and FedProx) is evaluated on the local test dataset (unseen at training). Table 1 shows the average weighted accuracy with weights proportional to local dataset sizes. We observe that FedEM obtains the best performance across all datasets.

Fairness across clients. FedEM's improvement in terms of average accuracy could be the result of learning particularly good models for some clients at the expense of bad models for other clients. Table 1 shows the bottom decile of the accuracy of local models, i.e., the $(T/10)$-th worst accuracy (the minimum accuracy is particularly noisy, notably because some local test datasets are very small). Even clients with the worst personalized models are still better off when FedEM is used for training.

Clients sampling. In cross-device federated learning, only a subset of clients may be available at each round. We ran CIFAR10 experiments with different levels of participation: at each round a given fraction of all clients were sampled uniformly without replacement. We restrict the comparison to FedEM and FedAvg+, as 1) FedAvg+ performed better than FedProx and FedAvg in the previous CIFAR10 experiments, 2) it is not clear how to extend pFedMe and clustered FL to handle client sampling. Results in Fig. 1 (App. M) show that FedEM is more robust to low clients' participation levels.

Generalization to unseen clients. As discussed in Sec. 3.2, FedEM allows new clients arriving after the distributed training to easily learn their personalized models. With the exception of FedAvg+, it is not clear if and how the other personalized FL algorithms can tackle the same goal. In order to evaluate the quality of new clients' personalized models, we performed an experiment where only 80% of the clients ("old" clients) participate to the training. The remaining 20% join the system in a second phase and use their local training datasets to learn their personalized weights. Table 4 shows that FedEM allows new clients to learn a personalized model at least as good as FedAvg's global one and always better than FedAvg+'s one. Unexpectedly, new clients achieve sometimes a significantly higher test accuracy than old clients (e.g., 47.5% against 44.1% on CIFAR100). Our investigation (App. M.2) suggests that, by selecting their mixture weights on local datasets that were not used to train the components, new clients can compensate for potential overfitting in the initial training phase.

[3] The effect of the number of components M on the test accuracy is explored in Appendix M.3 and Appendix M.4

Table 2. Average test accuracy across clients unseen at training (train accuracy in parenthesis).

DATASET	FEDAVG	FEDAVG+	FEDEM
FEMNIST	78.3 (80.9)	74.2 (84.2)	**79.1** (81.5)
EMNIST	83.4 (82.7)	83.7 (92.9)	**84.0** (83.3)
CIFAR10	77.3 (77.5)	80.4 (80.5)	**85.9** (90.7)
CIFAR100	41.1 (42.1)	36.5 (55.3)	**47.5** (46.6)
SHAKESPEARE	**46.7** (47.1)	40.2 (93.0)	**46.7** (46.6)
SYNTHETIC	68.6 (70.0)	69.1 (72.1)	**73.0** (74.1)

5. Conclusion

In this paper, we proposed a novel federated MTL approach based on the flexible assumption that local data distributions are mixtures of underlying distributions. Our EM-like algorithms allow clients to jointly learn shared component models and personalized mixture weights in client-server and fully decentralized settings. We proved convergence guarantees for our algorithms through a federated surrogate optimization framework which can be used to analyze other FL formulations. In practice, our approach learns models with higher accuracy and fairness than state-of-the-art FL algorithms, even for clients not present at training time.

In future work, we aim to reduce local computation and communication of our algorithms. Aside from standard compression schemes (Haddadpour et al., 2021), a promis-

ing direction is to limit the number of component models that a client updates/transmits at each step. This could be done in an adaptive manner based on the client's current mixture weights. A second interesting direction is to study personalized FL approaches under privacy constraints (quite unexplored until now with the notable exception of Bellet et al. (2018)). Some features of our algorithms may be beneficial for privacy (e.g., the fact that personalized weights are kept locally and that all users contribute to all shared models). We hope to design differentially private versions of our algorithms and characterize their privacy-utility trade-offs.

References

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(61):1817–1853, 2005.

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. Personalized and Private Peer-to-Peer Machine Learning. In *AISTATS*, 2018.

Ben-David, S., Lu, T., and Pál, D. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, 2008.

Boyd, S., Diaconis, P., and Xiao, L. Fastest mixing markov chain on a graph. *SIAM REVIEW*, 46:667–689, 2003.

Bubeck, S. Convex optimization: Algorithms and complexity, 2015.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Corinzia, L. and Buhmann, J. M. Variational federated multi-task learning, 2019.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. In *ALT*, 2008.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/59c33016884a62116be975a9bb8257e3-Paper.pdf.

Cyffers, E. and Bellet, A. Privacy amplification by decentralization, 2021.

Darnstädt, M., Simon, H. U., and Szörényi, B. Unlabeled data does provably help. In *STACS*, 2013.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Dinh, C. T., Tran, N. H., and Nguyen, T. D. Personalized federated learning with moreau envelopes. *arXiv preprint arXiv:2006.08848*, 2020.

Dinh, C. T., Vu, T. T., Tran, N. H., Dao, M. N., and Zhang, H. Fedu: A unified framework for federated multi-task learning with laplacian regularization. *arXiv preprint arXiv:2102.07148*, 2021.

Erdös, P. and Rényi, A. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. In *NeurIPS*, 2020a.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *arXiv preprint arXiv:2006.04088*, 2020b.

Göpfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I., and Urner, R. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pp. 1500–1518. PMLR, 2019.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Unified analysis and sharp guarantees. In *ICML*, 2021.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. 2020.

Hanzely, F., Hanzely, S., Horváth, S., and Richtárik, P. Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint arXiv:2010.02372*, 2020.

Jiang, Y., Konečnỳ, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

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

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pp. 5917–5928, 2019.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *ICML*, 2020.

Konečny, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Lauritzen, S. *Graphical models*. Number 17 in Oxford Statistical Science Series. Clarendon Press, 1996. ISBN 0198522193.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Li, W. and McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 577–584, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143917. URL https://doi.org/10.1145/1143844.1143917.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5336–5346, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous Decentralized Parallel Stochastic Gradient Descent. In *ICML*, 2018.

Mairal, J. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pp. 783–791, 2013.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL https://doi.org/10.1145/1873951.1874254.

Marfoq, O., Xu, C., Neglia, G., and Vidal, R. Throughput-optimal topology design for cross-silo federated learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19478–19487. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e29b722e35040b88678e25a1ec032a21-Paper.pdf.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625, 2019.

Nedić, A., Olshevsky, A., and Rabbat, M. G. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106 (5):953–976, 2018. doi: 10.1109/JPROC.2018.2817461.

Neglia, G., Calbi, G., Towsley, D., and Vardoyan, G. The role of network topology for distributed machine learning. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 2350–2358, 2019. doi: 10.1109/INFOCOM.2019.8737602.

Neglia, G., Xu, C., Towsley, D., and Calbi, G. Decentralized gradient methods: does topology matter? In *AISTATS*, 2020.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 1 edition, 2003. ISBN 1402075537,9781402075537. URL http://gen.lib.rus.ec/book/index.php?md5=488d3c36f629a6e021fc011675df02ef.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc,

F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A. S., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *ArXiv*, abs/1812.06127, 2018.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Sim, K. C., Zadrazil, P., and Beaufays, F. An investigation into on-device personalization of end-to-end automatic speech recognition models. *arXiv preprint arXiv:1909.06678*, 2019.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4427–4437, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Stich, S. U. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D^2: Decentralized Training over Decentralized Data. In *ICML*, 2018.

Vanhaesebrouck, P., Bellet, A., and Tommasi, M. Decentralized Collaborative Learning of Personalized Models over Networks. In *AISTATS*, 2017.

Vogel, R., Achab, M., Clémençon, S., and Tillier, C. Weighted emprirical risk minimization: Transfer learning based on importance sampling. In *ESANN*, 2020.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BkluqlSFDS.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020b.

Zantedeschi, V., Bellet, A., and Tommasi, M. Fully decentralized joint learning of personalized models and collaboration graphs. volume 108 of *Proceedings of Machine Learning Research*, pp. 864–874, Online, 8 2020. PMLR. URL http://proceedings.mlr.press/v108/zantedeschi20a.html.

Zhang, Y. and Yeung, D. Y. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pp. 733, 2010.

Zhou, J., Chen, J., and Ye, J. Clustered multi-task learning via alternating structure optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/a516a87cfcaef229b342c437fe2b95f7-Paper.pdf.

A. An Impossibility Result

We start by showing that some assumptions on the local distributions $p_t(\mathbf{x}, y),\ t \in \mathcal{T}$ are needed for federated learning to be possible, i.e., for each client to be able to take advantage of the data at other clients. This holds even if all clients are observed during the initial training phase (i.e., $\mathcal{T} = [T]$).

Our argument relies on a reduction to an impossibility result for semi-supervised learning (SSL). If clients have arbitrarily different label distributions, the information carried by $p_{t'}(y|\mathbf{x}),\ t' \in [T] \setminus \{t\}$ is not relevant for client t, and client t can only use the information carried by the marginals $p_{t'}(\mathbf{x})$. Assuming that these marginals are identical for all clients, federated learning with T clients is then equivalent to T SSL problems, where the SSL problem associated with client t relies on labeled samples in $\mathcal{S}_t$ and unlabeled samples in $\mathcal{U}_t = \cup_{t' \in [T]\setminus\{t\}} \{\mathbf{x} : (\mathbf{x}, y) \in \mathcal{S}_{t'}\}$. [4]

The authors of (Ben-David et al., 2008) conjectured that even when the quantity of unlabeled data goes to infinity, the worst-case sample complexity of SSL improves over supervised learning at most by a constant factor that only depends on the hypothesis class (Ben-David et al., 2008, Conjecture 4). Later work has shown the conjecture to hold for the realizable case and hypothesis classes of finite VC dimension (Darnstädt et al., 2013, Theorem 1), even when the marginal distribution is known (Göpfert et al., 2019, Theorem 2) (whether the conjecture in (Ben-David et al., 2008) holds in the agnostic case is still an open problem). The main consequence for FL is that, without further assumptions, a client cannot provably benefit from larger amounts of data available at other clients.

B. Proof of Proposition 2.1

Proposition 2.1. *Let $l(\cdot,\cdot)$ be the mean squared error loss, the logistic loss or the cross-entropy loss, and $\breve{\Theta}$ and $\breve{\Pi}$ be a solution of the following optimization problem:*

$$\underset{\Theta,\Pi}{\text{minimize}} \underset{t\sim D_{\mathcal{T}}}{\mathbb{E}} \underset{(\mathbf{x},y)\sim\mathcal{D}_t}{\mathbb{E}} \left[-\log p_t(\mathbf{x}, y|\Theta, \pi_t)\right], \tag{4}$$

where $D_{\mathcal{T}}$ is any distribution with support $\mathcal{T}$. Under Assumptions 1, 2, and 3, the predictors

$$h_t^* = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}, \quad \forall t \in \mathcal{T} \tag{5}$$

minimize $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[l(h_t(\mathbf{x}), y)\right]$ and thus solve Problem (1).

First, Lemma B.1 shows that the parameters $\breve{\Theta}$ and $\breve{\Pi}$, solving Problem (4), generate the same probability distribution as the parameters Θ^* and Π^* defined in Assumptions 1 and 3. Then, Lemmas B.2–B.4 exploit the particular form of the mean squared error loss, the logistic loss, and the cross entropy loss to prove that predictors $h_t^*,\ \forall t \in \mathcal{T}$, defined in (5), minimize $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[l(h_t(\mathbf{x}), y)\right]$.

Lemma B.1. *Suppose that Assumptions 1 and 3 hold, and consider $\breve{\Theta}$ and $\breve{\Pi}$ to be a solution of Problem (4). Then*

$$p_t(\mathbf{x}, y|\breve{\Theta}, \breve{\pi}_t) = p_t(\mathbf{x}, y|\Theta^*, \pi_t^*),\ \forall t \in \mathcal{T}. \tag{13}$$

Proof. For $t \in \mathcal{T}$,

$$\begin{aligned}
\underset{(\mathbf{x},y)\sim\mathcal{D}_t}{\mathbb{E}} \left[-\log p_t(\mathbf{x}, y|\breve{\Theta}, \breve{\pi}_t)\right] &= -\int_{\mathbf{x},y\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x}, y|\Theta^*, \pi_t^*) \cdot \log p_t(\mathbf{x}, y|\breve{\Theta}, \breve{\pi}_t) d\mathbf{x}dy && (14)\\
&= -\int_{\mathbf{x},y\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x}, y|\Theta^*, \pi_t^*) \cdot \log \frac{p_t(\mathbf{x}, y|\breve{\Theta}, \breve{\pi}_t)}{p_t(\mathbf{x}, y|\Theta^*, \pi_t^*)} d\mathbf{x}dy \\
&\quad -\int_{\mathbf{x},y\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x}, y|\Theta^*, \pi_t^*) \cdot \log p_t(\mathbf{x}, y|\Theta^*, \pi_t^*) d\mathbf{x}dy && (15)\\
&= \mathcal{KL}\left(p_t\left(\cdot|\Theta^*, \pi_t^*\right) \| p_t\left(\cdot|\breve{\Theta}, \breve{\pi}_t\right)\right) + H\left[p_t\left(\cdot|\Theta^*, \pi_t^*\right)\right], && (16)
\end{aligned}$$

[4] Note that in FL settings, we have the extra difficulty that client t cannot have direct access to samples $\mathcal{U}_t$, since local data cannot be moved across clients.

where

$$H\left[p_t\left(\cdot|\Theta^*,\pi_t^*\right)\right] = -\int_{\mathbf{x},y\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x},y|\Theta^*,\pi_t^*)\cdot\log p_t(\mathbf{x},y|\Theta^*,\pi_t^*)d\mathbf{x}dy, \tag{17}$$

is the entropy of $p_t\left(\cdot|\Theta^*,\pi_t^*\right)$, and

$$\mathcal{KL}\left(p_t\left(\cdot|\Theta^*,\pi_t^*\right)\|p_t\left(\cdot|\breve{\Theta},\breve{\pi}_t\right)\right) = \int_{\mathbf{x},y\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x},y|\Theta^*,\pi_t^*)\cdot\log\frac{p_t(\mathbf{x},y|\Theta^*,\pi_t^*)}{p_t(\mathbf{x},y|\breve{\Theta},\breve{\pi}_t)}d\mathbf{x}dy, \tag{18}$$

is the Kullback-Leibler divergence between $p_t\left(\cdot|\Theta^*,\pi_t^*\right)$ and $p_t\left(\cdot|\breve{\Theta},\breve{\pi}_t\right)$. Since the $\mathcal{KL}$ divergence is non-negative, we have

$$\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\breve{\Theta},\breve{\pi}_t)\right] \geq H\left[p_t\left(\cdot|\Theta^*,\pi_t^*\right)\right] = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\Theta^*,\pi_t^*)\right] \tag{19}$$

Taking the expectation over $t\sim\mathcal{D}_\mathcal{T}$, we write

$$\mathop{\mathbb{E}}_{t\sim\mathcal{D}_\mathcal{T}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\breve{\Theta},\breve{\pi}_t)\right] \geq \mathop{\mathbb{E}}_{t\sim\mathcal{D}_\mathcal{T}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\Theta^*,\pi_t^*)\right]. \tag{20}$$

Since $\breve{\Theta}$ and $\breve{\Pi}$ is a solution of Problem (4), we also have

$$\mathop{\mathbb{E}}_{t\sim\mathcal{D}_\mathcal{T}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\breve{\Theta},\breve{\pi}_t)\right] \leq \mathop{\mathbb{E}}_{t\sim\mathcal{D}_\mathcal{T}}\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[-\log p_t(\mathbf{x},y|\Theta^*,\pi_t^*)\right]. \tag{21}$$

Combining (20), (21), and (16), we have

$$\mathop{\mathbb{E}}_{t\sim\mathcal{D}_\mathcal{T}}\mathcal{KL}\left(p_t\left(\cdot|\Theta^*,\pi_t^*\right)\|p_t\left(\cdot|\breve{\Theta},\breve{\pi}_t\right)\right) = 0. \tag{22}$$

Since $\mathcal{KL}$ divergence is non-negative, and the support of $\mathcal{D}_\mathcal{T}$ is the countable set $\mathcal{T}$, it follows that

$$\forall t\in\mathcal{T},\quad \mathcal{KL}\left(p_t\left(\cdot|\Theta^*,\pi_t^*\right)\|p_t\left(\cdot|\breve{\Theta},\breve{\pi}_t\right)\right) = 0. \tag{23}$$

Thus,

$$p_t(\mathbf{x},y|\breve{\Theta},\breve{\pi}_t) = p_t(\mathbf{x},y|\Theta^*,\pi_t^*),\quad \forall t\in\mathcal{T}. \tag{24}$$

□

In the following, we prove Prop. 2.1 in each of the three possible cases for the loss function.

B.1. Case of Mean Squared Error Loss

Lemma B.2. *Suppose that Assumption 2 holds, $\mathcal{Y}=\mathbb{R}^d$ for some $d>0$, and for $t\in\mathcal{T}$ and $m\in[M]$,*

$$z_t\sim\mathcal{M}(\breve{\pi}_t);\qquad p_t(y|\mathbf{x},z=m) = \mathcal{N}\left(y|h_{\breve{\theta}_m}(\mathbf{x}),I_d\right),$$

where $\mathcal{N}\left(y|h_{\breve{\theta}_m}(\mathbf{x}),I_d\right)$ is the d-dimensional Gaussian distribution with mean $h_{\breve{\theta}_m}(\mathbf{x})$ and co-variance I_d. Then, h_t^* defined as

$$\forall\mathbf{x}\in\mathcal{X};\ \ h_t^*(\mathbf{x}) = \sum_{m=1}^{M}\breve{\pi}_{tm}h_{\breve{\theta}_m}(\mathbf{x}),$$

minimizes the mean squared error

$$\mathcal{L}_{\mathcal{D}_t}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\|h(\mathbf{x})-y\|^2.$$

Proof.

$$\mathcal{L}_{\mathcal{D}_t}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t}\frac{\|h(\mathbf{x})-y\|^2}{2} \tag{25}$$

$$= \sum_{m=1}^{M} \frac{\breve{\pi}_{tm}}{2} \cdot \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_m} \|h(\mathbf{x}) - y\|^2 \tag{26}$$

$$= \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x},y\in\mathcal{X}\times\mathbb{R}^d} \frac{\|h(\mathbf{x}) - y\|^2}{\sqrt{(2\pi)^d}} \exp\left[-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right] p_m(\mathbf{x})d\mathbf{x}dy \tag{27}$$

We compute the gradient of $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t} \|h(\mathbf{x}) - y\|^2$ with respect to h as,

$$\nabla_h \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \frac{\|h(\mathbf{x}) - y\|^2}{2} = \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x},y} \frac{h(\mathbf{x}) - y}{\sqrt{(2\pi)^d}} \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} p_m(\mathbf{x})d\mathbf{x}dy \tag{28}$$

We write first-order optimality condition at h_t^*,

$$\nabla_h \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \|h_t^*(\mathbf{x}) - y\|^2 = 0, \tag{29}$$

thus,

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x},y} \frac{h_t^*(\mathbf{x}) - y}{\sqrt{(2\pi)^d}} \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} p_m(\mathbf{x})d\mathbf{x}dy = 0, \tag{30}$$

rearranging the terms leads to,

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x},y} h_t^*(\mathbf{x}) \cdot \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} p_m(\mathbf{x})d\mathbf{x}dy =$$

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x},y} y \cdot \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} p_m(\mathbf{x})d\mathbf{x}dy, \tag{31}$$

then,

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} h_t^*(\mathbf{x}) \cdot \left\{\int_{y\in\mathbb{R}^d} \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} dy\right\} p_m(\mathbf{x})d\mathbf{x} = \tag{32}$$

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \left\{\int_{y\in\mathbb{R}^d} y \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} dy\right\} p_m(\mathbf{x})d\mathbf{x}, \tag{33}$$

using the fact that

$$\forall x \in \mathcal{X};\ \frac{1}{\sqrt{(2\pi)^d}} \int_{y\in\mathbb{R}} \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} dy = 1, \tag{34}$$

and that

$$\frac{1}{\sqrt{(2\pi)^d}} \int_{y\in\mathbb{R}} y \exp\left\{-\frac{\left\|h_{\breve{\theta}_m}(\mathbf{x}) - y\right\|^2}{2}\right\} dy = h_{\breve{\theta}_m}(\mathbf{x}), \tag{35}$$

it follows that,

$$\sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} h_t^*(\mathbf{x}) p_m(\mathbf{x})d\mathbf{x} = \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} h_{\breve{\theta}_m(\mathbf{x})} p_m(\mathbf{x})d\mathbf{x}. \tag{36}$$

Using Assumption 2 and the fact that $\sum_{m=1}^{M} \breve{\pi}_{tm} = 1;\ t \in [T]$, we have

$$\int_{\mathbf{x}\in\mathcal{X}} \left(h_t^*(\mathbf{x}) - \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x} = 0. \tag{37}$$

Eq. 37 suggest the following optimality conditions,

$$\forall \mathbf{x} \in \mathcal{X};\ \ h_t^*(\mathbf{x}) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}). \tag{38}$$

Finally, since $h \mapsto \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t} \|h(\mathbf{x}) - y\|^2$ is convex, it follows that h_t^*, defined in (5) minimizes $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t} \|h(\mathbf{x}) - y\|^2$. □

B.2. Case of logistic loss

Lemma B.3. *Suppose that Assumption 2 holds, $\mathcal{Y} = \{0, 1\}$ and for $t \in \mathcal{T}$ and $m \in [M]$,*

$$z_t \sim \mathcal{M}(\pi_t); \qquad p_t(y|\mathbf{x}, z = m) = \mathcal{B}\left(y|h_{\breve{\theta}_m}(\mathbf{x})\right),$$

where $\mathcal{B}\left(y|h_{\breve{\theta}_m}(\mathbf{x})\right)$ is Bernoulli distribution with parameter $h_{\breve{\theta}_m}(\mathbf{x})$. Then, h_t^* defined as

$$\forall \mathbf{x} \in \mathcal{X};\ \ h_t^*(\mathbf{x}) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}),$$

minimizes the logistic loss

$$\mathcal{L}_{\mathcal{D}_t}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \{-y \log[h(\mathbf{x})] - (1-y)\log[1-h(\mathbf{x})]\}.$$

Proof.

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_t}(h) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} (-y \log[h(\mathbf{x})] - (1-y)\log[1-h(\mathbf{x})]) && (39)\\
&= \sum_{m=1}^{M} \breve{\pi}_{tm} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_m} (-y \log[h(\mathbf{x})] - (1-y)\log[1-h(\mathbf{x})]) && (40)\\
&= \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \{-\log[1-h(\mathbf{x})] \cdot p_m(\mathbf{x}, y=1) - \log[h(\mathbf{x})] \cdot p_m(\mathbf{x}, y=0)\} d\mathbf{x} && (41)\\
&= \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \{-\log[1-h(\mathbf{x})] \cdot p_m(y=1|\mathbf{x}) - \log[h(\mathbf{x})] \cdot p_m(y=0|\mathbf{x})\} p_m(\mathbf{x}) d\mathbf{x} && (42)\\
&= \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \left\{-\log[1-h(\mathbf{x})] \cdot \left(1 - h_{\breve{\theta}_m}(\mathbf{x})\right) - \log[h(\mathbf{x})] \cdot h_{\breve{\theta}_m}(\mathbf{x})\right\} p_m(\mathbf{x}) d\mathbf{x} && (43)
\end{aligned}$$

Using Assumption 2, we have

$$\mathcal{L}_{\mathcal{D}_t}(h) = \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \left\{-\log[1-h(\mathbf{x})] \cdot \left(1 - h_{\breve{\theta}_m}(\mathbf{x})\right) - \log[h(\mathbf{x})] \cdot h_{\breve{\theta}_m}(\mathbf{x})\right\} p(\mathbf{x}) d\mathbf{x} \tag{44}$$

We compute the gradient of $\mathcal{L}_{\mathcal{D}_t}(h)$ with respect to h,

$$\nabla_h \mathcal{L}_{\mathcal{D}_t}(h) = \sum_{m=1}^{M} \breve{\pi}_{tm} \int_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{1 - h_{\breve{\theta}_m}(\mathbf{x})}{1 - h(\mathbf{x})} - \frac{h_{\breve{\theta}_m}(\mathbf{x})}{h(\mathbf{x})} \right\} p(\mathbf{x}) d\mathbf{x} \tag{45}$$

$$= \sum_{m=1}^{M} \breve{\pi}_{tm} \cdot \int_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{h(\mathbf{x}) - h_{\breve{\theta}_m}(\mathbf{x})}{(1-h(\mathbf{x}))\cdot h(\mathbf{x})} \right\} p(\mathbf{x}) d\mathbf{x} \tag{46}$$

$$= \int_{\mathbf{x}\in\mathcal{X}} \sum_{m=1}^{M} \breve{\pi}_{tm} \cdot \left\{ \frac{h(\mathbf{x}) - h_{\breve{\theta}_m}(\mathbf{x})}{(1-h(\mathbf{x}))\cdot h(\mathbf{x})} \right\} p(\mathbf{x}) d\mathbf{x} \tag{47}$$

$$= \int_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{h(\mathbf{x}) - \sum_{m=1}^{M} \breve{\pi}_{tm} \cdot h_{\breve{\theta}_m}(\mathbf{x})}{(1-h(\mathbf{x}))\cdot h(\mathbf{x})} \right\} p(\mathbf{x}) d\mathbf{x} \tag{48}$$

First-order optimality condition at h_t^* is,

$$\int_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{h_t^*(\mathbf{x}) - \sum_{m=1}^{M} \breve{\pi}_{tm} \cdot h_{\breve{\theta}_m}(\mathbf{x})}{(1-h_t^*(\mathbf{x}))\cdot h_t^*(\mathbf{x})} \right\} p(\mathbf{x}) d\mathbf{x} = 0 \tag{49}$$

Suggesting that the following candidate optimality condition,

$$\forall \mathbf{x} \in \mathcal{X};\ \ h_t^*(\mathbf{x}) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}). \tag{50}$$

Finally, exploiting the convexity of the logistic loss, h_t^* minimizes it. □

B.3. Case of Cross-Entropy loss

Lemma B.4. *Suppose that Assumption 2 holds and $\mathcal{Y} = \{0, \ldots, L\}$ for some $L > 2$, and for $t \in \mathcal{T}$ and $m \in [M]$,*

$$z_t \sim \mathcal{M}(\pi_t); \qquad p_t(y|\mathbf{x}, z = m) = \mathcal{M}\left(y|h_{\breve{\theta}_m}(\mathbf{x})\right),$$

where $\mathcal{M}\left(y|h_{\breve{\theta}_m}(\mathbf{x})\right)$ is the multinomial distribution with parameter $h_{\breve{\theta}_m}(\mathbf{x}) \in \Delta^L$. Then, h_t^* defined as

$$\forall \mathbf{x} \in \mathcal{X};\ \ h_t^*(\mathbf{x}) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}(\mathbf{x}),$$

minimizes the cross-entropy loss

$$\mathcal{L}_{\mathcal{D}_t}(h) = -\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \sum_{l=1}^{L} \mathbb{1}_{\{y=l\}} \log\left(h_{w_t}(\mathbf{x})\right)_l,$$

on Δ^L.

Proof. We express the fact that $h \in \Delta^L$, as

$$\forall \mathbf{x} \in \mathcal{X},\ \sum_{l=1}^{L} \left(h(\mathbf{x})\right)_l = 1,$$

obtaining the following problem

$$\begin{aligned} \underset{h}{\text{minimize}} \quad & -\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \sum_{l=1}^{L} \mathbb{1}_{\{y=l\}} \log\left(h(\mathbf{x})\right)_l \\ \text{subject to} \quad & \forall \mathbf{x} \in \mathcal{X}, \quad \sum_{l=1}^{L} \left(h(\mathbf{x})\right)_l = 1 \end{aligned}$$

which is equivalent to

$$\begin{aligned} \underset{h}{\text{minimize}} \quad & -\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \sum_{l=1}^{L} \mathbb{1}_{\{y=l\}} \log\left(h(\mathbf{x})\right)_l \\ \text{subject to} \quad & \int_{x\in\mathcal{X}} \left\{\sum_{l=1}^{L} \left(h\left(\mathbf{x}\right)\right)_l\right\} p\left(\mathbf{x}\right) d\mathbf{x} = 1 \end{aligned}$$

The Lagrangian of this problem is

$$L\left(h,\lambda\right) = -\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_t} \sum_{l=1}^{L} \mathbb{1}_{\{y=l\}} \log\left(h(\mathbf{x})\right)_l + \lambda\left[\int_{x\in\mathcal{X}} \left\{\sum_{l=1}^{L} \left(h\left(\mathbf{x}\right)\right)_l\right\} p\left(\mathbf{x}\right) d\mathbf{x} - 1\right]. \tag{51}$$

Using assumption 2, we can simplify the Lagrangian as

$$L\left(h,\lambda\right) = \int_{x\in\mathcal{X}} \left\{-\sum_{m=1}^{M} \breve{\pi}_{tm} \sum_{l=1}^{L} \left(h_{\breve{\theta}_m}\left(\mathbf{x}\right)\right)_l \cdot \log\left(h\left(\mathbf{x}\right)\right)_l + \lambda\left(\sum_{l=1}^{L} \left(h\left(\mathbf{x}\right)\right)_l - 1\right)\right\} p(\mathbf{x}) d\mathbf{x}. \tag{52}$$

KKT first order optimality conditions at h_t^* are written

$$\begin{cases} \nabla_h L\left(h_t^*,\lambda\right) & = 0 \\ \int_{x\in\mathcal{X}} \left\{\sum_{l=1}^{L} \left(h_t^*\left(\mathbf{x}\right)\right)_l - 1\right\} p\left(\mathbf{x}\right) d\mathbf{x} & = 0, \end{cases} \tag{53}$$

then,

$$\begin{cases} \int_{x\in\mathcal{X}} \sum_{l=1}^{L} \left\{-\sum_{m=1}^{M} \breve{\pi}_{tm} \frac{\left(h_{\breve{\theta}_m}(\mathbf{x})\right)_l}{\left(h_t^*(\mathbf{x})\right)_l} + \lambda\right\} \cdot p(\mathbf{x}) d\mathbf{x} & = 0 \\ \int_{x\in\mathcal{X}} \left\{\sum_{l=1}^{L} \left(h_t^*\left(\mathbf{x}\right)\right)_l - 1\right\} p\left(\mathbf{x}\right) d\mathbf{x} & = 0 \end{cases} \tag{54}$$

Leading to the following first-order optimality condition for any $\mathbf{x} \in \mathcal{X}$

$$\begin{cases} \lambda \cdot h_t\left(\mathbf{x}\right) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}\left(\mathbf{x}\right) \\ \sum_{l=1}^{L} \left(h_t\left(\mathbf{x}\right)\right)_l = 1 \end{cases} \tag{55}$$

Since $h_{\breve{\theta}_m}\left(\mathbf{x}\right) \in \Delta^L$ and $\pi_t \in \Delta^M$, it follows that $\lambda = 1$. Thus

$$h_{w_t}\left(\mathbf{x}\right) = \sum_{m=1}^{M} \breve{\pi}_{tm} h_{\breve{\theta}_m}\left(\mathbf{x}\right) \tag{56}$$

Finally, since the cross-entropy loss is convex in h, h_t^* is a minimizer. □

C. Relation with Other Multi-Task Learning Frameworks

In this appendix, we give more details about the relation of our formulation with existing frameworks for (federated) MTL sketched in Sec. 2.2. We suppose that Assumptions 1–3 hold and that each client learns a predictor of the form (5). Note that this is more general than (Zantedeschi et al., 2020), where each client learns a personal hypothesis as a weighted combination of a set of M base *known* hypothesis, since the base hypothesis and *not only the weights* are learned in our case.

Alternating Structure Optimization (Zhou et al., 2011). Alternating structure optimization (ASO) is a popular MTL approach that learns a shared low-dimensional predictive structure on hypothesis spaces from multiple related tasks, i.e., all tasks are assumed to share a common feature space $P \in \mathbb{R}^{d' \times d}$, where $d' \leq \min(T, d)$ is the dimensionality of the shared feature space and P has orthonormal columns ($PP^\intercal = I_{d'}$), i.e., P is *semi-orthogonal matrix*. ASO leads to the following formulation:

$$\underset{W,P:PP^\intercal=I_{d'}}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n_t} l\left(h_{w_t}\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right) + \alpha\left(\operatorname{tr}\left(WW^\intercal\right) - \operatorname{tr}\left(WP^\intercal PW^\intercal\right)\right) + \beta \operatorname{tr}\left(WW^\intercal\right), \tag{57}$$

where $\alpha \geq 0$ is the regularization parameter for task relatedness and $\beta \geq 0$ is an additional L2 regularization parameter.

When the hypothesis $(h_\theta)_\theta$ are assumed to be linear, Eq. (5) can be written as $W = \Pi\Theta$. Writing the LQ decomposition[5] of matrix Θ, i.e., $\Theta = LQ$, where $L \in \mathbb{R}^{M \times M}$ is a lower triangular matrix and $Q \in \mathbb{R}^{M \times d}$ is a semi-orthogonal matrix ($QQ^\intercal = I_M$), (5) becomes $W = \Pi LQ \in \mathbb{R}^{T \times d}$, thus, $W = WQ^\intercal Q$, leading to the constraint $\|W - WQ^\intercal Q\|_F^2 = \operatorname{tr}\left(WW^\intercal\right) - \operatorname{tr}\left(WQ^\intercal QW^\intercal\right) = 0$. If we assume $\|\theta_m\|_2^2$ to be bounded by a constant $B > 0$ for all $m \in [M]$, we get the constraint $\operatorname{tr}\left(WW^\intercal\right) \leq TB$. It means that minimizing $\sum_{t=1}^{T}\sum_{i=1}^{n_t} l\left(h_{w_t}\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right)$ under our Assumption 1 can be formulated as the following constrained optimization problem

$$\begin{aligned} \underset{W,Q:QQ^\intercal=I_M}{\text{minimize}} \quad & \sum_{t=1}^{T}\sum_{i=1}^{n_t} l\left(h_{w_t}\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right), \\ \text{subject to} \quad & \operatorname{tr}\left\{WW^\intercal\right\} - \operatorname{tr}\left\{WQ^\intercal QW^\intercal\right\} = 0, \\ & \operatorname{tr}\left(WW^\intercal\right) \leq TB. \end{aligned} \tag{58}$$

Thus, there exists Lagrange multipliers $\alpha \in \mathbb{R}$ and $\beta > 0$, for which Problem (58) is equivalent to the following regularized optimization problem

$$\underset{W,Q:QQ^\intercal=I_M}{\text{minimize}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} l\left(h_{w_t}\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right) + \alpha\left(\operatorname{tr}\left\{WW^\intercal\right\} - \operatorname{tr}\left\{WQ^\intercal QW^\intercal\right\}\right) + \beta \operatorname{tr}\left\{WW^\intercal\right\}, \tag{59}$$

which is exactly Problem (57).

Federated MTL via task relationships. The ASO formulation above motivated the authors of (Smith et al., 2017) to learn personalized models by solving the following problem

$$\min_{W,\Omega} \sum_{t=1}^{T}\sum_{i=1}^{n_t} l\left(h_{w_t}\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right) + \lambda \operatorname{tr}\left(W\Omega W^\intercal\right), \tag{60}$$

Two alternative MTL formulations are presented in (Smith et al., 2017) to justify Problem (60): MTL with probabilistic priors (Zhang & Yeung, 2010) and MTL with graphical models (Lauritzen, 1996). Both of them can be covered using our Assumption 1 as follows:

- Considering $T = M$ and $\Pi = I_M$ in Assumption 1 and introducing a prior on Θ of the form

$$\Theta \sim \left(\prod \mathcal{N}\left(0, \sigma^2 I_d\right)\right) \mathcal{MN}\left(I_d \otimes \Omega\right) \tag{61}$$

  lead to a formulation similar to MTL with probabilistic priors (Zhang & Yeung, 2010).

[5] Note that when Θ is a full rank matrix, this decomposition is unique.

- Two tasks t and t' are independent if $\langle \pi_t, \pi_{t'} \rangle = 0$, thus using $\Omega_{t,t'} = \langle \pi_t, \pi_{t'} \rangle$ leads to the same graphical model as in (Lauritzen, 1996)

Several personalized FL formulations, e.g., `pFedMe`(Dinh et al., 2020), `FedU` (Dinh et al., 2021) and the formulation studied in (Hanzely & Richtárik, 2020) and in (Hanzely et al., 2020), are special cases of formulation (61).

D. Centralized Expectation Maximization

Proposition 3.1. *Under Assumptions 1 and 2, at the k-th iteration the EM algorithm updates parameter estimates through the following steps:*

E-step:
$$q_t^{k+1}(z_t^{(i)} = m) \propto \pi_{tm}^k \cdot \exp\left(-l(h_{\theta_m^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right), \qquad t \in [T],\ m \in [M],\ i \in [n_t] \quad (8)$$

M-step:
$$\pi_{tm}^{k+1} = \frac{\sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m)}{n_t}, \qquad t \in [T],\ \ m \in [M] \quad (9)$$

$$\theta_m^{k+1} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \sum_{i=1}^{n_t} q_t^{k+1}(z_t^{(i)} = m) \cdot l\big(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)}\big), \qquad m \in [M] \quad (10)$$

Proof. The objective is to learn parameters $\{\breve{\Theta}, \breve{\Pi}\}$ from the data $\mathcal{S}_{1:T}$ by maximizing the likelihood $p\left(S_{1:T}|\Theta, \Pi\right)$. We introduce functions $q_t(z),\ t \in [T]$ such that $q_t \geq 0$ and $\sum_{z=1}^{M} q_t(z) = 1$ in the expression of the likelihood. For $\Theta \in \mathbb{R}^{M \times d}$ and $\Pi \in \Delta^{T \times M}$, we have

$$\log p(\mathcal{S}_{1:T}|\Theta, \Pi) = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \log p_t\left(s_t^{(i)}|\Theta, \pi_t\right) \quad (62)$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \log \left[\sum_{m=1}^{M} \left(\frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right)}{q_t\left(z_t^{(i)} = m\right)} \right) q_t\left(z_t^{(i)} = m\right) \right] \quad (63)$$

$$\geq \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log \frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right)}{q_t\left(z_t^{(i)} = m\right)} \quad (64)$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right) - \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log q_t\left(z_t^{(i)} = m\right) \quad (65)$$

$$\triangleq \mathfrak{L}(\Theta, \Pi, Q_{1:T}), \quad (66)$$

where we used Jensen's inequality because $\log$ is concave. $\mathfrak{L}(\Theta, \Pi, Q_{1:T})$ is an *evidence lower bound*. The centralized EM-algorithm corresponds to iteratively maximizing this bound with respect to $Q_{1:T}$ (E-step) and with respect to $\{\Theta, \Pi\}$ (M-step).

E-step. The difference between the log-likelihood and the evidence lower bound $\mathfrak{L}(\Theta, \Pi, Q_{1:T})$ can be expressed in terms of a sum of $\mathcal{KL}$ divergences:

$$\log p(\mathcal{S}_{1:T}|\Theta, \Pi) - \mathfrak{L}(\Theta, \Pi, Q_{1:T}) = \quad (67)$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left\{ \log p_t\left(s_t^{(i)}|\Theta, \pi_t\right) - \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \cdot \log \frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right)}{q_t\left(z_t^{(i)} = m\right)} \right\} \quad (68)$$

$$= \sum_{t=1}^{T} \sum_{t=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \left(\log p_t\left(s_t^{(i)}|\Theta, \pi_t\right) - \log \frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right)}{q_t\left(z_t^{(i)} = m\right)} \right) \quad (69)$$

$$= \sum_{t=1}^{T} \sum_{t=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log \frac{p_t\left(s_t^{(i)}|\Theta, \pi_t\right) \cdot q_t\left(z_t^{(i)} = m\right)}{p_t\left(s_t^{(i)}, z_t^{(i)} = m|\Theta, \pi_t\right)} \quad (70)$$

$$= \sum_{t=1}^{T} \sum_{t=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log \frac{q_t\left(z_t^{(i)} = m\right)}{p_t\left(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t\right)} \tag{71}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \mathcal{KL}\left(q_t\left(z_t^{(i)}\right) || p_t\left(z_t^{(i)} | s_t^{(i)}, \Theta, \pi_t\right)\right) \geq 0. \tag{72}$$

For fixed parameters $\{\Theta, \Pi\}$, the maximum of $\mathfrak{L}(\Theta, \Pi, Q_{1:T})$ is reached when

$$\sum_{t=1}^{T} \sum_{i=1}^{n_t} \mathcal{KL}\left(q_t\left(z_t^{(i)}\right) || p_t\left(z_t^{(i)} | s_t^{(i)}, \Theta, \pi_t\right)\right) = 0.$$

Thus for $t \in [T]$ and $i \in [n_t]$, we have:

$$q_t(z_t^{(i)} = m) = p_t(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t) \tag{73}$$

$$= \frac{p_t(s_t^{(i)} | z_t^{(i)} = m, \Theta, \pi_t) \times p_t(z_t^{(i)} = m, \Theta, \pi_t)}{p_t\left(s_t^{(i)} | \Theta, \pi_t\right)} \tag{74}$$

$$= \frac{p_m(s_t^{(i)} | \theta_m) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'}(s_t^{(i)}) \times \pi_{tm'}} \tag{75}$$

$$= \frac{p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) \times p_m\left(\mathbf{x}_t^{(i)}\right) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'}\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_{m'}\right) \times p_{m'}\left(\mathbf{x}_t^{(i)}\right) \times \pi_{tm'}} \tag{76}$$

$$= \frac{p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) \times p\left(\mathbf{x}_t^{(i)}\right) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'}\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, , \theta_{m'}\right) \times p\left(\mathbf{x}_t^{(i)}\right) \times \pi_{tm'}} \tag{77}$$

$$= \frac{p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'}\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_{m'}\right) \times \pi_{tm'}}, \tag{78}$$

where (77) relies on Assumption 2. It follows that

$$q_t(z_t^{(i)} = m) = p_t(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t) = \frac{p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'}\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_{m'}\right) \times \pi_{tm'}}. \tag{79}$$

M-step. We maximize now $\mathfrak{L}(\Theta, \Pi, Q_{1:T})$ with respect to $\{\Theta, \Pi\}$. By dropping the terms not depending on $\{\Theta, \Pi\}$ in the expression of $\mathfrak{L}(\Theta, \Pi, Q_{1:T})$ we write:

$$\mathfrak{L}(\Theta, \Pi, Q_{1:T}) = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \log p_t\left(s_t^{(i)}, z_t^{(i)} = m | \Theta, \pi_t\right) + c \tag{80}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \Big[\log p_t\left(s_t^{(i)} | z_t^{(i)} = m, \Theta, \pi_t\right) + \log p_t\left(z_t^{(i)} = m | \Theta, \pi_t\right) \Big] + c \tag{81}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \left[\log p_{\theta_m}\left(s_t^{(i)}\right) + \log \pi_{tm}\right] + c \tag{82}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t\left(z_t^{(i)} = m\right) \left[\log p_{\theta_m}\left(y_t^{(i)} | \mathbf{x}_t^{(i)}\right) + \log p_m\left(\mathbf{x}_t^{(i)}\right) + \log \pi_{tm}\right] + c \tag{83}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} q_t \left(z_t^{(i)} = m \right) \left[\log p_{\theta_m} \left(y_t^{(i)} | \mathbf{x}_t^{(i)} \right) + \log \pi_{tm} \right] + c', \tag{84}$$

$$\tag{85}$$

where c and c' are constant not depending on $\{\Theta, \Pi\}$.

Thus, for $t \in [T]$ and $m \in [M]$, by solving a simple optimization problem we update π_{tm} as follows:

$$\pi_{tm} = \frac{\sum_{i=1}^{n_t} q_t(z_t^{(i)} = m)}{n_t}. \tag{86}$$

On the other hand, for $m \in [M]$, we update θ_m by solving:

$$\theta_m \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{n_t} q_t(z_t^{(i)} = m) \times l \left(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)} \right). \tag{87}$$

□

E. Details on Federated Expectation-Maximization

As mentioned in Section 3.2, under the following standard assumptions (see e.g., (Wang et al., 2020b)), `FedEM` (see Alg. 7) converges to a stationary point of f. Below, we use the more compact notation $l(\theta; s_t^{(i)}) \triangleq l(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)})$.

Assumption 4. *The negative log-likelihood f is bounded below by $f^* \in \mathbb{R}$.*

Assumption 5. *(Smoothness) For all $t \in [T]$ and $i \in [n_t]$, the function $\theta \mapsto l(\theta; s_t^{(i)})$ is L-smooth and twice continuously differentiable.*

Assumption 6. *(Unbiased gradients and bounded variance) Each client $t \in [T]$ can sample a random batch ξ from $\mathcal{S}_t$ and compute an unbiased estimator $g_t(\theta, \xi)$ of the local gradient with bounded variance, i.e., $\mathbb{E}_\xi[g_t(\theta, \xi)] = \frac{1}{n_t}\sum_{i=1}^{n_t} \nabla_\theta l(\theta; s_t^{(i)})$ and $\mathbb{E}_\xi \|g_t(\theta, \xi) - \frac{1}{n_t}\sum_{i=1}^{n_t} \nabla_\theta l(\theta; s_t^{(i)})\|^2 \leq \sigma^2$.*

Assumption 7. *(Bounded dissimilarity) There exist β and G such that for any set of weights $\alpha \in \Delta^M$:*

$$\sum_{t=1}^{T} \frac{n_t}{n} \Big\| \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{m=1}^{M} \alpha_m \cdot l(\theta; s_t^{(i)}) \Big\|^2 \leq \qquad G^2 + \beta^2 \Big\| \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{m=1}^{M} \alpha_m \cdot l(\theta; s_t^{(i)}) \Big\|^2. \tag{88}$$

Assumption 7 limits the level of dissimilarity of the different tasks, similarly to what is done in (Wang et al., 2020b).

F. Fully Decentralized Federated Expectation-Maximization

F.1. Fully Decentralized Algorithm

In some cases, clients may want to communicate directly in a peer-to-peer fashion instead of relying on the central server mediation (see Kairouz et al., 2019, Section 2.1). In fact, fully decentralized schemes may provide stronger privacy guarantees (Cyffers & Bellet, 2021) and speed-up training as they better use communication resources (Lian et al., 2017; Marfoq et al., 2020) and reduce the effect of stragglers (Neglia et al., 2019). For these reasons, they have attracted significant interest recently in the machine learning community (Lian et al., 2017; Vanhaesebrouck et al., 2017; Lian et al., 2018; Tang et al., 2018; Bellet et al., 2018; Neglia et al., 2020; Marfoq et al., 2020; Koloskova et al., 2020). We refer to (Nedić et al., 2018) for a comprehensive survey of fully decentralized optimization (also known as consensus-based optimization), and to (Koloskova et al., 2020) for a unified theoretical analysis of decentralized SGD.

We propose `D-FedEM` (Alg. 3 in App. I.2), a *fully decentralized version* of our federated expectation maximization algorithm. As in `FedEM`, the M-step for Θ update is replaced by an approximate maximization step consisting of local updates. The global aggregation step in `FedEM` (Alg. 1, line 13) is replaced by a partial aggregation step, where each client computes a weighted average of its current components and those of a subset of clients (its *neighborhood*), which may vary over time. The convergence of decentralized optimization schemes requires certain assumptions to guarantee that each client can influence the estimates of other clients over time. In our paper, we consider the general assumption in (Koloskova et al., 2020, Assumption 4)

Assumption 8 (Koloskova et al. (2020, Assumption 4))**.** *Symmetric doubly stochastic mixing matrices are drawn at each round k from (potentially different) distributions $W^k \sim \mathcal{W}^k$ and there exists two constants $p \in (0,1]$, and integer $\tau \geq 1$ such that for all $\Xi \in \mathbb{R}^{M\times d\times T}$ and all integers $l \in \{0,\dots,K/\tau\}$:*

$$\mathbb{E}\left\|\Xi W_{l,\tau} - \bar{\Xi}\right\|_{\mathcal{F}}^2 \leq (1-p)\left\|\Xi - \bar{\Xi}\right\|_{\mathcal{F}}^2, \tag{89}$$

where $W_{l,\tau} \triangleq W^{(l+1)\tau-1}\dots W^{l\tau}$, $\bar{\Xi} \triangleq \Xi\frac{\mathbf{1}\mathbf{1}^\intercal}{T}$, and the expectation is taken over the random distributions $W^k \sim \mathcal{W}^k$.

Assumption 8 expresses the fact that the sequence of mixing matrices, on average and every τ communication rounds, brings the values in the columns of Ξ closer to their row-wise average (thereby mixing the clients' updates over time). For instance, the assumption is satisfied if the communication graph is strongly connected every τ rounds, i.e., the graph $([T],\mathcal{E})$, where the edge (i,j) belongs to the graph if $w_{i,j}^h > 0$ for some $h \in \{k+1,\dots,k+\tau\}$ is connected.

Under Assumption 8 and some other standard mild assumptions, `D-FedEM` converges to a stationary point of f.

Theorem F.1. *Under Assumptions 1–8, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, `D-FedEM`'s iterates satisfy the following inequalities after a large enough number of communication rounds K:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_\Theta f\left(\bar{\Theta}^k,\Pi^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \quad \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{n_t}{n}\mathcal{KL}\left(\pi_t^k,\pi_t^{k-1}\right) \leq \mathcal{O}\left(\frac{1}{K}\right), \tag{90}$$

where $\bar{\Theta}^k = \Theta^k\frac{\mathbf{1}\mathbf{1}^\intercal}{T}$. Moreover, individual estimates $\left(\Theta_t^k\right)_{1\leq t\leq T}$ converge to consensus, i.e., to $\bar{\Theta}^k$:

$$\min_{k\in[K]}\mathbb{E}\sum_{t=1}^{T}\left\|\Theta_t^k - \bar{\Theta}^k\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

G. Reminder on Basic (Centralized) Surrogate Optimization

In this appendix, we recall the (centralized) *first-order surrogate optimization* framework introduced in (Mairal, 2013). In this framework, given a continuous function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we are interested in solving

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

using the majoration-minimization scheme presented in Alg. 2.

Algorithm 2 Basic Surrogate Optimization

Input: $\theta^0 \in \mathbb{R}^d$; number of iteration K
for iterations $k = 1, \ldots, K$ **do**
  compute g^k, a surrogate function of f near θ^{k-1}
  update solution: $\theta^k \in \arg\min g^k(\theta)$
end for

This procedure relies on surrogate functions, that approximate well the objective function in a neighborhood of a point. Reference (Mairal, 2013) focuses on *first-order surrogate functions* defined below.

Definition G.1 (First-Order Surrogate (Mairal, 2013))**.** A function $g : \mathbb{R}^d \mapsto \mathbb{R}$ is a first order surrogate of f near $\theta^k \in \mathbb{R}^d$ when the following is satisfied:

- **Majorization:** we have $g(\theta') \geq f(\theta')$ for all $\theta' \in \arg\min_{\theta \in \mathbb{R}^d} g(\theta)$. When the more general condition $g \geq f$ holds, we say that g is a **majorant** function.
- **Smoothness:** the approximation error $r \triangleq g - f$ is differentiable, and its gradient is L-Lipschitz. Moreover, we have $r(\theta^k) = 0$ and $\nabla r(\theta^k) = 0$.

H. Federated Surrogate Optimization

In this appendix, we give more details on the federated surrogate optimization framework mentioned in Sec. 3.2.

Our novel federated surrogate optimization framework minimizes an objective function $(\mathbf{u}, \mathbf{v}_{1:T}) \mapsto f(\mathbf{u}, \mathbf{v}_{1:T})$ that can be written as a weighted sum $f(\mathbf{u}, \mathbf{v}_{1:T}) = \sum_{t=1}^{T} \omega_t f_t(\mathbf{u}, \mathbf{v}_t)$ of T functions. We suppose that each client $t \in [T]$ can compute a partial first order surrogate of f_t, defined as follows.

Definition 1. *[Partial first-order surrogate] A function $g(\mathbf{u}, \mathbf{v}) : \mathbb{R}^{d_u} \times \mathcal{V} \to \mathbb{R}$ is a partial first-order surrogate of $f(\mathbf{u}, \mathbf{v})$ wrt $\mathbf{u}$ near $(\mathbf{u}_0, \mathbf{v}_0) \in R^{d_u} \times \mathcal{V}$ when the following conditions are satisfied:*

1. *$g(\mathbf{u}, \mathbf{v}) \geq f(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathcal{V}$;*
2. *$r(\mathbf{u}, \mathbf{v}) \triangleq g(\mathbf{u}, \mathbf{v}) - f(\mathbf{u}, \mathbf{v})$ is differentiable and L-smooth with respect to $\mathbf{u}$. Moreover, we have $r(\mathbf{u}_0, \mathbf{v}_0) = 0$ and $\nabla_{\mathbf{u}} r(\mathbf{u}_0, \mathbf{v}_0) = 0$.*
3. *$g(\mathbf{u}, \mathbf{v}_0) - g(\mathbf{u}, \mathbf{v}) = d_{\mathcal{V}}(\mathbf{v}_0, \mathbf{v})$ for all $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \arg\min_{\mathbf{v}' \in \mathcal{V}} g(\mathbf{u}, \mathbf{v}')$, where $d_{\mathcal{V}}$ is non-negative and $d_{\mathcal{V}}(\mathbf{v}, \mathbf{v}') = 0 \iff \mathbf{v} = \mathbf{v}'$.*

Under the assumption that each client t can compute a partial first order surrogate of f_t, we propose algorithms for federated surrogate optimization in both the client-server setting (Alg. 5) and the fully decentralized one (Alg. 9). Both algorithms are iterative and distributed: at each iteration $k > 0$, client $t \in [T]$ computes a partial first-order surrogate g_t^k of f_t near $\left\{\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right\}$ (resp. $\left\{\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}\right\}$) for federated surrogate optimization in Alg. 5 (resp. for fully decentralized surrogate optimization in Alg 9).

The convergence of those two algorithms requires the following standard assumptions. Each of them generalizes one of the Assumptions 4–7 for our EM algorithms.

Assumption 4′. *The objective function f is bounded below by $f^* \in \mathbb{R}$.*

Assumption 5′. *(Smoothness) For all $t \in [T]$ and $k > 0$, g_t^k is L-smooth wrt to $\mathbf{u}$.*

Assumption 6′. *(Unbiased gradients and bounded variance) Each client $t \in [T]$ can sample a random batch ξ from $\mathcal{S}_t$ and compute an unbiased estimator $\nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v}; \xi)$ of the local gradient with bounded variance, i.e., $\mathbb{E}_\xi[\nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v}; \xi)] = \nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v})$ and $\mathbb{E}_\xi \|\nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v}; \xi) - \nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v})\|^2 \leq \sigma^2$.*

Assumption 7′. *(Bounded dissimilarity) There exist β and G such that*

$$\sum_{t=1}^{T} \omega_t \cdot \left\|\nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v})\right\|^2 \leq G^2 + \beta^2 \left\|\sum_{t=1}^{T} \omega_t \cdot \nabla_{\mathbf{u}} g_t^k(\mathbf{u}, \mathbf{v})\right\|^2.$$

Under these assumptions a parallel result to Thm. 3.2 holds for the client-server setting.

Theorem 3.2′. *Under Assumptions 4′–7′, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, the iterates of federated surrogate optimization (Alg. 5) satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \qquad \frac{1}{K}\sum_{k=1}^{K} \Delta_{\mathbf{v}} f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right), \tag{91}$$

where the expectation is over the random batches samples, and $\Delta_v f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) \triangleq f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^k\right) - f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^{k+1}\right) \geq 0$.

In the fully decentralized setting, if in addition to Assumptions 5′-7′, we suppose that Assumption 8 holds, a parallel result to Thm. F.1 holds.

Theorem F.1′. *Under Assumptions 4′–7′ and Assumption 8, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, the iterates of fully decentralized federated surrogate optimization (Alg. 9) satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\bar{\mathbf{u}}^k, v_{1:T}^k\right)\right\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \qquad \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T} \omega_t \cdot d_{\mathcal{V}}\left(\mathbf{v}_t^k, \mathbf{v}_t^{k+1}\right) \leq \mathcal{O}\left(\frac{1}{K}\right), \tag{92}$$

where $\bar{\mathbf{u}}^k = \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t^k$. *Moreover, local estimates* $\left(\mathbf{u}_t^k\right)_{1\le t\le T}$ *converge to consensus, i.e., to* $\bar{\mathbf{u}}^k$*:*

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|^2 \le \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

The proofs of Theorem 3.2′ and Theorem F.1′ are in Sec. J.1 and Sec. J.2, respectively.

I. Detailed Algorithms

I.1. Client-Server Algorithm

Algorithm 7 is a detailed version of Algorithm 1, with local SGD used as local solver.

Algorithm 5 gives our general algorithm for federated surrogate optimization, from which Algorithm 7 is derived.

Algorithm 3 `FedEM`: Federated Expectation-Maximization

Input: data $\mathcal{S}_{1:T}$; number of mixture distributions M; number of communication rounds K; number of local steps J
{Initialization}
for task $t = 1, \dots, T$ in parallel over T clients **do**
  randomly initialize $\Theta_t = (\theta_{m,t})_{1 \leq m \leq M} \in \mathbb{R}^{M \times d}$
  randomly initialize $\pi_t^0 \in \Delta^M$
end for
{Main loop}
for iterations $k = 1, \dots, K$ **do**
  sample $W^{k-1} \sim \mathcal{W}^{k-1}$
  for tasks $t = 1, \dots, T$ in parallel over T clients **do**
    for component $m = 1, \dots, M$ **do**
      {E-step}
      for sample $i = 1, \dots n_t$ **do**
        $q_t^k\left(z_t^{(i)} = m\right) \leftarrow \frac{\pi_{tm}^k \cdot \exp\left(-l(h_{\theta_m^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right)}{\sum_{m'=1}^M \pi_{tm'}^k \cdot \exp\left(-l(h_{\theta_{m'}^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right)}$
      end for
      {M-step}
      $\pi_{tm}^k \leftarrow \frac{\sum_{i=1}^{n_t} q_t^k(z_t^{(i)}=m)}{n_t}$
      $\theta_{m,t}^{k-1/2} \leftarrow$ `LocalSolver`(J, θ_m^{k-1}, $\left\{q_t^k\left(z_i^t = m\right)\right\}_{i \in [n_t]}$, $\mathcal{S}_t$, $\frac{n_t}{n}$)
    end for
    send $\theta_{m,t}^{k-1/2}$, $1 \leq m \leq M$ to neighbors
    receive $\theta_{m,s}^{k-1/2}$, $1 \leq m \leq M$ from neighbors $s \in \mathcal{N}_t$
    for component $m = 1, \dots, M$ **do**
      $\theta_{m,t}^k \leftarrow \sum_{s=1}^T w_{s,t}^{k-1} \cdot \theta_{m,s}^{k-1/2}$
    end for
  end for
end for

Algorithm 4 SGD solver used with `FedEM`

Input: number of iterations J; θ; samples' weights $q_{1:|\mathcal{S}|}$; data $\mathcal{S}$
for task $j = 0, \dots, J-1$ **do**
  sample indexes $\mathcal{I}$ from $1, \dots, |\mathcal{S}|$
  $\theta \leftarrow \theta - \eta_{k-1,j} \sum_{i \in \mathcal{I}} q_i \cdot \nabla_\theta l\left(h_\theta\left(\mathbf{x}^{(i)}\right), y^{(i)}\right)$
end for

Algorithm 5 Federated Surrogate Optimization

Input: $\mathbf{u}^0 \in \mathbb{R}^{d_u}$; $V^0 = \left(\mathbf{v}_t^0\right)_{1 \le t \le T} \in \mathcal{V}^T$; number of iterations K; number of local steps J
for iterations $k = 1, \ldots, K$ **do**
  server broadcast $\mathbf{u}^{k-1}$, $1 \le m \le M$ to the T clients
  for tasks $t = 1, \ldots, T$ in parallel over T clients **do**
    compute partial first-order surrogate function g_t^k pf f_t near $\{\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\}$
    $\mathbf{v}_t^k \leftarrow \arg\min_{\mathbf{v} \in \mathcal{V}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}\right)$
    $\mathbf{u}_t^k \leftarrow \texttt{LocalSolver}(J, \mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}, g_t^k, \mathcal{S}_t, \omega_t)$
    client t sends $\mathbf{u}_t^k$ to the server
  end for
  client t sends $\mathbf{u}_t^k$ to the server
  for component $m = 1, \ldots, M$ **do**
    $\mathbf{u}^k \leftarrow \sum_{t=1}^T \omega_t \cdot \mathbf{u}_t^k$
  end for
end for

Algorithm 6 SGD solver used with federated surrogate optimization

Input: number of iterations J; u; v; g
for task $j = 0, \ldots, J-1$ **do**
  sample batch $\xi^{k-1,j}$
  $\mathbf{u} \leftarrow \mathbf{u} - \eta_{k-1,j} \cdot \nabla_{\mathbf{u}} g\left(\mathbf{u}, \mathbf{v}, \xi^{k-1,j}\right)$
end for

I.2. Fully Decentralized Algorithm

Algorithm 3 shows `D-FedEM`, the fully decentralization version of our federated expectation maximization algorithm.

Algorithm 9 gives our general fully decentralized algorithm for federated surrogate optimization, from which Algorithm 3 is derived.

Algorithm 7 `D-FedEM`: Fully Decentralized Federated Expectation-Maximization

Input: data $\mathcal{S}_{1:T}$; number of mixture distributions M; number of communication rounds K; number of local steps J; mixing matrix distribution $\mathcal{W}^k$ for $k \in [K]$ {Initialization}
for task $t = 1, \dots, T$ in parallel over T clients **do**
  randomly initialize $\Theta_t^0 = (\theta_{m,t}^0)_{1 \leq m \leq M} \in \mathbb{R}^{M \times d}$
  randomly initialize $\pi_t^0 \in \Delta^M$
end for
{Main loop}
for iterations $k = 1, \dots, K$ **do**
  server broadcast $\theta_m^{k-1},\ 1 \leq m \leq M$ to the T clients
  for tasks $t = 1, \dots, T$ in parallel over T clients **do**
    for component $m = 1, \dots, M$ **do**
      {E-step}
      for sample $i = 1, \dots n_t$ **do**
        $q_t^k\left(z_t^{(i)} = m\right) \leftarrow \frac{\pi_{tm}^k \cdot \exp\left(-l(h_{\theta_m^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right)}{\sum_{m'=1}^M \pi_{tm'}^k \cdot \exp\left(-l(h_{\theta_{m'}^k}(\mathbf{x}_t^{(i)}), y_t^{(i)})\right)}$
      end for
      {M-step}
      $\pi_{tm}^k \leftarrow \frac{\sum_{i=1}^{n_t} q_t^k(z_t^{(i)} = m)}{n_t}$
      $\theta_{m,t}^k \leftarrow$ `LocalSolver`(J, θ_m^{k-1}, $\left\{q_t^k\left(z_i^t = m\right)\right\}_{i \in [n_t]}$, $\mathcal{S}_t$)
    end for
  end for
  client t sends $\theta_{m,t}^k,\ 1 \leq m \leq M$, to the server
  for component $m = 1, \dots, M$ **do**
    $\theta_m^k \leftarrow \sum_{t=1}^T \frac{n_t}{n} \cdot \theta_{m,t}^k$
  end for
end for

Algorithm 8 SGD solver used with `D-FedEM`

Input: number of iterations J; θ; samples' weights $q_{1:|\mathcal{S}|}$; data $\mathcal{S}$; $\frac{n_t}{n}$
for task $j = 0, \dots, J-1$ **do**
  sample indexes $\mathcal{I}$ from $1, \dots, |\mathcal{S}|$
  $\theta \leftarrow \theta - \frac{n_t}{n} \cdot \eta_{k-1,j} \sum_{i \in \mathcal{I}} q_i \cdot \nabla_\theta l\left(h_\theta\left(\mathbf{x}^{(i)}\right), y^{(i)}\right)$
end for

Algorithm 9 Fully-Decentralized Federated Surrogate Optimization

Input: $\mathbf{u}^0 \in \mathbb{R}^{d_u}$; $V^0 = \left(\mathbf{v}_t^0\right)_{1 \le t \le T} \in \mathcal{V}^T$; number of iterations K; number of local steps J; mixing matrix distribution $\mathcal{W}^k$ for $k \in [K]$
for iterations $k = 1, \ldots, K$ **do**
  sample $W^{k-1} \sim \mathcal{W}^{k-1}$
  server broadcast $\mathbf{u}^{k-1},\ 1 \le m \le M$ to the T clients
  for tasks $t = 1, \ldots, T$ in parallel over T clients **do**
    compute partial first-order surrogate function g_t^k pf f_t near $\{\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\}$
    $\mathbf{v}_t^k \leftarrow \arg\min_{\mathbf{v} \in \mathcal{V}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}\right)$
    $\mathbf{u}_t^{k-1/2} \leftarrow \texttt{LocalSolver}(J, \mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}, g_t^k, \omega_t)$
    send $\mathbf{u}_t^{k-1/2}$ to neighbors
    receive $\mathbf{u}_s^{k-1/2},\ 1 \le m \le M$ from neighbors $s \in \mathcal{N}_t$
    $\mathbf{u}_t^k \leftarrow \sum_{s=1}^T w_{s,t}^{k-1} \cdot \mathbf{u}_s^{k-1/2}$
  end for
end for

Algorithm 10 SGD solver used with fully decentralized federated surrogate optimization

Input: number of iterations J; $\mathbf{u}$; $\mathbf{v}$; g; ω_t
for task $j = 0, \ldots, J-1$ **do**
  sample batch $\xi^{k-1,j}$
  $\mathbf{u} \leftarrow \mathbf{u} - \omega_t \cdot \eta_{k-1,j} \cdot \nabla_{\mathbf{u}} g\left(\mathbf{u}, \mathbf{v}, \xi^{k-1,j}\right)$
end for

J. Convergence Proofs

We study the client-server setting and the fully decentralized setting in Sec. J.1 and Sec. J.2, respectively. In both cases, we first prove the more general result for surrogate optimization and then derive the specific result for `FedEM` and `D-FedEM`.

In this section, for conciseness we do not use bold fonts to denote vectors.

J.1. Client-Server Setting

J.1.1. Additional Notations

At iteration $k > 0$, we use $\mathbf{u}_t^{k-1,j}$ to denote the j-th iterate of the local solver at client $t \in [T]$, thus

$$\mathbf{u}_t^{k-1,0} = \mathbf{u}^{k-1}, \tag{93}$$

and

$$\mathbf{u}^k = \sum_{t=1}^{T} \omega_t \cdot \mathbf{u}_t^{k-1,J}. \tag{94}$$

At iteration $k > 0$, the local solver's updates at client $t \in [T]$ can be written as (for $0 \leq j \leq J-1$):

$$\mathbf{u}_t^{k-1,j+1} = \mathbf{u}_t^{k-1,j} - \eta_{k-1,j} \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right), \tag{95}$$

where $\xi_t^{k-1,j}$ is the batch drawn at the j-th local update of u_t^{k-1}.

We introduce $\eta_{k-1} = \sum_{j=0}^{J-1} \eta_{k-1,j}$, and we define the normalized update of the local solver at client $t \in [T]$ as,

$$\hat{\delta}_t^{k-1} \triangleq -\frac{\mathbf{u}_t^{k-1,J} - \mathbf{u}_t^{k-1,0}}{\eta_{k-1}} = \frac{\sum_{j=0}^{J-1} \eta_{k-1,j} \cdot \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right)}{\sum_{j=0}^{J-1} \eta_{k-1,j}}, \tag{96}$$

and also define

$$\delta_t^{k-1} \triangleq \frac{\sum_{j=0}^{J-1} \eta_{k-1,j} \cdot \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right)}{\eta_{k-1}}. \tag{97}$$

With this notation,

$$\mathbf{u}^k - \mathbf{u}^{k-1} = -\eta_{k-1} \cdot \sum_{t=1}^{T} \omega_t \cdot \hat{\delta}_t^{k-1}. \tag{98}$$

Finally, we define $g^k,\ k > 0$ as

$$g^k \left(\mathbf{u}, \mathbf{v}_{1:T} \right) = \sum_{t=1}^{T} \omega_t \cdot g_t^k \left(\mathbf{u}, \mathbf{v}_t \right). \tag{99}$$

Note that g^k is a convex combination of functions $g_t^k,\ t \in [T]$.

J.1.2. Proof of Theorem 3.2′

Lemma J.1. *Suppose that Assumptions 5′–7′ hold. Then, for $k > 0$, and $(\eta_{k,j})_{0 \leq j \leq J-1}$ such that $\eta_k \triangleq \sum_{j=0}^{J-1} \eta_{k,j} \leq \min \left\{ \frac{1}{2\sqrt{2}L}, \frac{1}{4L\beta} \right\}$, the updates of federated surrogate optimization (Alg 5) verify*

$$\begin{aligned} \mathbb{E}\left[\frac{f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) - f(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{\eta_{k-1}} \right] \leq -\frac{1}{4} \mathbb{E} \left\| \nabla_{\mathbf{u}} f \left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 - \frac{1}{\eta_{k-1}} \sum_{t=1}^{T} \omega_t \cdot d_{\mathcal{V}} \left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^k \right) \\ + 2\eta_{k-1} L \left(\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}^2}{\eta_{k-1}} L + 1 \right) \sigma^2 + 4\eta_{k-1}^2 L^2 G^2. \end{aligned} \tag{100}$$

Proof. This proof uses standard techniques from distributed stochastic optimization. It is inspired by (Wang et al., 2020b, Theorem 1).

For $k > 0$, g^k is L-smooth wrt $\mathbf{u}$, because it is a convex combination of L-smooth functions $g_t^k,\ t \in [T]$. Thus, we write

$$g^k\left(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}\right) - g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) \leq \left\langle \mathbf{u}^k - \mathbf{u}^{k-1}, \nabla_{\mathbf{u}} g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right\rangle + \frac{L}{2}\left\|\mathbf{u}^k - \mathbf{u}^{k-1}\right\|^2, \tag{101}$$

where $< \mathbf{u}, \mathbf{u}' >$ denotes the scalar product of vectors $\mathbf{u}$ and $\mathbf{u}'$. Using equation (98), and taking the expectation over random batches $\left(\xi_t^{k-1,j}\right)_{\substack{0\leq j\leq J-1\\ 1\leq t\leq T}}$, we have

$$\begin{aligned}\mathbb{E}\left[g^k\left(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}\right) - g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right] \leq \\ - \eta_{k-1} \underbrace{\mathbb{E}\left\langle \sum_{t=1}^T \omega_t \cdot \hat{\delta}_t^{k-1}, \nabla_{\mathbf{u}} g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right\rangle}_{\triangleq T_1} + \frac{L\eta_{k-1}^2}{2} \cdot \underbrace{\mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \hat{\delta}_t^{k-1} \right\|^2}_{\triangleq T_2}.\end{aligned} \tag{102}$$

We bound each of those terms separately, For T_1 we have

$$T_1 = \mathbb{E}\left\langle \sum_{t=1}^T \omega_t \cdot \hat{\delta}_t^{k-1}, \nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) \right\rangle \tag{103}$$

$$\begin{aligned}&= \mathbb{E}\left\langle \sum_{t=1}^T \omega_t \cdot \left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right), \nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) \right\rangle \\ &\quad + \mathbb{E}\left\langle \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1}, \nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) \right\rangle.\end{aligned} \tag{104}$$

Because stochastic gradients are unbiased (Assumption 6′), we have

$$\mathbb{E}\left[\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right] = 0, \tag{105}$$

thus,

$$T_1 = \mathbb{E}\left\langle \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1}, \nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) \right\rangle \tag{106}$$

$$= \frac{1}{2}\left(\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1} \right\|^2 \right) - \frac{1}{2}\mathbb{E}\left\| \nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) - \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1} \right\|^2. \tag{107}$$

For T_2 we have for $k > 0$,

$$T_2 = \mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \hat{\delta}_t^{k-1} \right\|^2 \tag{108}$$

$$= \mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right) + \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1} \right\|^2 \tag{109}$$

$$\leq 2\,\mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right) \right\|^2 + 2\,\mathbb{E}\left\| \sum_{t=1}^T \omega_t \cdot \delta_t^{k-1} \right\|^2 \tag{110}$$

$$= 2\sum_{t=1}^T \omega_t^2 \cdot \mathbb{E}\left\| \hat{\delta}_t^{k-1} - \delta_t^{k-1} \right\|^2 + 2 \sum_{1\leq s\neq t\leq T} \omega_t \omega_s\, \mathbb{E}\left\langle \hat{\delta}_t^{k-1} - \delta_t^{k-1}, \hat{\delta}_s^{k-1} - \delta_s^{k-1} \right\rangle + 2\,\mathbb{E}\left\| \sum_{t=1}^T \omega_t \delta_t^{k-1} \right\|^2. \tag{111}$$

Since clients sample batches independently, and stochastic gradients are unbiased (Assumption 6′), we have

$$\mathbb{E}\left\langle \hat{\delta}_t^{k-1} - \delta_t^{k-1}, \hat{\delta}_s^{k-1} - \delta_s^{k-1} \right\rangle = 0, \tag{112}$$

thus,

$$T_2 \leq 2\sum_{t=1}^{T} \omega_t^2 \cdot \mathbb{E}\left\| \hat{\delta}_t^{k-1} - \delta_t^{k-1} \right\|^2 + 2\,\mathbb{E}\left\| \sum_{t=1}^{T} \omega_t \delta_t^{k-1} \right\|^2 \tag{113}$$

$$= 2\sum_{t=1}^{T} \omega_t^2\, \mathbb{E}\left\| \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left[\nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right) \right] \right\|^2 \tag{114}$$

$$+ 2\,\mathbb{E}\left\| \sum_{t=1}^{T} \omega_t \delta_t^{k-1} \right\|^2. \tag{115}$$

Using Jensen inequality, we have

$$\left\| \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left[\nabla_u g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) - \nabla_u g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right) \right] \right\|^2 \leq$$

$$\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right) \right\|^2, \tag{116}$$

and since the variance of stochastic gradients is bounded by σ^2 (Assumption 6′), it follows that

$$\mathbb{E}\left\| \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left[\nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j} \right) \right] \right\|^2$$

$$\leq \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \sigma^2 = \sigma^2. \tag{117}$$

Replacing back in the expression of T_2, we have

$$T_2 \leq 2\sum_{t=1}^{T} \omega_t^2 \sigma^2 + 2\,\mathbb{E}\left\| \sum_{t=1}^{T} \omega_t \cdot \delta_t^{k-1} \right\|^2. \tag{118}$$

Finally, since $0 \leq \omega_t \leq 1,\ t \in [T]$ and $\sum_{t=1}^{T} \omega_t = 1$, we have

$$T_2 \leq 2\sigma^2 + 2\,\mathbb{E}\left\| \sum_{t=1}^{T} \omega_t \cdot \delta_t^{k-1} \right\|^2. \tag{119}$$

Having bounded T_1 and T_2, we can replace Eq. (107) and Eq. (119) in Eq. (102), and we get

$$\mathbb{E}\left[g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right] \leq -\frac{\eta_{k-1}}{2} \left\| \nabla_{\mathbf{u}} g^k \left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 + \eta_{k-1}^2 L \sigma^2$$

$$- \frac{\eta_{k-1}}{2} \left(1 - 2L\eta_{k-1} \right) \cdot \mathbb{E}\left\| \sum_{t=1}^{T} \omega_t \cdot \delta_t^{k-1} \right\|^2$$

$$+ \frac{\eta_{k-1}}{2} \mathbb{E}\left\| \nabla_{\mathbf{u}} g^k \left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) - \sum_{t=1}^{T} \omega_t \cdot \delta_t^{k-1} \right\|^2. \tag{120}$$

As $\eta_{k-1} \leq \frac{1}{2\sqrt{2}L} \leq \frac{1}{2L}$, we have

$$\mathbb{E}\Big[g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})\Big] \leq -\frac{\eta_{k-1}}{2}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \eta_{k-1}^2 L\sigma^2 + \frac{\eta_{k-1}}{2}\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) - \sum_{t=1}^{T}\omega_t \delta_t^{k-1}\right\|^2. \tag{121}$$

Replacing $\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) = \sum_{t=1}^{T}\omega_t \cdot \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right)$, and using Jensen inequality to bound the last term in the RHS of Eq. (121), we have

$$\mathbb{E}\Big[g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})\Big] \leq -\frac{\eta_{k-1}}{2}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \eta_{k-1}^2 L\sigma^2 + \frac{\eta_{k-1}}{2}\sum_{t=1}^{T}\omega_t \cdot \underbrace{\mathbb{E}\left\|\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - \delta_t^{k-1}\right\|^2}_{\triangleq T_3}. \tag{122}$$

We now bound the term T_3:

$$T_3 = \mathbb{E}\left\|\nabla_u g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - \delta_t^{k-1}\right\|^2 \tag{123}$$

$$= \mathbb{E}\left\|\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - \sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right)\right\|^2 \tag{124}$$

$$= \mathbb{E}\left\|\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}\left[\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right)\right]\right\|^2 \tag{125}$$

$$\leq \sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}\mathbb{E}\left\|\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right)\right\|^2 \tag{126}$$

$$\leq \sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}L^2\mathbb{E}\left\|\mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j}\right\|^2, \tag{127}$$

where the first inequality follows from Jensen inequality and the second one follow from the L-smoothness of g_t^k (Assumption 5′). We bound now the term $\mathbb{E}\left\|\mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j}\right\|$ for $j \in \{0, \ldots, J-1\}$ and $t \in [T]$,

$$\mathbb{E}\left\|\mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j}\right\|^2 = \mathbb{E}\left\|\mathbf{u}_t^{k-1,j} - \mathbf{u}_t^{k-1,0}\right\|^2 \tag{128}$$

$$= \mathbb{E}\left\|\sum_{l=0}^{j-1}\left(\mathbf{u}_t^{k-1,l+1} - \mathbf{u}_t^{k-1,l}\right)\right\|^2 \tag{129}$$

$$= \mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,l}\right)\right\|^2 \tag{130}$$

$$\leq 2\mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\left[\nabla_u g_t^k\left(\mathbf{u}_t^{k-1,l}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,l}\right) - \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,l}, \mathbf{v}_t^{k-1}\right)\right]\right\|^2 + 2\mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,l}, \mathbf{v}_t^{k-1}\right)\right\|^2 \tag{131}$$

$$= 2\sum_{l=0}^{j-1}\eta_{k-1,l}^2\mathbb{E}\left\|\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,l}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,l}\right) - \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,l}, \mathbf{v}_t^{k-1}\right)\right\|^2$$

$$+ 2\mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2 \tag{132}$$

$$\leq 2\sigma^2\sum_{l=0}^{j-1}\eta_{k-1,l}^2 + 2\mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2, \tag{133}$$

where, in the last two steps, we used the fact that stochastic gradients are unbiased and have bounded variance (Assumption 6′). We bound now the last term in the RHS of Eq. (133),

$$\mathbb{E}\left\|\sum_{l=0}^{j-1}\eta_{k-1,l}\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2 =$$

$$\mathbb{E}\left\|\left(\sum_{l'=0}^{j-1}\eta_{k-1,l'}\right)\cdot\sum_{l=0}^{j-1}\frac{\eta_{k-1,l}}{\sum_{l'=0}^{j-1}\eta_{k-1,l'}}\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2 \tag{134}$$

$$\leq\left(\sum_{l'=0}^{j-1}\eta_{k-1,l'}\right)^2\cdot\sum_{l=0}^{j-1}\frac{\eta_{k-1,l}}{\sum_{l'=0}^{j-1}\eta_{k-1,l'}}\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2 \tag{135}$$

$$=\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\cdot\sum_{l=0}^{j-1}\eta_{k-1,l}\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\right\|^2 \tag{136}$$

$$=\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\cdot\sum_{l=0}^{j-1}\eta_{k-1,l}\mathbb{E}\Big\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,0},\mathbf{v}_t^{k-1}\right)$$
$$-\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,0},\mathbf{v}_t^{k-1}\right)+\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)\Big\|^2 \tag{137}$$

$$\leq 2\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\cdot\sum_{l=0}^{j-1}\eta_{k-1,l}\cdot\left[\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,0},\mathbf{v}_t^{k-1}\right)\right\|^2\right.$$
$$\left.+\,\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,0},\mathbf{v}_t^{k-1}\right)\right\|^2\right] \tag{138}$$

$$=2\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\cdot\sum_{l=0}^{j-1}\eta_{k-1,l}\cdot\left[\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2\right.$$
$$\left.+\,\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}_t^{k-1,l},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2\right] \tag{139}$$

$$\leq 2\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\sum_{l=0}^{j-1}\eta_{k-1,l}\left[\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2+L^2\mathbb{E}\left\|\mathbf{u}_t^{k-1,l}-\mathbf{u}^{k-1}\right\|^2\right] \tag{140}$$

$$=2L^2\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)\sum_{l=0}^{j-1}\eta_{k-1,l}\cdot\mathbb{E}\left\|\mathbf{u}_t^{k-1,l}-\mathbf{u}^{k-1}\right\|^2$$
$$+\,2\left(\sum_{l=0}^{j-1}\eta_{k-1,l}\right)^2\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2, \tag{141}$$

where the first inequality is obtained using Jensen inequality, and the last one is a result of the L-smoothness of g_t (Assumption 5′). Replacing Eq. (141) in Eq. (133), we have

$$\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}\cdot\mathbb{E}\left\|\mathbf{u}^{k-1}-\mathbf{u}_t^{k-1,j}\right\|^2\leq 2\sigma^2\left(\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}}{\eta_{k-1}}\cdot\sum_{l=0}^{j-1}\eta_{k-1,l}^2\right)$$

$$+ 4L^2 \sum_{j=0}^{J-1} \left(\frac{\eta_{k-1,j}}{\eta_{k-1}} \sum_{l=0}^{j-1} \eta_{k-1,l} \right) \cdot \left(\sum_{l=0}^{j-1} \eta_{k-1,l} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,l} - \mathbf{u}_t^{k-1} \right\|^2 \right)$$
$$+ 4 \left(\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left(\sum_{l=0}^{j-1} \eta_{k-1,l} \right)^2 \right) \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{142}$$

Since $\sum_{l=0}^{j-1} \eta_{k-1,l} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,l} - \mathbf{u}_t^{k-1} \right\|^2 \leq \sum_{j=0}^{J-1} \eta_{k-1,j} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,j} - \mathbf{u}_t^{k-1} \right\|^2$, we have

$$\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j} \right\|^2 \leq 2\sigma^2 \left(\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \sum_{l=0}^{j-1} \eta_{k-1,l}^2 \right)$$
$$+ 4L^2 \left(\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \sum_{l=0}^{j-1} \eta_{k-1,l} \right) \cdot \left(\sum_{j=0}^{J-1} \eta_{k-1,j} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,j} - \mathbf{u}^{k-1} \right\|^2 \right)$$
$$+ 4 \left(\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \left(\sum_{l=0}^{j-1} \eta_{k-1,l} \right)^2 \right) \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{143}$$

We use Lemma J.11 to simplify the last expression, obtaining

$$\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j} \right\|^2 \leq 2\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\}$$
$$+ 4\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 + 4\eta_{k-1} L^2 \cdot \sum_{j=0}^{J-1} \eta_{k-1,j} \mathbb{E} \left\| \mathbf{u}_t^{k-1,j} - \mathbf{u}^{k-1} \right\|^2 . \tag{144}$$

Rearranging the terms, we have

$$\left(1 - 4\eta_{k-1}^2 L^2\right) \cdot \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}^{k-1} - \mathbf{u}_t^{k-1,j} \right\|^2 \leq 2\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} + 4\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{145}$$

Finally, replacing Eq. (145) into Eq. (127), we have

$$\left(1 - 4\eta_{k-1}^2 L^2\right) \cdot T_3 \leq 2\sigma^2 L^2 \cdot \left(\sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right) + 4\eta_{k-1}^2 L^2 \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{146}$$

For η_{k-1} small enough, in particular if $\eta_{k-1} \leq \frac{1}{2\sqrt{2}L}$, then $\frac{1}{2} \leq 1 - 4\eta_{k-1}^2 L^2$, thus

$$\frac{T_3}{2} \leq 2\sigma^2 L^2 \cdot \left(\sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right) + 4\eta_{k-1}^2 L^2 \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{147}$$

Replacing the bound of T_3 from Eq. (147) into Eq. (122), we have obtained

$$\mathbb{E} \left[g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right] \leq -\frac{\eta_{k-1}}{2} \mathbb{E} \left\| \nabla_{\mathbf{u}} g^k \left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 + 2\eta_{k-1} L \left(\sum_{j=0}^{J-1} \eta_{k-1,j}^2 L + \eta_{k-1} \right) \cdot \sigma^2$$
$$+ 4\eta_{k-1}^3 L^2 \sum_{t=1}^{T} \omega_t \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1} \right) \right\|^2 . \tag{148}$$

Using Assumption 7′, we have

$$\mathbb{E}\Big[g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})\Big] \leq -\frac{\eta_{k-1}}{2}\,\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + 4\eta_{k-1}^3 L^2\beta^2 \cdot \mathbb{E}\left\|\sum_{t=1}^{T}\omega_t \cdot \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right)\right\|^2 + 2\eta_{k-1}L\left(\sum_{j=0}^{J-1}\eta_{k-1,j}^2 L + \eta_{k-1}\right)\cdot\sigma^2 + 4\eta_{k-1}^3 L^2 G^2. \tag{149}$$

Dividing by η_{k-1}, we get

$$\mathbb{E}\Big[\frac{g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{\eta_{k-1}}\Big] \leq \frac{8\eta_{k-1}^2 L^2\beta^2 - 1}{2}\,\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + 2\eta_{k-1}L\left(\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}^2}{\eta_{k-1}}L + 1\right)\cdot\sigma^2 + 4\eta_{k-1}^2 L^2 G^2. \tag{150}$$

For η_{k-1} small enough, if $\eta_{k-1} \leq \frac{1}{4L\beta}$, then $8\eta_{k-1}^2 L^2\beta^2 - 1 \leq \frac{1}{2}$. Thus,

$$\mathbb{E}\Big[\frac{g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{\eta_{k-1}}\Big] \leq -\frac{1}{4}\,\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + 2\eta_{k-1}L\left(\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}^2}{\eta_{k-1}}L + 1\right)\cdot\sigma^2 + 4\eta_{k-1}^2 L^2 G^2. \tag{151}$$

Since for $t \in [T]$, g_t^k is a pseudo first-order surrogate of f_t near $\left\{\mathbf{u}^{k-1}, v_t^{k-1}\right\}$, we have (see Def. 1)

$$g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) = f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right), \tag{152}$$

$$\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) = \nabla_u f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right), \tag{153}$$

$$g_t^k\left(\mathbf{u}^{k}, \mathbf{v}_t^{k-1}\right) = g_t^k\left(\mathbf{u}^{k}, \mathbf{v}_t^{k}\right) + d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^{k}\right). \tag{154}$$

Multiplying by ω_t and summing over $t \in [T]$, we have

$$g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) = f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right), \tag{155}$$

$$\nabla_u g^k\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) = \nabla_u f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right), \tag{156}$$

$$g^k\left(\mathbf{u}^{k}, \mathbf{v}_{1:T}^{k-1}\right) = g^k\left(\mathbf{u}^{k}, \mathbf{v}_{1:T}^{k}\right) + \sum_{t=1}^{T}\omega_t \cdot d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^{k}\right). \tag{157}$$

Replacing Eq. (155), Eq. (156) and Eq. (157) in Eq. (151), we have

$$\mathbb{E}\left[\frac{g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^{k}) - f(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{\eta_{k-1}}\right] \leq -\frac{1}{4}\,\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 - \frac{1}{\eta_{k-1}}\sum_{t=1}^{T}\omega_t \cdot d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^{k}\right) + 2\eta_{k-1}L\left(\left\{\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}^2}{\eta_{k-1}}\right\}L + 1\right)\cdot\sigma^2 + 4\eta_{k-1}^2 L^2 G^2. \tag{158}$$

Using again Def. 1, we have

$$g^k(\mathbf{u}^k, \mathbf{v}_{1:T}^k) \geq f(\mathbf{u}^k, \mathbf{v}_{1:T}^k), \tag{159}$$

thus,

$$\mathbb{E}\left[\frac{f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) - f(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{\eta_{k-1}}\right] \leq$$
$$-\frac{1}{4}\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 - \frac{1}{\eta_{k-1}}\sum_{t=1}^{T}\omega_t \cdot d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^k\right)$$
$$+ 2\eta_{k-1}L\left(\sum_{j=0}^{J-1}\frac{\eta_{k-1,j}^2}{\eta_{k-1}}L + 1\right)\cdot\sigma^2 + 4\eta_{k-1}^2L^2G^2. \tag{160}$$

□

Lemma J.2. *For $k \geq 0$ and $t \in [T]$, the iterates of Alg. 5 verify*

$$0 \leq d_{\mathcal{V}}\left(\mathbf{v}_t^{k+1}, \mathbf{v}_t^k\right) \leq f_t\left(\mathbf{u}^k, \mathbf{v}_t^k\right) - f_t(\mathbf{u}^k, \mathbf{v}_t^{k+1}) \tag{161}$$

Proof. Since $\mathbf{v}_t^{k+1} \in \arg\min_{v\in V} g_t^k\left(\mathbf{u}^{k-1}, v\right)$, and g_t^k is a pseudo first-order surrogate of f_t near $\{\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\}$, we have

$$g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^k\right) = d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^k\right), \tag{162}$$

thus,

$$f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) - f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^k\right) \geq d_{\mathcal{V}}\left(\mathbf{v}_t^{k-1}, \mathbf{v}_t^k\right), \tag{163}$$

where we used the fact that

$$g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right) = f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right), \tag{164}$$

and,

$$g_t^k\left(\mathbf{u}^{k-1}, \mathbf{v}_t^k\right) \geq f_t\left(\mathbf{u}^{k-1}, \mathbf{v}_t^k\right). \tag{165}$$

□

Theorem 3.2′. *Under Assumptions 4′–7′, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, the iterates of federated surrogate optimization (Alg. 5) satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \qquad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\Delta_v f(\mathbf{u}^k, \mathbf{v}_{1:T}^k)\right] \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right), \tag{91}$$

where the expectation is over the random batches samples, and $\Delta_v f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) \triangleq f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^k\right) - f\left(\mathbf{u}^k, \mathbf{v}_{1:T}^{k+1}\right) \geq 0$.

Proof. For K large enough, $\eta = \frac{a_0}{\sqrt{K}} \leq \frac{1}{J}\min\left\{\frac{1}{2\sqrt{2}L}, \frac{1}{4L\beta}\right\}$, thus the assumptions of Lemma J.1 are satisfied. Lemma J.1 and non-negativity of $d_{\mathcal{V}}$ lead to

$$\mathbb{E}\Big[\frac{f(\mathbf{u}^k, \mathbf{v}_{1:T}^k) - f(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1})}{J\eta}\Big] \leq -\frac{1}{4}\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2$$
$$+ 2\eta L\left(\eta L + 1\right)\cdot\sigma^2 + 4J^2\eta^2L^2G^2. \tag{166}$$

Rearranging the terms and summing for $k \in [K]$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2$$
$$\leq 4\mathbb{E}\Big[\frac{f(\mathbf{u}^0, \mathbf{v}_{1:T}^0) - f(\mathbf{u}^K, \mathbf{v}_{1:T}^K)}{J\eta K}\Big] + 8\frac{\eta L\left(\eta L + 1\right)\cdot\sigma^2 + 2J^2\eta^2L^2G^2}{K} \tag{167}$$
$$\leq 4\mathbb{E}\Big[\frac{f(\mathbf{u}^0, \mathbf{v}_{1:T}^0) - f^*}{J\eta K}\Big] + 8\frac{\eta L\left(\eta L + 1\right)\cdot\sigma^2 + 2J^2\eta^2L^2G^2}{K} \tag{168}$$

Thus,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2=\mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \tag{169}$$

To prove the second part of Eq. (91), we first decompose $\Delta_{\mathbf{v}}\triangleq f\left(\mathbf{u}^k,\mathbf{v}_{1:T}^k\right)-f\left(\mathbf{u}^k,\mathbf{v}_{1:T}^{k+1}\right)\geq 0$as follow,

$$\Delta_{\mathbf{v}}=\underbrace{f\left(\mathbf{u}^k,\mathbf{v}_{1:T}^k\right)-f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)}_{\triangleq T_1^k}+\underbrace{f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)-f\left(\mathbf{u}^k,\mathbf{v}_{1:T}^{k+1}\right)}_{\triangleq T_2^k}. \tag{170}$$

Using Eq. (100), and Eq. (169), it follows that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[T_1^k\right]\leq\mathcal{O}\left(\frac{1}{K}\right). \tag{171}$$

We use the fact that f is $2L$-smooth (Lemma J.12) w.r.t. u and Cauchy-Schwartz inequality. Thus, for $k>0$, we write

$$T_2^k=f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)-f\left(\mathbf{u}^k,\mathbf{v}_{1:T}^{k+1}\right) \tag{172}$$

$$\leq\left\|\nabla_{\mathbf{u}}f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)\right\|\cdot\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|+2L^2\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|^2. \tag{173}$$

Summing over k and taking expectation:

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[T_2^k\right]&\leq\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_{\mathbf{u}}f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)\right\|\cdot\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|\right]\\&\quad+\frac{1}{K}\sum_{k=1}^{K}2L^2\,\mathbb{E}\left[\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|^2\right] \qquad (174)\\&\leq\frac{1}{K}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_{\mathbf{u}}f\left(\mathbf{u}^{k+1},\mathbf{v}_{1:T}^{k+1}\right)\right\|^2\right]}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|^2\right]}\\&\quad+\frac{1}{K}\sum_{k=1}^{K}2L^2\,\mathbb{E}\left[\left\|\mathbf{u}^{k+1}-\mathbf{u}^k\right\|^2\right], \qquad (175)\end{aligned}$$

where the second inequality follows from Cauchy-Schwarz inequality. From Eq. (145), with $\eta_{k-1}=J\eta$, we have for $t\in[T]$

$$\mathbb{E}\left\|\mathbf{u}^k-\mathbf{u}_t^{k-1,J}\right\|^2\leq 4\sigma^2J\eta^2+8J^3\eta^2\cdot\mathbb{E}\left\|\nabla_u g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2. \tag{176}$$

Multiplying the previous by ω_t and summing for $t\in[T]$, we have

$$\sum_{t=1}^{T}\omega_t\cdot\mathbb{E}\left\|\mathbf{u}^{k-1}-\mathbf{u}_t^{k-1,J}\right\|^2\leq 4J^2\sigma^2\eta^2+8J^3\eta^2\cdot\sum_{t=1}^{T}\omega_t\mathbb{E}\left\|\nabla_u g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2. \tag{177}$$

Using Assumption 7′, it follows that

$$\sum_{t=1}^{T}\omega_t\mathbb{E}\left\|\mathbf{u}^{k-1}-\mathbf{u}_t^{k-1,J}\right\|^2\leq 4J^2\eta^2\left(2JG^2+\sigma^2\right)+8J^3\eta^2\beta^2\mathbb{E}\left\|\sum_{t=1}^{T}\omega_t\nabla_u g_t^k\left(\mathbf{u}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2. \tag{178}$$

Finally using Jensen inequality and the fact that g_t^k is a pseudo-first order of f_t near $\left\{u^{k-1},v_t^{k-1}\right\}$, we have

$$\mathbb{E}\left\|\mathbf{u}^{k-1}-\mathbf{u}^k\right\|^2\leq 4J^2\eta^2\left(2JG^2+\sigma^2\right)+8J^3\eta^2\beta^2\mathbb{E}\left\|\nabla_u f\left(\mathbf{u}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2. \tag{179}$$

From Eq. (169) and $\eta \leq \mathcal{O}(1/\sqrt{K})$, we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\mathbf{u}^{k-1}-\mathbf{u}^{k}\right\|^{2} \leq \mathcal{O}\left(1\right), \tag{180}$$

Replacing the last inequality in Eq. (175) and using again Eq. (169), we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[T_2^k\right] \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right). \tag{181}$$

Combining Eq. (171) and Eq. (181), it follows that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\Delta_v f(\mathbf{u}^k, \mathbf{v}_{1:T}^k)\right] \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right). \tag{182}$$

□

J.1.3. Proof of Theorem 3.2

In this section f denotes the negative log-likelihood function defined in Eq. (6). Moreover, we introduce the negative log-likelihood at client t as follows

$$f_t(\Theta,\Pi) \triangleq -\frac{\log p(\mathcal{S}_t|\Theta,\Pi)}{n} \triangleq -\frac{1}{n_t}\sum_{i=1}^{n_t}\log p(s_t^{(i)}|\Theta,\pi_t). \tag{183}$$

Theorem 3.2. *Under Assumptions 1–7, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, `FedEM`'s iterates satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_\Theta f\left(\Theta^k,\Pi^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \tag{11}$$

$$\frac{1}{K}\sum_{k=1}^{K}\Delta_\Pi f(\Theta^k,\Pi^k) \leq \mathcal{O}\left(\frac{1}{K^{3/4}}\right), \tag{12}$$

where the expectation is over the random batches samples, and $\Delta_\Pi f(\Theta^k,\Pi^k) \triangleq f\left(\Theta^k,\Pi^k\right) - f\left(\Theta^k,\Pi^{k+1}\right) \geq 0$.

Proof. We prove this result as a particular case of Theorem 3.2′. To this purpose, in this section, we consider that $\mathcal{V} \triangleq \Delta^M$, $u = \Theta \in \mathbb{R}^{dM}$, $v_t = \pi_t$, and $\omega_t = n_t/n$ for $t \in [T]$. For $k > 0$, we define g_t^k as follow,

$$g_t^k\Big(\Theta,\pi_t\Big) = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)}=m\right)\cdot\Bigg(l\left(h_{\theta_m}(\mathbf{x}_t^{(i)}),y_t^{(i)}\right) - \log p_m(\mathbf{x}_t^{(i)}) - \log\pi_t + \log q_t^k\left(z_t^{(i)}=m\right) - c\Bigg), \tag{184}$$

where c is the same constant appearing in Assumption 3, Eq. (3). With this definition, it is easy to check that the federated surrogate optimization algorithm (Alg. 5) reduces to `FedEM` (Alg. 7). Theorem 3.2 follows immediately from Theorem 3.2′, once we verify that $\left(g_t^k\right)_{1\leq t\leq T}$ verify the assumptions of Theorem 3.2′.

Assumption 4′, Assumption 6′, and Assumption 7′ follow directly from Assumption 4, Assumption 6, and Assumption 7, respectively. Lemma J.3 shows that for $k > 0$, g^k is smooth w.r.t. Θ and then Assumption 5′ is satisfied. Finally, Lemmas J.4–J.6 show that for $t \in [T]$ g_t^k is a partial first-order surrogate of f_t w.r.t. Θ near $\left\{\Theta^{k-1},\pi_t\right\}$ with $d_{\mathcal{V}}(\cdot,\cdot) = \mathcal{KL}(\cdot\|\cdot)$. □

Lemma J.3. *Under Assumption 5, for $t \in [T]$ and $k > 0$, g_t^k is L-smooth w.r.t Θ.*

Proof. g_t^k is a convex combination of L-smooth function $\theta \mapsto l(\theta; s_t^{(i)}),\ i \in [n_t]$. Thus it is also L-smooth. □

Lemma J.4. *Suppose that Assumptions 1–3, hold. Then, for $t \in [T]$, $\Theta \in \mathbb{R}^{M\times d}$ and $\pi_t \in \Delta^M$*

$$r_t^k(\Theta, \pi_t) \triangleq g_t^k(\Theta, \pi_t) - f_t(\Theta, \pi_t) = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{KL}\left(q_t^k\left(z_i^{(t)}\right) \| p_t\left(z_i^{(t)} | s_i^{(t)}, \Theta, \pi_t\right)\right),$$

where $\mathcal{KL}$ is Kullback–Leibler divergence

Proof. Let $k > 0$ and $t \in [T]$, and consider $\Theta \in \mathbb{R}^{M\times d}$ and $\pi_t \in \Delta^M$, then

$$\begin{aligned}
g_t^k\Big(\Theta, \pi_t\Big) &= \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \cdot \Bigg(l\left(h_{\theta_m}(\mathbf{x}_t^{(i)}), y_t^{(i)}\right) - \log p_m(\mathbf{x}_t^{(i)}) - \log \pi_t \\
&\qquad + \log q_t^k\left(z_t^{(i)} = m\right) - c\Bigg), && (185)\\
&= \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \cdot \Bigg(- \log p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) - \log p_m(\mathbf{x}_t^{(i)}) - \log \pi_t \\
&\qquad + \log q_t^k\left(z_t^{(i)} = m\right)\Bigg) && (186)\\
&= \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \cdot \Bigg(- \log p_m\left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m\right) \cdot p_m(\mathbf{x}_t^{(i)}) \cdot p_t\left(z_t^{(i)} = m\right) \\
&\qquad + \log q_t^k\left(z_t^{(i)} = m\right)\Bigg) && (187)\\
&= \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \cdot \left(\log q_t^k\left(z_t^{(i)} = m\right) - \log p_t\left(s_t^{(i)}, z_t^{(i)} = m \middle| \Theta, \pi_t\right)\right) && (188)\\
&= \frac{1}{n_t}\sum_{t=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \log \frac{q_t^k\left(z_t^{(i)} = m\right)}{p_t\left(s_t^{(i)}, z_t^{(i)} = m | \Theta, \pi_t\right)}. && (189)
\end{aligned}$$

Thus,

$$\begin{aligned}
r_t^k\Big(\Theta, \pi_t\Big) &\triangleq g_t^k(\Theta, \pi_t) - f_t(\Theta, \pi_t) && (190)\\
&= -\frac{1}{n_t}\sum_{t=1}^{n_t}\sum_{m=1}^{M}\left(q_t^k\left(z_t^{(i)} = m\right) \cdot \log \frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m | \Theta, \pi_t\right)}{q_t^k\left(z_t^{(i)} = m\right)}\right) \\
&\qquad + \frac{1}{n_t}\sum_{i=1}^{n_t} \log p_t\left(s_t^{(i)} | \Theta, \pi_t\right) && (191)\\
&= \frac{1}{n_t}\sum_{t=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right)\Bigg(\log p_t\left(s_t^{(i)} | \Theta, \pi_t\right) \\
&\qquad - \log \frac{p_t\left(s_t^{(i)}, z_t^{(i)} = m | \Theta, \pi_t\right)}{q_t^k\left(z_t^{(i)} = m\right)}\Bigg) && (192)\\
&= \frac{1}{n_t}\sum_{t=1}^{n_t}\sum_{m=1}^{M} q_t^k\left(z_t^{(i)} = m\right) \log \frac{p_t\left(s_t^{(i)} | \Theta, \pi_t\right) \cdot q_t^k\left(z_t^{(i)} = m\right)}{p_t\left(s_t^{(i)}, z_t^{(i)} = m | \Theta, \pi_t\right)} && (193)
\end{aligned}$$

$$= \frac{1}{n_t} \sum_{t=1}^{n_t} \sum_{m=1}^{M} q_t^k \left(z_t^{(i)} = m \right) \cdot \log \frac{q_t^k \left(z_t^{(i)} = m \right)}{p_t \left(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t \right)}. \tag{194}$$

Thus,

$$r_t^k \left(\Theta, \pi_t \right) = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{KL} \left(q_t^k(\cdot) \| p_t(\cdot | s_i^{(t)}, \Theta, \pi_t) \right) \geq 0 \tag{195}$$

□

The following lemma shows that g_t^k and g^k (as defined in Eq. 99) satisfy the first two properties in Definition 1.

Lemma J.5. *Suppose that Assumptions 1–3 and Assumption 5 hold. For all $k \geq 0$ and $t \in [T]$, g_t^k is a majorant of f_t and $r_t^k \triangleq g_t^k - f_t$ is L-smooth in Θ. Moreover $r_t^k \left(\Theta^{k-1}, \pi_t^{k-1} \right) = 0$ and $\nabla_\Theta r_t^k \left(\Theta^{k-1}, \pi_t^{k-1} \right) = 0$.*

The same holds for g^k, i.e., g^k is a majorant of f, $r^k \triangleq g^k - f$ is L-smooth in Θ, $r^k \left(\Theta^{k-1}, \Pi^{k-1} \right) = 0$ and $\nabla_\Theta r^k \left(\Theta^{k-1}, \Pi^{k-1} \right) = 0$

Proof. For $t \in [T]$, consider $\Theta \in \mathbb{R}^{M \times d}$ and $\pi_t \in \Delta^M$, we have (Lemma J.4)

$$r_t^k \left(\Theta, \pi_t \right) \triangleq g_t^k \left(\Theta, \pi_t \right) - f_t \left(\Theta, \pi_t \right) = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{KL} \left(q_t^k \left(z_i^{(t)} \right) \| p_t \left(z_t^{(i)} | s_t^{(i)}, \Theta, \pi_t \right) \right) \tag{196}$$

Since $\mathcal{KL}$ divergence is non-negative, it follows that g_t^k is a majorant of f_t, i.e.,

$$\forall\, \Theta \in \mathbb{R}^{M \times d},\ \pi_t \in \Delta^M;\ \ g_t^k \left(\Theta, \pi \right) \geq f_t \left(\Theta, \pi_t \right) \tag{197}$$

Moreover since, $q_t^k \left(z_t^{(i)} \right) = p_t \left(z_t^{(i)} | s_t^{(i)}, \Theta^{k-1}, \pi_t^{k-1} \right)$ for $k > 0$, it follows that

$$r_t^k \left(\Theta^{k-1}, \pi_t^{k-1} \right) = 0 \tag{198}$$

For $i \in [n_t]$ and $m \in [M]$, from Eq. 79, we have

$$p_t \left(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t \right) = \frac{p_m \left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_m \right) \times \pi_{tm}}{\sum_{m'=1}^{M} p_{m'} \left(y_t^{(i)} | \mathbf{x}_t^{(i)}, \theta_{m'} \right) \times \pi_{tm'}} \tag{199}$$

$$= \frac{\exp \left[-l \left(h_{\theta_m}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) \right] \times \pi_{tm}}{\sum_{m'=1}^{M} \exp \left[-l \left(h_{\theta_{m'}}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) \right] \times \pi_{tm'}} \tag{200}$$

$$= \frac{\exp \left[-l \left(h_{\theta_m}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) + \log \pi_{tm} \right]}{\sum_{m'=1}^{M} \exp \left[-l \left(h_{\theta_{m'}}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) + \log \pi_{tm'} \right]}, \tag{201}$$

Thus,

$$\mathcal{KL} \left(q_t^k \left(z_i^{(t)} \right) \| p_t \left(z_t^{(i)} | s_t^{(i)}, \Theta, \pi_t \right) \right) = \underbrace{\sum_{m=1}^{M} q_t^k \left(z_i^{(t)} \right) \cdot \left(\log q_t^k \left(z_i^{(t)} \right) + l \left(h_{\theta_m}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) - \log \pi_{tm} \right)}_{L\text{-smooth, because convex combination of } L\text{-smooth functions}}$$
$$+ \log \left(\sum_{m'=1}^{M} \exp \left[-l \left(h_{\theta_{m'}}(\mathbf{x}_t^{(i)}), y_t^{(i)} \right) + \log \pi_{tm'} \right] \right). \tag{202}$$

For ease of notation, we introduce

$$l_i(\theta) \triangleq l \left(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)} \right), \qquad \theta \in \mathbb{R}^d,\ m \in [M],\ i \in [n_t], \tag{203}$$

$$\gamma_m(\Theta) \triangleq p_t\left(z_t^{(i)} = m | s_t^{(i)}, \Theta, \pi_t\right), \qquad m \in [M], \tag{204}$$

and,

$$\varphi_i(\Theta) \triangleq \log\left(\sum_{m'=1}^{M} -\exp\left[-l\left(h_{\theta_{m'}}(\mathbf{x}_t^{(i)}), y_t^{(i)}\right) + \log \pi_{tm'}\right]\right), \qquad i \in [n_t]. \tag{205}$$

To prove the L-smoothness of r_t^k, it is enough to prove that $\varphi_i,\ i \in [n_t]$ is L-smooth. For $i \in [n_t]$, function l_i is differentiable because smooth (Assum. 5), thus, φ_i is differentiable and its gradient is given by

$$\nabla_{\theta_m} \varphi_i(\Theta) = \gamma_m(\Theta) \cdot \nabla l_i(\theta_m), \quad m \in [M] \tag{206}$$

We also know that $\gamma_m,\ m \in [M]$ is differentiable as the composition of the softmax function and the function $\{\Theta \mapsto -l_i(\Theta) + \log \pi_{tm}\}$. Its gradient is given by

$$\begin{cases} \nabla_{\theta_m} \gamma_m(\Theta) = -\gamma_m(\Theta)(1 - \gamma_m(\Theta)) \cdot \nabla l_i(\theta_m) \\ \nabla_{\theta_{m'}} \gamma_m(\Theta) = \gamma_m(\Theta)\gamma_{m'}(\Theta) \cdot \nabla l_i(\theta_m); \qquad m' \neq m \end{cases} \tag{207}$$

We use $\mathbf{H}(\varphi_i(\Theta)) \in \mathbb{R}^{dM \times dM}$ (resp. $\mathbf{H}(l_i(\theta))$) to denote the hessian of φ (resp. l_i) at Θ (resp. θ). The hessian of φ_i is a block matrix given by

$$\begin{cases} \Big(\mathbf{H}(\varphi_i(\Theta))\Big)_{m,m} = -\gamma_m(\Theta) \cdot (1 - \gamma_m(\Theta)) \cdot (\nabla l_i(\theta_m)) \cdot (\nabla l_i(\theta_m))^\intercal + \gamma_m(\Theta) \cdot \mathbf{H}(l_i(\theta_m)) \\ \Big(\mathbf{H}(\varphi_i(\Theta))\Big)_{m,m'} = \gamma_m(\Theta) \cdot \gamma_{m'}(\Theta) \cdot (\nabla l_i(\theta_{m'})) \cdot (\nabla l_i(\theta_m))^\intercal; \qquad m' \neq m. \end{cases} \tag{208}$$

We introduce the block matrix $\tilde{\mathbf{H}} \in \mathbb{R}^{dM \times dM}$, defined by

$$\begin{cases} \tilde{\mathbf{H}}_{m,m} = -\gamma_m(\Theta) \cdot (1 - \gamma_m(\Theta)) \cdot (\nabla l_i(\theta_m)) \cdot (\nabla l_i(\theta_m))^\intercal \\ \tilde{\mathbf{H}}_{m,m'} = \gamma_m(\Theta) \cdot \gamma_m(\Theta) \cdot (\nabla_\theta l_i(\theta_m)) \cdot (\nabla l_i(\theta_{m'}))^\intercal; \qquad m' \neq m, \end{cases} \tag{209}$$

Eq. (208) can be written,

$$\begin{cases} \Big(\mathbf{H}(\varphi_i(\Theta))\Big)_{m,m} - \tilde{\mathbf{H}}_{m,m} = \gamma_m(\Theta) \cdot \mathbf{H}(l_i(\theta_m)) \\ \Big(\mathbf{H}(\varphi_i(\Theta))\Big)_{m,m'} - \tilde{\mathbf{H}}_{m,m'} = 0; \qquad m' \neq m. \end{cases} \tag{210}$$

We recall that a twice differentiable function is L smooth if and only if the eigenvalues of its Hessian are smaller then L, see e.g., (Nesterov, 2003, Lemma 1.2.2) or (Bubeck, 2015, Section 3.2). Since l_i is L-smooth (Assumption 5), we have for $\theta \in \mathbb{R}^d$,

$$\mathbf{H}(l_i(\theta)) \preccurlyeq L \cdot I_d. \tag{211}$$

Using Lemma J.15, we can conclude that matrix $\tilde{\mathbf{H}}$ is semi-definite negative, thus

$$\mathbf{H}(\varphi_i(\Theta)) \preccurlyeq L \cdot I_{dM}. \tag{212}$$

The last equation proves that φ_i is L-smooth. Thus r_t^k is L-smooth with respect to Θ as the average of L-smooth function, i.e.,

$$r_t^k(\Theta) = \frac{1}{n_t} \sum_{i=1}^{n_t} \varphi_i(\Theta)$$

Moreover, since $r_t^k(\Theta^{k-1}, \pi_t^{k-1}) = 0$ and $\forall \Theta, \Pi;\ \ r_t^k(\Theta, \pi_t) \geq 0$, it follows that Θ^{k-1} is a minimizer of $\left\{\Theta \mapsto r_t^k\left(\Theta, \pi_t^{k-1}\right)\right\}$. Thus, $\nabla_\Theta r_t^k(\Theta^{k-1}, \pi_t^{k-1}) = 0$.

For $\Theta \in \mathbb{R}^{M\times d}$ and $\Pi \in \Delta^{T\times M}$, we have

$$r^k(\Theta,\Pi) \triangleq g^k(\Theta,\Pi) - f(\Theta,\Pi) \tag{213}$$

$$\triangleq \sum_{t=1}^{T} \frac{n_t}{n} \cdot \left[g_t^k(\Theta,\pi_t) - f_t(\Theta,\pi_t)\right] \tag{214}$$

$$= \sum_{t=1}^{T} \frac{n_t}{n} r_t^k(\Theta,\pi_t). \tag{215}$$

We see that r^k is a weighted average of $\left(r_t^k\right)_{1\leq t\leq T}$. Thus, r_t^k is L-smooth in Θ, $r^k(\Theta,\Pi) \geq 0$, moreover $r_t^k\left(\Theta^{k-1},\Pi^{k-1}\right) = 0$ and $\nabla_\Theta r_t^k\left(\Theta^{k-1},\Pi^{k-1}\right) = 0$. □

The following lemma shows that g_t^k and g^k satisfy the third property in Definition 1.

Lemma J.6. *Suppose that Assumption 1 holds and consider $\Theta \in \mathbb{R}^{M\times d}$ and $\Pi \in \Delta^{T\times M}$, for $k > 0$, the iterates of Alg. 5 verify*

$$g^k(\Theta,\Pi) = g^k\left(\Theta,\Pi^k\right) + \sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k,\pi_t\right).$$

Proof. For $t \in [T]$ and $k > 0$, consider $\Theta \in \mathbb{R}^{M\times d}$ and $\pi_t \in \Delta^M$ such that $\forall m \in [M]; \pi_{tm} \neq 0$, we have

$$g_t^k(\Theta,\pi_t) - g_t^k\left(\Theta,\pi_t^k\right) = \sum_{m=1}^{M} \underbrace{\left\{\frac{1}{n_t}\sum_{i=1}^{n_t} q_t^k\left(z_t^{(i)} = m\right)\right\}}_{=\pi_{tm}^k \text{ (Prop. 3.1)}} \times \left(\log \pi_{tm}^k - \log \pi_{tm}\right) \tag{216}$$

$$= \sum_{m=1}^{M} \pi_{tm}^k \log \frac{\pi_{tm}^k}{\pi_{tm}} \tag{217}$$

$$= \mathcal{KL}\left(\pi_t^k,\pi_t\right). \tag{218}$$

We multiply by $\frac{n_t}{n}$ and some for $t \in [T]$. It follows that

$$g^k\left(\Theta,\Pi^k\right) + \sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k,\pi_t\right) = g^k(\Theta,\Pi). \tag{219}$$

□

J.2. Fully Decentralized Setting

J.2.1. Additional Notations

Remark 2. *For convenience and without loss of generality, we suppose in this section that* $\omega_t = \frac{1}{T}$.

We introduce the following matrix notation:

$$\mathbf{U}^k \triangleq \left[\mathbf{u}_1^k, \dots, \mathbf{u}_T^k\right] \in \mathbb{R}^{d_u \times T} \tag{220}$$

$$\bar{\mathbf{U}}^k \triangleq \left[\bar{\mathbf{u}}^k, \dots, \bar{\mathbf{u}}^k\right] \in \mathbb{R}^{d_u \times T} \tag{221}$$

$$\partial g^k\left(\mathbf{U}^k, \mathbf{v}_{1:T}^k; \xi^k\right) \triangleq \left[\nabla_{\mathbf{u}} g_1^k\left(\mathbf{u}_1^k, \mathbf{v}_1^k; \xi_1^k\right), \dots, \nabla_{\mathbf{u}} g_T^k\left(\mathbf{u}_T^k, \mathbf{v}_T^k; \xi_T^k\right)\right] \in \mathbb{R}^{d_u \times T} \tag{222}$$

where $\bar{\mathbf{u}}^k = \frac{1}{T}\sum_{t=1}^T \mathbf{u}_t^k$ and $\mathbf{v}_{1:T}^k = \left(\mathbf{v}_t^k\right)_{1 \le t \le T} \in \mathcal{V}^T$

We denote by $\mathbf{u}_t^{k-1,j}$ the j-th iterate of the local solver at global iteration k at client $t \in [T]$, and by $\mathbf{U}^{k-1,j}$ the matrix whose column t is $\mathbf{u}_t^{k-1,j}$, thus,

$$\mathbf{u}_t^{k-1,0} = \mathbf{u}_t^{k-1}; \qquad \mathbf{U}^{k-1,0} = \mathbf{U}^{k-1} \tag{223}$$

and,

$$\mathbf{u}_t^k = \sum_{s=1}^T w_{ts}^{k-1} \mathbf{u}_s^{k-1,J}; \qquad \mathbf{U}^k = \mathbf{U}^{k-1,J} W^{k-1} \tag{224}$$

Using this notation, the updates of Alg. 9 can be summarized as

$$\mathbf{U}^k = \left[\mathbf{U}^{k-1} - \sum_{j=0}^{J-1} \eta_{k-1,j} \partial g^k\left(\mathbf{U}^{k-1,j}, \mathbf{v}_{1:T}; \xi^{k-1,j}\right)\right] W^{k-1} \tag{225}$$

We also define, the normalized update of local solver at client $t \in [T]$as,

$$\hat{\delta}_t^{k-1} \triangleq -\frac{\mathbf{u}_t^{k-1,J} - \mathbf{u}_t^{k-1,0}}{\eta_{k-1}} = \frac{\sum_{j=0}^{J-1} \eta_{k-1,j} \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^k; \xi_t^{k-1,j}\right)}{\sum_{j=0}^{J-1} \eta_{k-1,j}} \tag{226}$$

and,

$$\delta_t^{k-1} \triangleq \frac{\sum_{j=0}^{J-1} \eta_{k-1,j} \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^k\right)}{\eta_{k-1}} \tag{227}$$

Because clients updates are independent, and stochastic gradient are unbiased, it is clear that

$$\mathbb{E}\left[\delta_t^{k-1} - \hat{\delta}_t^{k-1}\right] = 0 \tag{228}$$

and that

$$\forall\, t, s \in [T] \text{ s.t. } s \neq t, \;\; \mathbb{E}\langle \delta_t^{k-1} - \hat{\delta}_t^{k-1}, \delta_s^{k-1} - \hat{\delta}_s^{k-1}\rangle = 0 \tag{229}$$

We introduce the matrix notation,

$$\hat{\Upsilon}^{k-1} \triangleq \left[\hat{\delta}_1^{k-1}, \dots, \hat{\delta}_T^{k-1}\right] \in \mathbb{R}^{d_u \times T}; \qquad \Upsilon^{k-1} \triangleq \left[\delta_1^{k-1}, \dots, \delta_T^{k-1}\right] \in \mathbb{R}^{d_u \times T} \tag{230}$$

Using this notation, Eq. (225) becomes

$$\mathbf{U}^k = \left[\mathbf{U}^{k-1} - \eta_{k-1} \hat{\Upsilon}^{k-1}\right] W^{k-1} \tag{231}$$

J.2.2. Proof of Theorem F.1′

In fully decentralized optimization, proving the convergence usually consists in deriving a recurrence on a term measuring the optimality of the average iterate (in our case this term is $\mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^k\right)\right\|^2$) and a term measuring the distance to consensus, i.e., $\mathbb{E}\sum_{t=1}^T \left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|^2$. In what follows we obtain those two recurrences, and then prove the convergence.

Lemma J.7 (Average iterate term recursion)**.** *Suppose that Assumptions 5′–7′ and Assumption 8 hold. Then, for $k > 0$, and $(\eta_{k,j})_{1\leq j\leq J-1}$ such that $\eta_k \triangleq \sum_{j=0}^{J-1}\eta_{k,j} \leq \min\left\{\frac{1}{2\sqrt{2}L}, \frac{1}{8L\beta}\right\}$, the updates of fully decentralized federated surrogate optimization (Alg. 9) verify*

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^k) - f(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] &\leq -\frac{1}{T}\sum_{t=1}^T d_{\mathcal{V}}\left(\mathbf{v}_t^k, \mathbf{v}_t^{k-1}\right)\\ &- \frac{\eta_{k-1}}{8}\mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^T \mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2\\ &+ \frac{\eta_{k-1}^2 L}{T}\left(4\sum_{j=0}^{J-1}\frac{L\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 + \frac{16\eta_{k-1}^3 L^2}{T}G^2. \end{aligned} \tag{232}$$

Proof. We multiply both sides of Eq. (231) by $\frac{\mathbf{1}\mathbf{1}^\intercal}{T}$, thus for $k > 0$ we have,

$$\mathbf{U}^k\cdot\frac{\mathbf{1}\mathbf{1}^\intercal}{T} = \left[\mathbf{U}^{k-1} - \eta_{k-1}\hat{\Upsilon}^{k-1}\right]W^{k-1}\frac{\mathbf{1}\mathbf{1}^\intercal}{T}, \tag{233}$$

since W^{k-1} is doubly stochastic (Assumption 8), i.e., $W^{k-1}\frac{\mathbf{1}\mathbf{1}^\intercal}{T} = \frac{\mathbf{1}\mathbf{1}^\intercal}{T}$, is follows that,

$$\bar{\mathbf{U}}^k = \bar{\mathbf{U}}^{k-1} - \eta_{k-1}\hat{\Upsilon}^{k-1}\cdot\frac{\mathbf{1}\mathbf{1}^\intercal}{T}, \tag{234}$$

thus,

$$\bar{\mathbf{u}}^k = \bar{\mathbf{u}}^{k-1} - \frac{\eta_{k-1}}{T}\cdot\sum_{t=1}^T\hat{\delta}_t^{k-1}. \tag{235}$$

Using the fact that g^k is L-smooth (Assumption 5′), we write

$$\mathbb{E}\left[g^k\left(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}\right)\right] = \mathbb{E}\left[g^k\left(\bar{\mathbf{u}}^{k-1} - \frac{\eta_{k-1}}{T}\sum_{t=1}^T\hat{\delta}_t^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right] \tag{236}$$

$$\begin{aligned}&\leq g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) - \mathbb{E}\left\langle\nabla_{\mathbf{u}} g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}), \frac{\eta_{k-1}}{T}\sum_{t=1}^T\hat{\delta}_t^{k-1}\right\rangle\\ &\quad + \frac{L}{2}\mathbb{E}\left\|\frac{\eta_{k-1}}{T}\sum_{t=1}^T\hat{\delta}_t^{k-1}\right\|^2 \end{aligned} \tag{237}$$

$$\begin{aligned}&= g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) - \eta_{k-1}\underbrace{\mathbb{E}\left\langle\nabla_{\mathbf{u}} g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}), \frac{1}{T}\sum_{t=1}^T\hat{\delta}_t^{k-1}\right\rangle}_{\triangleq T_1}\\ &\quad + \frac{\eta_{k-1}^2\cdot L}{2T^2}\underbrace{\mathbb{E}\left\|\sum_{t=1}^T\hat{\delta}_t^{k-1}\right\|^2}_{\triangleq T_2}, \end{aligned} \tag{238}$$

where the expectation is taken over local random batches. As in the centralized case, we bound the terms T_1 and T_2. First, we bound T_1, for $k > 0$, we have

$$T_1 = \mathbb{E}\left\langle\nabla_{\mathbf{u}} g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}), \frac{1}{T}\sum_{t=1}^T\hat{\delta}_t^{k-1}\right\rangle \tag{239}$$

$$= \underbrace{\mathbb{E}\left\langle \nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right), \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right)\right\rangle}_{=0, \text{ because } \mathbb{E}\left[\delta_t^{k-1} - \hat{\delta}_t^{k-1}\right]=0}$$

$$+ \mathbb{E}\left\langle \nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right), \frac{1}{T}\sum_{t=1}^{T}\delta_t^{k-1}\right\rangle \tag{240}$$

$$= \mathbb{E}\left\langle \nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right), \frac{1}{T}\sum_{t=1}^{T}\delta_t^{k-1}\right\rangle \tag{241}$$

$$= \frac{1}{2}\,\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{T}\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2$$

$$- \frac{1}{2}\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right) - \frac{1}{T}\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{242}$$

We bound now T_2. For $k > 0$, we have,

$$T_2 = \mathbb{E}\left\|\sum_{t=1}^{T}\hat{\delta}_t^{k-1}\right\|^2 \tag{243}$$

$$= \mathbb{E}\left\|\sum_{t=1}^{T}\left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right) + \sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{244}$$

$$\leq 2\mathbb{E}\left\|\sum_{t=1}^{T}\left(\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right)\right\|^2 + 2\cdot\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{245}$$

$$= 2\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right\|^2 + 2\sum_{1\leq t\neq s\leq T}\underbrace{\mathbb{E}\left\langle \hat{\delta}_t^{k-1} - \delta_t^{k-1}, \hat{\delta}_s^{k-1} - \delta_s^{k-1}\right\rangle}_{=0;\text{ because of Eq. (229)}}$$

$$+ 2\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{246}$$

$$= 2\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\hat{\delta}_t^{k-1} - \delta_t^{k-1}\right\|^2 + 2\cdot\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{247}$$

$$= 2\cdot\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 + 2\cdot\sum_{t=1}^{T}\left(\frac{\omega_t^2}{\eta_{k-1}^2}\,\mathbb{E}\left\|\sum_{j=0}^{J-1}\eta_{k-1,j}\cdot\left[\nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right)\right.\right.\right.$$

$$\left.\left.\left. - \nabla_{\mathbf{u}} g_t^k\left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}; \xi_t^{k-1,j}\right)\right]\right\|^2\right). \tag{248}$$

Since batches are sampled independently, stochastic gradients are unbiased and have finite variance (Assumption 6′), the last term in the RHS of the previous can be bounded using σ^2, leading to

$$T_2 \leq 2\cdot\sum_{t=1}^{T}\left[\omega_t^2\cdot\frac{\sum_{j=0}^{J-1}\eta_{k-1,j}^2}{\eta_{k-1}^2}\sigma^2\right] + 2\cdot\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{249}$$

$$= 2\cdot\sigma^2\cdot\left(\sum_{t=1}^{T}\omega_t^2\cdot\frac{\sum_{j=0}^{J-1}\eta_{k-1,j}^2}{\eta_{k-1}^2}\right) + 2\mathbb{E}\left\|\sum_{t=1}^{T}\delta_t^{k-1}\right\|^2 \tag{250}$$

$$\leq 2 \cdot \sigma^2 + 2 \cdot \mathbb{E} \left\| \sum_{t=1}^{T} \delta_t^{k-1} \right\|^2 . \tag{251}$$

Replacing Eq. (242) and Eq. (251) in Eq. (238), we have

$$\begin{aligned}
\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right] \leq \\
- \frac{\eta_{k-1}}{2} \mathbb{E} \left\| \nabla_{\mathbf{u}} g^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 - \frac{\eta_{k-1}}{2} \left(1 - 2L\eta_{k-1} \right) \mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^{T} \delta_t^{k-1} \right\|^2 \\
+ \frac{L}{T} \eta_{k-1}^2 \sigma^2 + \frac{\eta_{k-1}}{2} \mathbb{E} \left\| \nabla_{\mathbf{u}} g^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) - \frac{1}{T} \sum_{t=1}^{T} \delta_t^{k-1} \right\|^2 .
\end{aligned} \tag{252}$$

For η_{k-1} small enough, in particular for $\eta_{k-1} \leq \frac{1}{2L}$, we have

$$\begin{aligned}
\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right] \leq \\
- \frac{\eta_{k-1}}{2} \mathbb{E} \left\| \nabla_{\mathbf{u}} g^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 + \frac{L}{T} \eta_{k-1}^2 \sigma^2 \\
+ \frac{\eta_{k-1}}{2} \mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^{T} \left(\nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \delta_t^{k-1} \right) \right\|^2 .
\end{aligned} \tag{253}$$

We use Jensen inequality to bound the lest term in the RHS of the previous equation, leading to

$$\begin{aligned}
\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) \right] \leq \\
- \frac{\eta_{k-1}}{2} \mathbb{E} \left\| \nabla_{\mathbf{u}} g^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 + \frac{L}{T} \eta_{k-1}^2 \sigma^2 \\
+ \frac{\eta_{k-1}}{2T} \cdot \sum_{t=1}^{T} \underbrace{\mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \delta_t^{k-1} \right\|^2}_{T_3} .
\end{aligned} \tag{254}$$

We bound now the term T_3,

$$T_3 = \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \delta_t^{k-1} \right\|^2 \tag{255}$$

$$= \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \frac{\sum_{j=0}^{J-1} \eta_{k-1,j} \cdot \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right)}{\eta_{k-1}} \right\|^2 \tag{256}$$

$$= \mathbb{E} \left\| \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \left[\nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) \right] \right\|^2 . \tag{257}$$

Using Jensen inequality, it follows that

$$T_3 \leq \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1} \right) \right\|^2 \tag{258}$$

$$= \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1} \right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1} \right) \right.$$

$$+ \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right) \Big\|^2 \tag{259}$$

$$\leq 2 \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 2 \cdot \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1}, \mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}} g_t^k \left(\mathbf{u}_t^{k-1,j}, \mathbf{v}_t^{k-1}\right) \right\|^2 \tag{260}$$

$$\leq 2L^2 \cdot \mathbb{E} \left\| \bar{\mathbf{u}}^{k-1} - \mathbf{u}_t^{k-1} \right\|^2 + 2L^2 \cdot \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,j} - \mathbf{u}_t^{k-1,0} \right\|^2, \tag{261}$$

where we used the L-smoothness of g_t^k (Assumption 5′) to obtain the last inequality. As in the centralized case (lemma J.1), we bound terms $\left\| \mathbf{u}_t^{k-1,j} - \mathbf{u}_t^{k-1,0} \right\|^2$, $j \in \{0, \ldots, J-1\}$. Using exactly the same steps as in the proof of lemma J.1, Eq. (145) holds with $\mathbf{u}^{k-1}$ with $\mathbf{u}_t^{k-1,0}$, i.e.,

$$\left(1 - 4\eta_{k-1}^2 L^2\right) \cdot \sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,0} - \mathbf{u}_t^{k-1,j} \right\|^2 \leq 2\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} + 4\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\mathbf{u}_t^{k-1,0}, \mathbf{v}_t^{k-1}\right) \right\|^2. \tag{262}$$

For η_{k-1} small enough, in particular for $\eta_{k-1} \leq \frac{1}{2\sqrt{2}L}$, we have

$$\sum_{j=0}^{J-1} \frac{\eta_{k-1,j}}{\eta_{k-1}} \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1,0} - \mathbf{u}_t^{k-1,j} \right\|^2$$

$$\leq 8\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\mathbf{u}_t^{k-1,0}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 4\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} \tag{263}$$

$$\leq 8\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\mathbf{u}_t^{k-1,0}, \mathbf{v}_t^{k-1}\right) - \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) + \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 4\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} \tag{264}$$

$$\leq 16\eta_{k-1}^2 \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\mathbf{u}_t^{k-1,0}, \mathbf{v}_t^{k-1}\right) - \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 16\eta_{k-1}^2 \cdot \left\| \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 4\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} \tag{265}$$

$$\leq 16\eta_{k-1}^2 L^2 \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1} \right\|^2 + 16\eta_{k-1}^2 \cdot \left\| \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 4\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\}, \tag{266}$$

where the last inequality follows from the L-smoothness of g_t^k. Replacing Eq. (266) in Eq. (261), we have

$$T_3 \leq 32\eta_{k-1}^2 L^4 \cdot \mathbb{E} \left\| \mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1} \right\|^2 + 8L^2\sigma^2 \cdot \left\{ \sum_{j=0}^{J-1} \eta_{k-1,j}^2 \right\} + 32\eta_{k-1}^2 L^2 \cdot \mathbb{E} \left\| \nabla_u g_t^k \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right) \right\|^2 + 2L^2 \cdot \mathbb{E} \left\| \bar{\mathbf{u}}^{k-1} - \mathbf{u}_t^{k-1} \right\|^2, \tag{267}$$

where the last inequality follows from the L-smoothness of g_t^k. For η_k small enough, in particular if $\eta_k \leq \frac{1}{2\sqrt{2}L}$ we have,

$$T_3 \leq 6L^2\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 + 8L^2\sigma^2\sum_{j=0}^{J-1}\eta_{k-1,j}^2 + 32\eta_{k-1}^2L^2\left\|\nabla_u g_t^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right)\right\|^2. \tag{268}$$

Replacing Eq. (268) in Eq. (254), we have

$$\begin{aligned}\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] \leq \\ \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 + \frac{\eta_{k-1}^2L}{T}\left(4\sum_{j=0}^{J-1}\frac{LT\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 \\ - \frac{\eta_{k-1}}{2}\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{16\eta_{k-1}^3L^2}{T}\sum_{t=1}^{T}\left\|\nabla_u g_t^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_t^{k-1}\right)\right\|^2.\end{aligned} \tag{269}$$

We use now Assumption 7′ to bound the last term in the RHS of the previous equation, leading to

$$\begin{aligned}\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] \leq \\ \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 + \frac{\eta_{k-1}^2L}{T}\left(4\sum_{j=0}^{J-1}\frac{LT\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 \\ - \frac{\eta_{k-1}\cdot\left(1 - 32\eta_{k-1}^2L^2\beta^2\right)}{2}\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{16\eta_{k-1}^3L^2}{T}G^2.\end{aligned} \tag{270}$$

For η_{k-1} small enough, in particular, if $\eta_{k-1} \leq \frac{1}{8L\beta}$, we have

$$\begin{aligned}\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - g^k(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] \leq \\ - \frac{\eta_{k-1}}{4}\mathbb{E}\left\|\nabla_{\mathbf{u}} g^k\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 \\ + \frac{\eta_{k-1}^2L}{T}\left(4\sum_{j=0}^{J-1}\frac{LT\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 + \frac{16\eta_{k-1}^3L^2}{T}G^2.\end{aligned} \tag{271}$$

We use Lemma J.14 to get

$$\begin{aligned}\mathbb{E}\left[g^k(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^{k-1}) - f(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] \leq \\ - \frac{\eta_{k-1}}{8}\mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}\right)\right\|^2 + \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 \\ + \frac{\eta_{k-1}^2L}{T}\left(4\sum_{j=0}^{J-1}\frac{L\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 + \frac{16\eta_{k-1}^3L^2}{T}G^2.\end{aligned} \tag{272}$$

Finally, since g_t^k is a pseudo first-order surrogate of f_t near $\left\{\mathbf{u}^{k-1}, \mathbf{v}_t^{k-1}\right\}$, we have

$$\mathbb{E}\left[f(\bar{\mathbf{u}}^k, \mathbf{v}_{1:T}^k) - f(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1})\right] \leq -\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\, d_{\mathcal{V}}\left(\mathbf{v}_t^k, \mathbf{v}_t^{k-1}\right)$$

$$- \frac{\eta_{k-1}}{8} \mathbb{E} \left\| \nabla_{\mathbf{u}} f \left(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1} \right) \right\|^2 + \frac{3\eta_{k-1} L^2}{T} \cdot \sum_{t=1}^{T} \mathbb{E} \left\| \mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1} \right\|^2$$
$$+ \frac{\eta_{k-1}^2 L}{T} \left(4 \sum_{j=0}^{J-1} \frac{L \cdot \eta_{k-1,j}^2}{\eta_{k-1}} + 1 \right) \sigma^2 + \frac{16 \eta_{k-1}^3 L^2}{T} G^2. \tag{273}$$

□

Lemma J.8 (Recursion for consensus distance, part 1)**.** *Suppose that Assumptions 5′–7′ and Assumption 8 hold. Consider* $m = \left\lfloor \frac{k}{\tau} \right\rfloor - 1$*, then, for* $k > 0$*, and* $(\eta_{k,j})_{1 \le j \le J-1}$ *such that* $\eta_k \triangleq \sum_{j=0}^{J-1} \eta_{k,j} \le \min \left\{ \frac{1}{2\sqrt{2}L}, \frac{1}{4L\beta} \right\}$*, the updates of fully decentralized federated surrogate optimization (Alg 9) verify*

$$\mathbb{E} \sum_{t=1}^{T} \left\| \mathbf{u}_t^k - \bar{\mathbf{u}}^k \right\|_F^2 \le$$
$$(1 - \frac{p}{2}) \beta^2 \mathbb{E} \left\| \mathbf{U}^{m\tau} - \bar{\mathbf{U}}^{m\tau} \right\|_F^2 + 20\tau \left(1 + \frac{2}{p} \right) L^2 \sum_{l=m\tau}^{k-1} \eta_l^2 \, \mathbb{E} \left\| \mathbf{U}^l - \bar{\mathbf{U}}^l \right\|_F^2$$
$$+ \left(1 + 16 L^2 \tau \left(1 + \frac{2}{p} \right) \cdot \left\{ \sum_{j=0}^{J-1} \eta_{l,j}^2 \right\} \right) \cdot T \cdot \sigma^2 + 16\tau \left(1 + \frac{2}{p} \right) T^2 G^2 \sum_{l=m\tau}^{k-1} \eta_l^2$$
$$+ 16\tau \left(1 + \frac{2}{p} \right) T^2 \beta^2 \sum_{l=m\tau}^{k-1} \eta_l^2 \, \mathbb{E} \left\| \nabla_{\mathbf{u}} f \left(\bar{\mathbf{u}}^{l,j}, \mathbf{v}_{1:T}^{l+1} \right) \right\|^2. \tag{274}$$

Proof. For $k \ge \tau$, and $m = \left\lfloor \frac{k}{\tau} \right\rfloor - 1$, we have

$$\mathbb{E} \sum_{t=1}^{T} \left\| \mathbf{u}_t^k - \bar{\mathbf{u}}^k \right\|_F^2 = \mathbb{E} \left\| \mathbf{U}^k - \bar{\mathbf{U}}^k \right\|_F^2 \tag{275}$$
$$= \mathbb{E} \left\| \mathbf{U}^k - \bar{\mathbf{U}}^{m\tau} + \bar{\mathbf{U}}^{m\tau} - \bar{\mathbf{U}}^k \right\|_F^2 \tag{276}$$
$$\le \mathbb{E} \left\| \mathbf{U}^k - \bar{\mathbf{U}}^{m\tau} \right\|_F^2. \tag{277}$$

Using Eq. (231) recursively, we have

$$\mathbf{U}^k = \mathbf{U}^{m\tau} \left\{ \prod_{l'=m\tau}^{k-1} W^{l'} \right\} - \sum_{l=m\tau}^{k-1} \eta_l \hat{\Upsilon}^l \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\}. \tag{278}$$

Thus,

$$\mathbb{E} \sum_{t=1}^{T} \left\| \mathbf{u}_t^k - \bar{\mathbf{u}}^k \right\|_F^2 \le \mathbb{E} \left\| \mathbf{U}^{m\tau} \left\{ \prod_{l'=m\tau}^{k-1} W^{l'} \right\} - \bar{\mathbf{U}}^{m\tau} - \sum_{l=m\tau}^{k-1} \eta_l \hat{\Upsilon}^l \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \right\|_F^2 \tag{279}$$
$$= \mathbb{E} \left\| \mathbf{U}^{m\tau} \left\{ \prod_{l'=m\tau}^{k-1} W^{l'} \right\} - \bar{\mathbf{U}}^{m\tau} - \sum_{l=m\tau}^{k-1} \eta_l \Upsilon^l \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \sum_{l=m\tau}^{k-1} \eta_l \left(\Upsilon^l - \hat{\Upsilon}^l \right) \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \right\|_F^2 \tag{280}$$
$$= \mathbb{E} \left\| \mathbf{U}^{m\tau} \left\{ \prod_{l'=m\tau}^{k-1} W^{l'} \right\} - \bar{\mathbf{U}}^{m\tau} - \sum_{l=m\tau}^{k-1} \eta_l \Upsilon^l \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \right\|_F^2 + \mathbb{E} \left\| \sum_{l=m\tau}^{k-1} \eta_l \left(\Upsilon^l - \hat{\Upsilon}^l \right) \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \right\|_F^2$$
$$+ 2\mathbb{E} \left\langle \mathbf{U}^{m\tau} \left\{ \prod_{l'=m\tau}^{k-1} W^{l'} \right\} - \bar{\mathbf{U}}^{m\tau} - \sum_{l=m\tau}^{k-1} \eta_l \Upsilon^l \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\}, \sum_{l=m\tau}^{k-1} \eta_l \left(\Upsilon^l - \hat{\Upsilon}^l \right) \left\{ \prod_{l'=l}^{k-1} W^{l'} \right\} \right\rangle_F. \tag{281}$$

Since stochastic gradients are unbiased, the last term in the RHS of the previous equation is equal to zero. Using the following standard inequality for euclidean norm with $\alpha > 0$,

$$\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\alpha)\|\mathbf{a}\|^2 + \left(1+\alpha^{-1}\right)\|\mathbf{b}\|^2, \tag{282}$$

we have

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq \tag{283}$$

$$(1+\alpha)\,\mathbb{E}\left\|\mathbf{U}^{m\tau}\left\{\prod_{l'=m\tau}^{k-1} W^{l'}\right\} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + \left(1+\alpha^{-1}\right)\mathbb{E}\left\|\sum_{l=m\tau}^{k-1}\eta_l\Upsilon^l\left\{\prod_{l'=l}^{k-1} w^{l'}\right\}\right\|_F^2$$
$$+\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\left(\Upsilon^l - \hat{\Upsilon}^l\right)\left\{\prod_{l'=l}^{k-1} W^{l'}\right\}\right\|_F^2. \tag{284}$$

Since $k \geq (m+1)\tau$ and matrices $\left(W^l\right)_{l\geq 0}$ are doubly stochastic, we have

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq (1+\alpha)\,\mathbb{E}\left\|\mathbf{U}^{m\tau}\left\{\prod_{l'=m\tau}^{(m+1)\tau-1} W^{l'}\right\} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + \left(1+\alpha^{-1}\right)\mathbb{E}\left\|\sum_{l=m\tau}^{k-1}\eta_l\Upsilon^l\right\|_F^2$$
$$+\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l - \hat{\Upsilon}^l\right\|_F^2 \tag{285}$$

$$\leq (1+\alpha)\,\mathbb{E}\left\|\mathbf{U}^{m\tau}\left\{\prod_{l'=m\tau}^{(m+1)\tau-1} W^{l'}\right\} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + \left(1+\alpha^{-1}\right)\cdot(k-m\tau)\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l\right\|_F^2$$
$$+\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l - \hat{\Upsilon}^l\right\|_F^2 \tag{286}$$

Using Assumption 8 to bound the first of the RHS of the previous equation and the fact that that $k \leq (m+2)\tau$, it follows that

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq (1+\alpha)(1-p)\mathbb{E}\left\|\mathbf{U}^{m\tau} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + 2\tau\left(1+\alpha^{-1}\right)\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l\right\|_F^2 + \sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l - \hat{\Upsilon}^l\right\|_F^2. \tag{287}$$

We use the fact that stochastic gradients have bounded variance (Assumption 6′) to bound $\mathbb{E}\left\|\Upsilon^l - \hat{\Upsilon}^l\right\|_F^2$ as follow,

$$\mathbb{E}\left\|\Upsilon^l - \hat{\Upsilon}^l\right\|_F^2 = \sum_{t=1}^{T}\mathbb{E}\left\|\delta_t^l - \hat{\delta}_t^l\right\|^2 \tag{288}$$

$$= \sum_{t=1}^{T}\mathbb{E}\left\|\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\left(\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^{l};\xi_t^{l,j}\right)\right)\right\| \tag{289}$$

$$\leq \sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\left(\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^{k-1}\right) - \nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^{l};\xi_t^{l,j}\right)\right)\right\| \tag{290}$$

$$\leq \sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\sigma^2 \tag{291}$$

$$= T\cdot\sigma^2, \tag{292}$$

where we used Jensen inequality to obtain the first inequality and Assumption 6′ to obtain the second inequality. Replacing back in Eq. (287), we have

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k-\bar{\mathbf{u}}^k\right\|_F^2\leq$$

$$(1+\alpha)(1-p)\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+2\tau\left(1+\alpha^{-1}\right)\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l\right\|_F^2+T\cdot\sigma^2. \tag{293}$$

The last step of the proof consists in bounding $\mathbb{E}\left\|\Upsilon^l\right\|_F$ for $l\in\{m\tau,\dots k-1\}$,

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2=\sum_{t=1}^{T}\mathbb{E}\left\|\delta_t^l\right\|^2 \tag{294}$$

$$=\sum_{t=1}^{T}\mathbb{E}\left\|\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)\right\|^2 \tag{295}$$

$$\leq\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)\right\|^2 \tag{296}$$

$$\leq\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)+\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)\right\|^2 \tag{297}$$

$$\leq 2\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)\right\|^2$$
$$+2\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)\right\|^2. \tag{298}$$

Since g_t^{l+1} is a first order surrogate of f near $\left\{u_t^l,v_t^l\right\}$, we have

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2\leq 2\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,0},\mathbf{v}_t^l\right)\right\|^2$$
$$+2\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^l,\mathbf{v}_t^l\right)+\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^l,\mathbf{v}_t^l\right)\right\|^2 \tag{299}$$

$$\leq 2\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,j},\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}g_t^{l+1}\left(\mathbf{u}_t^{l,0},\mathbf{v}_t^l\right)\right\|^2$$
$$+4\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}f_t\left(\mathbf{u}_t^l,\mathbf{v}_t^l\right)-\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^l,\mathbf{v}_t^l\right)\right\|^2+4\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^l,\mathbf{v}_t^l\right)\right\|^2. \tag{300}$$

Since f is $2L$-smooth w.r.t $\mathbf{u}$ (Lemma J.12) and g is L-smooth w.r.t $\mathbf{u}$ (Assumption 5′), we have

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2\leq 2\sum_{t=1}^{T}\sum_{j=0}^{J-1}\frac{\eta_{l,j}}{\eta_l}\cdot L^2\,\mathbb{E}\left\|\mathbf{u}_t^{l,j}-\mathbf{u}_t^{l,0}\right\|^2++16L^2\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^l-\bar{\mathbf{u}}^l\right\|^2$$
$$+4\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^l,\mathbf{v}_t^l\right)\right\|^2. \tag{301}$$

We use Eq. (266) to bound the first term in the RHS of the previous equation, leading to

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2\leq 32\eta_l^2L^2\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}}g_t^{l+1}\left(\bar{\mathbf{u}}^{l,j},v_t^{l+1}\right)\right\|^2+16L^2\left(1+2\eta_l^2L^2T\right)\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^l-\bar{\mathbf{u}}^l\right\|^2$$

$$+ 4\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}} f_t\left(\bar{\mathbf{u}}^l, \mathbf{v}_t^l\right)\right\|^2 + 8L^2\sigma^2 \cdot \left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\}. \tag{302}$$

Using Lemma J.14, we have

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2 \leq 4\left(1+16\eta_l^2L^2\right)\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}} f_t\left(\bar{\mathbf{u}}^{l,j}, \mathbf{v}_t^{l+1}\right)\right\|^2$$
$$+ 16L^2\left(3+2\eta_l^2L^2\right)\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^l - \bar{\mathbf{u}}^l\right\|^2 + 8L^2\sigma^2T\cdot\left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\}. \tag{303}$$

For η_l small enough, in particular, for $\eta_l \leq \frac{1}{2\sqrt{2}L}$, we have

$$\mathbb{E}\left\|\Upsilon^l\right\|_F^2 \leq 8\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}} f_t\left(\bar{\mathbf{u}}^{l,j}, v_t^{l+1}\right)\right\|^2 + 10L^2\,\mathbb{E}\left\|\mathbf{U}^l - \bar{\mathbf{U}}^l\right\|_F^2 + 8L^2\sigma^2T\left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\} \tag{304}$$

Replacing Eq. (304) in Eq. (293), we have

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq$$
$$(1+\alpha)(1-p)\mathbb{E}\left\|\mathbf{U}^{m\tau} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + 20\tau\left(1+\alpha^{-1}\right)L^2\sum_{l=m\tau}^{k-1}\eta_l^2\,\mathbb{E}\left\|\mathbf{U}^l - \bar{\mathbf{U}}^l\right\|_F^2$$
$$+ 16\tau\left(1+\alpha^{-1}\right)\sum_{l=m\tau}^{k-1}\eta_l^2\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{\mathbf{u}} f_t\left(\bar{\mathbf{u}}^{l,j}, v_t^{l+1}\right)\right\|^2$$
$$+ \left(1+16L^2\tau\left(1+\alpha^{-1}\right)\cdot\left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\}\right)\cdot T\cdot\sigma^2. \tag{305}$$

Finally, using Lemma J.13 and considering $\alpha = \frac{p}{2}$, we have

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq$$
$$(1-\frac{p}{2})\beta^2\mathbb{E}\left\|\mathbf{U}^{m\tau} - \bar{\mathbf{U}}^{m\tau}\right\|_F^2 + 20\tau\left(1+\frac{2}{p}\right)L^2\sum_{l=m\tau}^{k-1}\eta_l^2\,\mathbb{E}\left\|\mathbf{U}^l - \bar{\mathbf{U}}^l\right\|_F^2$$
$$+ \left(1+16L^2\tau\left(1+\frac{2}{p}\right)\cdot\left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\}\right)\cdot T\cdot\sigma^2 + 16\tau\left(1+\frac{2}{p}\right)T^2G^2\sum_{l=m\tau}^{k-1}\eta_l^2$$
$$+ 16\tau\left(1+\frac{2}{p}\right)T^2\beta^2\sum_{l=m\tau}^{k-1}\eta_l^2\,\mathbb{E}\left\|\nabla_{\mathbf{u}} f\left(\bar{\mathbf{u}}^{l,j}, \mathbf{v}_{1:T}^{l+1}\right)\right\|^2. \tag{306}$$

□

Lemma J.9 (Recursion for consensus distance, part 2)**.** *Suppose that Assumptions 5′–7′ and Assumption 8 hold. Consider $m\tau \leq k < (m+1)\tau$, then, for $(\eta_{k,j})_{1\leq j\leq J-1}$ such that $\eta_k \triangleq \sum_{j=0}^{J-1}\eta_{k,j} \leq \min\left\{\frac{1}{2\sqrt{2}L}, \frac{1}{4L\beta}\right\}$, the updates of fully decentralized federated surrogate optimization (Alg 9) verify*

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k - \bar{\mathbf{u}}^k\right\|_F^2 \leq$$

$$(1+\alpha)(1+\frac{p}{2})\beta^2\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+20\tau\left(1+\alpha^{-1}\right)L^2\sum_{l=m\tau}^{k-1}\eta_l^2\,\mathbb{E}\left\|\mathbf{U}^{l}-\bar{\mathbf{U}}^{l}\right\|_F^2$$
$$+\left(1+16L^2\tau\left(1+\alpha^{-1}\right)\cdot\left\{\sum_{j=0}^{J-1}\eta_{l,j}^2\right\}\right)\cdot T\cdot\sigma^2+16\tau\left(1+\alpha^{-1}\right)T^2G^2\sum_{l=m\tau}^{k-1}\eta_l^2$$
$$+16\tau\left(1+\alpha^{-1}\right)T^2\beta^2\sum_{l=m\tau}^{k-1}\eta_l^2\,\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{l,j},v_{1:T}^{l+1}\right)\right\|^2. \tag{307}$$

Proof. We use exactly the same proof as in Lemma J.8, with the only difference that Eq. (287) is replaced by

$$\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{u}_t^k-\bar{\mathbf{u}}^k\right\|_F^2\leq$$
$$(1+\alpha)\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+2\tau\left(1+\alpha^{-1}\right)\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l\right\|_F^2$$
$$+\sum_{l=m\tau}^{k-1}\eta_l^2\mathbb{E}\left\|\Upsilon^l-\hat{\Upsilon}^l\right\|_F^2, \tag{308}$$

resulting from the fact that $\left\{\prod_{l'=m\tau}^{(m+1)\tau-1}W^{l'}\right\}$ is a doubly stochastic matrix. □

Lemma J.10. *Under Assum. 5′-7′ and Assum 8. For $\eta_{k,j}=\frac{\eta}{J}$ with*

$$\eta\leq min\left\{\frac{1}{2\sqrt{2}L},\frac{p}{256\tau L},\frac{1}{16}\sqrt{\frac{p^2}{24\tau^2\beta^2}}\right\}, \tag{309}$$

the iterates of Alg. 9 verifies

$$\sum_{k=0}^{K}\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq\frac{1}{2}\sum_{k=0}^{K}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^k,\mathbf{v}_{1:T}^k\right)\right\|^2+64A\frac{\tau}{p}K\eta^2. \tag{310}$$

Proof. Using Lemma J.8 and Lemma J.9, we have and using the fact that $p\leq 1$, we have for $m=\left\lfloor\frac{k}{\tau}\right\rfloor-1$

$$\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq(1-\frac{p}{2})\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+\frac{168\tau}{p}L^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\mathbf{U}^{l}-\bar{\mathbf{U}}^{l}\right\|_F^2 \tag{311}$$
$$+\eta^2\underbrace{(k-m\tau)\left\{\frac{\sigma^2}{E}\left(1+8\tau L^2\left(1+\frac{2}{p}\right)\right)+24\tau\left(1+\frac{2}{p}\right)G^2\right\}}_{\triangleq A} \tag{312}$$
$$+\frac{24\tau}{p}\beta^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{l},\mathbf{v}_{1:T}^{l}\right)\right\|^2 \tag{313}$$
$$\tag{314}$$

and for $m\tau\leq k<(m+1)\tau$,

$$\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq(1+\frac{p}{2})\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+\frac{168\tau}{p}L^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\mathbf{U}^{l}-\bar{\mathbf{U}}^{l}\right\|_F^2 \tag{315}$$
$$+\eta^2\underbrace{(k-m\tau)\left\{\frac{\sigma^2}{E}\left(1+8\tau L^2\left(1+\frac{2}{p}\right)\right)+24\tau\left(1+\frac{2}{p}\right)G^2\right\}}_{\triangleq A} \tag{316}$$

$$+\frac{24\tau}{p}\beta^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^l,\mathbf{v}_{1:T}^l\right)\right\|^2. \tag{317}$$

Using the fact that $\eta\leq\frac{p}{256\tau L}$, it follows that for $m=\left\lfloor\frac{k}{\tau}\right\rfloor-1$

$$\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq(1-\frac{p}{2})\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+\frac{p}{16\tau}L^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\mathbf{U}^l-\bar{\mathbf{U}}^l\right\|^2 \tag{318}$$

$$+\eta^2A+\frac{24\tau}{p}\beta^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^l,\mathbf{v}_{1:T}^l\right)\right\|^2, \tag{319}$$

and for $m\tau\leq k<(m+1)\tau$,

$$\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq(1+\frac{p}{2})\mathbb{E}\left\|\mathbf{U}^{m\tau}-\bar{\mathbf{U}}^{m\tau}\right\|_F^2+\frac{p}{64\tau}\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\mathbf{U}^l-\bar{\mathbf{U}}^l\right\|_F^2 \tag{320}$$

$$+\eta^2A+\frac{24\tau}{p}\beta^2\eta^2\sum_{l=m\tau}^{k-1}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^l,\mathbf{v}_{1:T}^l\right)\right\|^2. \tag{321}$$

The rest of the proof follows from (Koloskova et al., 2020, Lemma 14). □

Theorem F.1′. *Under Assumptions 4′–7′ and Assumption 8, when clients use SGD as local solver with learning rate $\eta=\frac{a_0}{\sqrt{K}}$, after a large enough number of communication rounds K, the iterates of fully decentralized federated surrogate optimization (Alg. 9) satisfy:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^k,\mathbf{v}_{1:T}^k\right)\right\|^2\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \tag{322}$$

and,

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\omega_t\cdot\mathbb{E}\,d_{\mathcal{V}}\left(\mathbf{v}_t^k,\mathbf{v}_t^{k+1}\right)\leq\mathcal{O}\left(\frac{1}{K}\right), \tag{323}$$

where $\bar{\mathbf{u}}^k=\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t^k$. Moreover, local estimates $\left(\mathbf{u}_t^k\right)_{1\leq t\leq T}$ converge to consensus, i.e., to $\bar{\mathbf{u}}^k$:

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^k-\bar{\mathbf{u}}^k\right\|^2\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right). \tag{324}$$

Proof. We prove first the convergence to a stationary point in $\mathbf{u}$, i.e. Eq. (322), using (Koloskova et al., 2020, Lemma 17), then we prove Eq. (323) and Eq. (324).

Proof of Eq. 322. The result follows immediately from Lemma J.7 and Lemma J.10 by using (Koloskova et al., 2020, Lemma 17).

Proof of Eq. 324. We multiply Eq. (310) (Lemma J.10) by $\frac{1}{K+1}$, and we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq\frac{1}{2(K+1)}\sum_{k=0}^{K}\mathbb{E}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^k,\mathbf{v}_{1:T}^k\right)\right\|_F^2+\frac{64A\tau}{p(K+1)}K\eta^2, \tag{325}$$

since $\eta\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, using Eq. (322), it follows that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\mathbf{U}^k-\bar{\mathbf{U}}^k\right\|_F^2\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \tag{326}$$

Thus,

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^k-\bar{\mathbf{u}}^k\right\|_F^2\leq\mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \tag{327}$$

Proof of Eq. 323. Using the result of Lemma J.7 we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\, d_{\mathcal{V}}\left(v_t^k, v_t^{k-1}\right) \leq & \mathbb{E}\left[f(\bar{\mathbf{u}}^{k-1}, \mathbf{v}_{1:T}^{k-1}) - f(\bar{\mathbf{u}}^{k}, \mathbf{v}_{1:T}^{k-1})\right] \\ & + \frac{3\eta_{k-1}L^2}{T}\cdot\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{u}_t^{k-1} - \bar{\mathbf{u}}^{k-1}\right\|^2 \\ & + \frac{\eta_{k-1}^2 L}{T}\left(4\sum_{j=0}^{J-1}\frac{L\cdot\eta_{k-1,j}^2}{\eta_{k-1}} + 1\right)\sigma^2 + \frac{16\eta_{k-1}^3 L^2}{T}G^2. \end{aligned} \tag{328}$$

The final result follows from the fact that $\eta = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ and Eq. (324). □

J.2.3. Proof of Theorem F.1

We state the formal version of Theorem F.1, for which only an informal version was given in the main text.

Theorem F.1. *Under Assumptions 1–8, when clients use SGD as local solver with learning rate $\eta = \frac{a_0}{\sqrt{K}}$, `D-FedEM`'s iterates satisfy the following inequalities after a large enough number of communication rounds K:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_\Theta f\left(\bar{\Theta}^k, \Pi^k\right)\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \quad \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{n_t}{n}\mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) \leq \mathcal{O}\left(\frac{1}{K}\right), \tag{329}$$

where $\bar{\Theta}^k = \Theta^k\frac{\mathbf{1}\mathbf{1}^\intercal}{T}$. Moreover, individual estimates $\left(\Theta_t^k\right)_{1\leq t\leq T}$ converge to consensus, i.e., to $\bar{\Theta}^k$:

$$\min_{k\in[K]}\mathbb{E}\sum_{t=1}^{T}\left\|\Theta_t^k - \bar{\Theta}^k\right\|_F^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Proof. We use exactly the same proof as in Appendix J.1.3, showing that `D-FedEM` can be obtained from fully decentralized federated surrogate optimization. □

J.3. Auxiliary Lemmas

Lemma J.11. *Consider $J \geq 2$ and positive real numbers, $\eta_j,\ j = 0, \dots, J-1$, then:*

$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{j-1} \eta_l \right\} \leq \sum_{j=0}^{J-2} \eta_j$$
$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{j-1} \eta_l^2 \right\} \leq \sum_{j=0}^{J-2} {\eta_j}^2$$
$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \left(\sum_{l=0}^{j-1} \eta_l \right)^2 \right\} \leq \sum_{j=0}^{J-1} \eta_j \cdot \sum_{j=0}^{J-2} \eta_j$$

Proof. For the first inequality,

$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{j-1} \eta_l \right\} \leq \frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{J-2} \eta_l \right\} = \cdot \sum_{l=0}^{J-2} \eta_l. \tag{330}$$

For the second inequality

$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{j-1} \eta_l^2 \right\} \leq \frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \sum_{l=0}^{J-2} \eta_l^2 \right\} = \sum_{l=0}^{J-2} \eta_l^2. \tag{331}$$

For the third inequality,

$$\frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \left(\sum_{l=0}^{j-1} \eta_l \right)^2 \right\} \leq \frac{1}{\sum_{j=0}^{J-1} \eta_j} \cdot \sum_{j=0}^{J-1} \left\{ \eta_j \cdot \left(\sum_{l=0}^{J-2} \eta_l \right)^2 \right\} \tag{332}$$
$$\leq \left(\sum_{j=0}^{J-2} \eta_j \right)^2 \tag{333}$$
$$\leq \sum_{j=0}^{J-1} \eta_j \cdot \sum_{j=0}^{J-2} \eta_j \tag{334}$$

□

Lemma J.12. *Suppose that g is a pseudo first-order surrogate of f, and that g is L-smooth, then f is $2L$-smooth.*

Proof. The difference between f and g is L-smooth, and g is L-smooth, thus f is L-smooth as the sum of two L-smooth functions. □

Lemma J.13. *Consider $f = \sum_{t=1}^{T} \omega_t \cdot f_t$, for weights $\omega \in \Delta^T$. Suppose that for all $\{\mathbf{u}^0, \mathbf{v}^0\} \in \mathbb{R}^{d_u} \times \mathcal{V}$, and $t \in [T]$, f_t admits a pseudo first-order surrogate $g_t^{\{\mathbf{u}^0, \mathbf{v}^0\}}$ near $\{\mathbf{u}^0, \mathbf{v}^0\}$, and that $g^{\{\mathbf{u}^0, \mathbf{v}^0\}} = \sum_{t=1}^{T} \omega_t \cdot g_t^{\{\mathbf{u}^0, \mathbf{v}^0\}}$ verifies Assumption 7′ for $t \in [T]$. Then f also verifies Assumption 7′.*

Proof. Consider arbitrary $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d_u} \times \mathcal{V}$, and for $t \in [T]$, consider $g^{\{\mathbf{u},\mathbf{v}\}}$ to be a pseudo first-order surrogate of f_t near $\{\mathbf{u}, \mathbf{v}\}$. We write Assumption 7′ for $g^{\{\mathbf{u},\mathbf{v}\}}$,

$$\sum_{t=1}^{T} \omega_t \cdot \left\| \nabla_{\mathbf{u}} g_t^{\{\mathbf{u},\mathbf{v}\}}(\mathbf{u}, \mathbf{v}) \right\|^2 \leq G^2 + \beta^2 \left\| \sum_{t=1}^{T} \omega_t \cdot \nabla_{\mathbf{u}} g_t^{\{\mathbf{u},\mathbf{v}\}}(\mathbf{u}, \mathbf{v}) \right\|^2. \tag{335}$$

Since $g_t^{\{\mathbf{u},\mathbf{v}\}}$ is a pseudo first-order surrogate of f_t near $\{\mathbf{u},\mathbf{v}\}$, it follows that

$$\sum_{t=1}^{T}\omega_t\cdot\left\|\nabla_{\mathbf{u}}f_t(\mathbf{u},\mathbf{v})\right\|^2\leq G^2+\beta^2\left\|\sum_{t=1}^{T}\omega_t\cdot\nabla_{\mathbf{u}}f_t(\mathbf{u},\mathbf{v})\right\|^2. \tag{336}$$

□

Lemma J.14. *For $k>0$, the iterates of Alg. 9, verifies*

$$g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\leq f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)+\frac{L}{2}\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2,$$

and,

$$\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2\leq 2\left\|\nabla_{\mathbf{u}}g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2-2L^2\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2,$$

and,

$$\left\|\nabla_{\mathbf{u}}g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2\leq 2\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2+2L^2\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2,$$

Proof. For $k>0$ and $t\in[T]$, we have

$$\begin{aligned}
&g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)=\\
&\quad g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right) &(337)\\
&\quad=f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right) &(338)\\
&\quad=f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)+r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right) &(339)
\end{aligned}$$

Since $g_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)=f_t\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)$, it follows that

$$g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)=f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right) \tag{340}$$

Because r_t^k is L-smooth in $\mathbf{u}$ (Lemma J.5), we have

$$\begin{aligned}
r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)\leq&\left\langle\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right),\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\rangle\\
&+\frac{L}{2}\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2
\end{aligned} \tag{341}$$

Because g_t^k is a first order surrogate of f_t near $\left\{u_t^{k-1},v_t^{k-1}\right\}$, we have $\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)=0$, thus

$$g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)\leq f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+\frac{L}{2}\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2 \tag{342}$$

Multiplying by ω_t and summing for $t\in[T]$, we have

$$g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\leq f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)+\frac{L}{2}\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2. \tag{343}$$

Writing the gradient of Eq. (340), we have

$$\nabla_{\mathbf{u}}g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)=\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)+\nabla_{\mathbf{u}}r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right) \tag{344}$$

Multiplying by ω_t and summing for $t\in[T]$, we have

$$\nabla_{\mathbf{u}}g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)=\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)+$$

$$+\sum_{t=1}^{T}\omega_t\left[\nabla_{\mathbf{u}}r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)\right] \tag{345}$$

thus,

$$\left\|\nabla_{\mathbf{u}}g^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2 =$$

$$\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)+\sum_{t=1}^{T}\omega_t\left[\nabla_{\mathbf{u}}r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)\right]\right\|^2 \tag{346}$$

$$\geq\frac{1}{2}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2-\left\|\sum_{t=1}^{T}\omega_t\left[\nabla_{\mathbf{u}}r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)\right]\right\|^2 \tag{347}$$

$$\geq\frac{1}{2}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2-\sum_{t=1}^{T}\omega_t\left\|\nabla_{\mathbf{u}}r_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)-\nabla_{\mathbf{u}}r_t^k\left(\mathbf{u}_t^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2 \tag{348}$$

$$\geq\frac{1}{2}\left\|\nabla_{\mathbf{u}}f\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_{1:T}^{k-1}\right)\right\|^2-L^2\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2 \tag{349}$$

Thus,

$$\left\|\nabla_{\mathbf{u}}f_t\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2\leq 2\left\|\nabla_{\mathbf{u}}g_t^k\left(\bar{\mathbf{u}}^{k-1},\mathbf{v}_t^{k-1}\right)\right\|^2-2L^2\sum_{t=1}^{T}\omega_t\left\|\bar{\mathbf{u}}^{k-1}-\mathbf{u}_t^{k-1}\right\|^2 \tag{350}$$

The last equation, follows exactly like the second one. □

Lemma J.15. *Consider* $\mathbf{u}_1,\ldots,\mathbf{u}_M\in\mathbb{R}^d$ *and* $\alpha=(\alpha_1,\ldots,\alpha_M)\in\Delta^M$. *Define the block matrix* $\mathbf{H}$ *with*

$$\begin{cases}\mathbf{H}_{m,m}=-\alpha_m\cdot(1-\alpha_m)\cdot\mathbf{u}_m\cdot\mathbf{u}_m^\intercal\\ \mathbf{H}_{m,m'}=\alpha_m\cdot\alpha_{m'}\cdot\mathbf{u}_m\cdot\mathbf{u}_{m'}^\intercal; & m'\neq m,\end{cases} \tag{351}$$

then $\mathbf{H}$ *is semi-definite negative matrix.*

Proof. Consider $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_M]\in\mathbb{R}^{dM}$, we want to prove that

$$\mathbf{X}^\intercal\cdot\mathbf{H}\cdot\mathbf{X}\leq 0. \tag{352}$$

We have,

$$\mathbf{X}^\intercal\cdot\mathbf{H}\cdot\mathbf{X}=\sum_{m=1}^{M}\sum_{m'=1}^{M}\mathbf{x}_m^\intercal\cdot\mathbf{H}_{m,m'}\cdot\mathbf{x}_{m'} \tag{353}$$

$$=\sum_{m=1}^{M}\left[\mathbf{x}_m^\intercal\cdot\mathbf{H}_{m,m}\cdot\mathbf{x}_m+\sum_{\substack{m'=1\\m'\neq m}}^{M}\mathbf{x}_m^\intercal\cdot\mathbf{H}_{m,m}\cdot\mathbf{x}_{m'}\right] \tag{354}$$

$$=\sum_{m=1}^{M}\left[-\alpha_m\cdot(1-\alpha_m)\cdot\mathbf{x}_m^\intercal\cdot\mathbf{u}_m\cdot\mathbf{u}_m^\intercal\cdot\mathbf{x}_m+\sum_{\substack{m'=1\\m'\neq m}}^{M}\left(\alpha_m\cdot\alpha_{m'}\cdot\mathbf{x}_m^\intercal\cdot\mathbf{u}_m\cdot\mathbf{u}_{m'}^\intercal\cdot\mathbf{x}_{m'}\right)\right] \tag{355}$$

$$=\sum_{m=1}^{M}\left[-\alpha_m\cdot(1-\alpha_m)\cdot\langle\mathbf{x}_m,\mathbf{u}_m\rangle^2+\alpha_m\cdot\langle\mathbf{x}_m,\mathbf{u}_m\rangle\sum_{\substack{m'=1\\m'\neq m}}^{M}\alpha_{m'}\cdot\langle\mathbf{x}_{m'},\mathbf{u}_{m'}\rangle\right]. \tag{356}$$

Since $\alpha \in \Delta^M$, we have $\sum_{m=1}^M \alpha_m = 1$, thus,

$$\mathbf{X}^\intercal \cdot \mathbf{H} \cdot \mathbf{X} = \sum_{m=1}^M \alpha_m \cdot \langle \mathbf{x}_m, \mathbf{u}_m \rangle \cdot \sum_{\substack{m'=1 \\ m' \neq m}}^M \alpha_{m'} \Big(\langle \mathbf{x}_{m'}, \mathbf{u}_{m'} \rangle - \langle \mathbf{x}_m, \mathbf{u}_m \rangle \Big) \tag{357}$$

$$= \sum_{m=1}^M \alpha_m \cdot \langle \mathbf{x}_m, \mathbf{u}_m \rangle \cdot \sum_{m'=1}^M \alpha_{m'} \Big(\langle \mathbf{x}_{m'}, \mathbf{u}_{m'} \rangle - \langle \mathbf{x}_m, \mathbf{u}_m \rangle \Big) \tag{358}$$

$$= \left(\sum_{m=1}^M \alpha_m \cdot \langle \mathbf{x}_m, \mathbf{u}_m \rangle \right)^2 - \sum_{m=1}^M \alpha_m \cdot \langle \mathbf{x}_m, \mathbf{u}_m \rangle^2. \tag{359}$$

Using Jensen inequality, $\mathbf{X}^\intercal \cdot \mathbf{H} \cdot \mathbf{X} \leq 0$. □

K. Distributed Surrogate Optimization with Black-Box Solver

In this section, we cover the scenario, when the local SGD solver used in our algorithms (Alg. 5 and Alg. 9) is replaced by a (possibly non-iterative) black-box solver that is guaranteed to provide a *local inexact solution*. We introduce the following additional notation.

$$\begin{aligned}\psi_{m,t}(\pi, q) \triangleq \frac{1}{n_t}\sum_{i=1}^{n_t} q^{(i)}\left(z_t^{(i)}=m\right)\times\left[\log p_m(\mathbf{x}_t^{(i)})+\log\pi\right]\\ -\frac{1}{n_t}\sum_{i=1}^{n_t} q^{(i)}\left(z_t^{(i)}=m\right)\times\log q^{(i)}\left(z_t^{(i)}=m\right)+c,\end{aligned} \tag{360}$$

and,

$$\phi_{m,t}(\theta, q) \triangleq \frac{1}{n_t}\sum_{i=1}^{n_t} q^{(i)}(z_t^{(i)}=m)\times l\left(h_\theta(\mathbf{x}_t^{(i)}), y_t^{(i)}\right). \tag{361}$$

For distributions $Q_{1:T}=(q_t)_{1\leq t\leq T}$ over $[M]$, $\Pi\in\Delta^{T\times M}$ and $\Theta\in\mathbb{R}^{M\times d}$, we define

$$\begin{aligned}\Psi_t(\pi_t, q_t) &\triangleq \sum_{m=1}^{M}\psi_{m,t}(\pi_{tm}, q_t), \qquad \Phi_t(\Theta, q_t) \triangleq \sum_{m=1}^{M}\phi_{m,t}(\theta_m, q_t)\\ \Psi(\Pi, Q_{1:T}) &\triangleq \sum_{t=1}^{T}\frac{n_t}{n}\Psi_t(\pi_t, q_t); \qquad \Phi(\Theta, Q_{1:T}) \triangleq \sum_{t=1}^{T}\frac{n_t}{n}\Phi_t(\Theta, q_t)\end{aligned} \tag{362}$$

It holds:

$$g_t(\Theta, \pi_t, q_t) = \Phi_t(\Theta, q_t) - \Psi_t(\pi_t, q_t); \qquad g(\Theta, \Pi, Q_{1:T}) = \Phi(\Theta, Q_{1:T}) - \Psi(\Pi, Q_{1:T}) \tag{363}$$

As we mentioned above, the black-solver provides approximate solutions, as captured by the following assumption¿

Assumption 9 (Local inexact solution)**.** *There exists $0<\alpha<1$ such that for $t\in[T]$ and $k>0$,*

$$\Phi_t^k(\Theta_t^k) - \Phi_t^k(\Theta_{t,*}^k) \leq \alpha\times\left(\Phi_t^k(\Theta^{k-1}) - \Phi_t^k(\Theta_{t,*}^k)\right),$$

where $\Phi_t^k(\Theta)=\Phi_t\left(\Theta, q_t^k\right)$ for $\Theta\in\mathbb{R}^{M\times d}$ and $\Theta_{t,*}^k\in\arg\min_{\Theta\in\mathbb{R}^{M\times d}}\Phi_t^k(\Theta)$.

We further assume strong convexity,

Assumption 10. *For $t\in[T]$ and $i\in[n_t]$, we suppose that $\theta\mapsto l\left(h_\theta\left(\mathbf{x}_t^{(i)}\right), y_t^{(i)}\right)$ is μ-strongly convex.*

Assum. 9 is equivalent to the γ-inexact solution used in (Li et al., 2020) (Lemma. K.1), when local functions $(\Phi_t)_{1\leq t\leq T}$ are assumed to be convex. In addition to Assum. 9, we need to have $G^2=0$ in order to ensure the convergence of Alg. 7 and Alg. 3 to a stationary point of f, as shown by (Wang et al., 2020b, Thm. 2).

We analyse Alg. 5 if the `LocalSolver` verifies Assum. 9, and we prove that under Assum. 4-Assum. 10, with $G^2=0$, Alg. 5 converges to a stationary point of f.

First, we prove the following result

Lemma K.1. *Under Assum. 9, 5 and 10, the iterates of Alg. 7 verify for $k>0$ and $t\in[T]$,*

$$\left\|\nabla\Phi_t^k(\Theta_t^k)\right\|_F \leq \sqrt{\alpha}\kappa\left\|\nabla\Phi_t^k(\Theta^{k-1})\right\|_F,$$

where $\kappa=L/\mu$.

Proof. Since Φ_t^k is L-smooth, we have

$$\left\|\nabla\Phi_t^k\left(\Theta_t^k\right)\right\|_F^2 \leq 2L\left(\Phi_t^k\left(\Theta_t^k\right) - \Phi_t^k\left(\Theta_{t,*}^k\right)\right) \tag{364}$$

$$\leq 2L\alpha\left(\Phi_t^k\left(\Theta_t^{k-1}\right) - \Phi_t^k\left(\Theta_{t,*}^k\right)\right). \tag{365}$$

Since Φ_t^k is μ-strongly convex, we can use Polyak-Lojasiewicz (PL) inequality,

$$\Phi_t^k\left(\Theta_t^{k-1}\right) - \frac{1}{2\mu}\left\|\nabla\Phi_t^k\left(\Theta^{k-1}\right)\right\|_F^2 \leq \Phi_t^k\left(\Theta_t^{k-1}\right), \tag{366}$$

thus,

$$2\mu\left(\Phi_t^k\left(\Theta_t^{k-1}\right) - \Phi_t^k\left(\Theta_{t,*}^k\right)\right) \leq \left\|\nabla\Phi_t^k\left(\Theta^{k-1}\right)\right\|_F^2. \tag{367}$$

Combining Eq. (365) and Eq. (367), we have

$$\left\|\nabla\Phi_t^k\left(\Theta^{k-1}\right)\right\|_F^2 \leq \frac{L}{\mu}\alpha\left\|\nabla\Phi_t^{k-1}\left(\Theta^{k-1}\right)\right\|_F^2, \tag{368}$$

thus,

$$\left\|\nabla\Phi_t^k(\Theta_t^k)\right\|_F \leq \sqrt{\alpha}\kappa\left\|\nabla\Phi_t^k(\Theta^{k-1})\right\|_F. \tag{369}$$

□

Lemma K.2. *Suppose that Assum. 5, 7, 10 and Assum. 9 hold, and that $G^2 = 0$. Then,*

$$g^k\left(\Theta^k, \Pi^k\right) - g^k\left(\Theta_*^k, \Pi^k\right) \leq \tilde{\alpha} \times \left\{g^k\left(\Theta^{k-1}, \Pi^{k-1}\right) - g^k\left(\Theta_*^k, \Pi^k\right)\right\},$$

where $\tilde{\alpha} = \beta^2\kappa^5\alpha$

Proof. For $k > 0$ and $t \in [T]$, define $\Theta_*^k = \arg\min_\Theta \Phi^k(\Theta)$. Φ_t^k is μ-strongly convex, because it is a convex combination of the μ-strongly convex functions , we have

$$\begin{aligned}
\left\|\Theta_t^k - \Theta_*^k\right\|_F &\leq \frac{1}{\mu}\left\|\nabla\Phi_t^k\left(\Theta_t^k\right) - \nabla\Phi_t^k\left(\Theta_*^k\right)\right\|_F && (370)\\
&\leq \frac{1}{\mu}\left\|\nabla\Phi_t^k\left(\Theta_t^k\right)\right\|_F + \frac{1}{\mu}\left\|\nabla\Phi_t^k\left(\Theta_*^k\right)\right\|_F && (371)\\
&\leq \sqrt{\alpha}\frac{\kappa}{\mu}\left\|\nabla\Phi_t^k\left(\Theta^{k-1}\right)\right\|_F + \frac{1}{\mu}\left\|\nabla\Phi_t^k\left(\Theta_*^k\right)\right\|_F && (372)
\end{aligned}$$

where last inequality is a result of lemma K.1. Using Jensen inequality, we have

$$\begin{aligned}
\left\|\Theta^k - \Theta_*^k\right\|_F &= \left\|\sum_{t=1}^T \frac{n_t}{n}\left(\Theta_t^k - \Theta_*^k\right)\right\|_F && (373)\\
&\leq \sum_{t=1}^T \frac{n_t}{n}\left\|\Theta_t^k - \Theta_*^k\right\|_F && (374)\\
&\leq \sum_{t=1}^T \frac{n_t}{n}\left\{\sqrt{\alpha}\frac{\kappa}{\mu}\left\|\nabla\Phi_t^k\left(\Theta^{k-1}\right)\right\|_F + \frac{1}{\mu}\left\|\nabla\Phi_t^k\left(\Theta_*^k\right)\right\|_F\right\} && (375)
\end{aligned}$$

Using Assum. 7, it follows that

$$\left\|\Theta^k - \Theta_*^k\right\|_F \leq \beta\sqrt{\alpha}\frac{\kappa}{\mu}\left\|\nabla\Phi^k\left(\Theta^{k-1}\right)\right\|_F + \frac{\beta}{\mu}\left\|\nabla\Phi^k\left(\Theta_*^k\right)\right\|_F = \beta\sqrt{\alpha}\frac{\kappa}{\mu}\left\|\nabla\Phi^k\left(\Theta^{k-1}\right)\right\|_F \tag{376}$$

Since Φ^k is L-smooth (lemma J.3), we have

$$\begin{aligned}
\left\|\nabla\Phi^k\left(\Theta^k\right)\right\|_F &= \left\|\nabla\Phi^k\left(\Theta^{k-1}\right) - \nabla\Phi^k\left(\Theta_*^k\right)\right\|_F && (377)\\
&\leq L\left\|\Theta^k - \Theta_*^k\right\|_F && (378)\\
&\leq \beta\sqrt{\alpha}\kappa^2\left\|\nabla\Phi^k\left(\Theta^{k-1}\right)\right\|_F && (379)
\end{aligned}$$

Since Φ^k is strongly convex (Because of Assumption 10), we have

$$\Phi^k\left(\Theta^k\right) - \Phi^k\left(\Theta_*^k\right) \leq \frac{1}{2\mu}\left\|\nabla\Phi^k\left(\Theta^k\right)\right\|_F^2 \leq \frac{\beta^2\alpha\kappa^4}{2\mu}\left\|\nabla\Phi^k\left(\Theta^{k-1}\right)\right\|_F^2 \tag{380}$$

Using the L-smoothness of Φ^k (lemma J.3) we have

$$\left\|\nabla\Phi^k\left(\Theta^{k-1}\right)\right\|_F^2 \leq 2L\left(\Phi^k\left(\Theta^{k-1}\right) - \Phi^k\left(\Theta_*^k\right)\right) \tag{381}$$

Thus,

$$\Phi^k\left(\Theta^k\right) - \Phi^k\left(\Theta_*^k\right) \leq \underbrace{\beta^2\kappa^5\alpha}_{\triangleq\tilde{\alpha}}\left(\Phi^k\left(\Theta^{k-1}\right) - \Phi^k\left(\Theta_*^k\right)\right) \tag{382}$$

Moreover, by definition, we have that

$$\Psi_t^k\left(\pi_t^k\right) \triangleq \Psi_t\left(\pi_t^k, q^k\right) \geq \Psi_t^k\left(\pi_t^{k-1}\right) \tag{383}$$

Thus,

$$g^k\left(\Theta^k, \Pi^k\right) - g^k\left(\Theta_*^k, \Pi^k\right) \leq \tilde{\alpha} \times \left\{g^k\left(\Theta^{k-1}, \Pi^{k-1}\right) - g^k\left(\Theta_*^k, \Pi^k\right)\right\} \tag{384}$$

□

Lemma K.3. *Suppose that Assum. 1, 4, 10 and Assum. 5 hold and*

$$g^k(\Theta^k, \Pi^k) \leq g^k(\Theta^{k-1}, \Pi^{k-1});\ k > 0,$$

then

$$r^k(\Theta^k, \Pi^k) \xrightarrow[k\to+\infty]{} 0 \tag{385}$$

$$\sum_{t=1}^{T}\frac{n_t}{n}\mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) \xrightarrow[k\to+\infty]{} 0 \tag{386}$$

$$\left\|\nabla_\Theta r^k(\Theta^k, \Pi^k)\right\|_F^2 \xrightarrow[k\to+\infty]{} 0 \tag{387}$$

If we moreover suppose that there exists $0 < \tilde{\alpha} < 1$ such that for all $k > 0$,

$$g^k(\Theta^k, \Pi^k) - g^k(\Theta_*^k, \Pi^k) \leq \tilde{\alpha} \times \left\{g^k(\Theta^{k-1}, \Pi^{k-1}) - g^k(\Theta_*^k, \Pi^k)\right\},$$

then,

$$\left\|\Theta^k - \Theta_*^k\right\|_F^2 \xrightarrow[k\to+\infty]{} 0 \tag{388}$$

where Θ_*^k is the minimizer of $\Theta \mapsto g^k\left(\Theta, \Pi^k\right)$.

Proof. From Prop. J.5 it follows that, for $k \geq 0$, g^k is a majorant of f and that $g^k(\Theta^{k-1}, \Pi^{k-1}) = f(\Theta^{k-1}, \Pi^{k-1})$. Thus, the following holds,

$$f(\Theta^k, \Pi^k) \leq g^k(\Theta^k, \Pi^k) \leq g^k(\Theta^{k-1}, \Pi^{k-1}) = f(\Theta^{k-1}, \Pi^{k-1}), \tag{389}$$

It follows that the sequence $\left(f\left(\Theta^k, \Pi^k\right)\right)_{k\geq 0}$ is a non-increasing sequence. Since f is bounded below (Assum. 4), it follows that $\left(f\left(\Theta^k, \Pi^k\right)\right)_{k\geq 0}$ is convergent. Denote by f^∞ its limit. The sequence $\left(g^k(\Theta^k, \Pi^k)\right)_{k\geq 0}$ also converges to f^∞.

Proof of Eq. 385 Using again the result of Prop. J.5 and the fact that $g^k(\Theta^k, \Pi^k) \leq g^k(\Theta^{k-1}, \Pi^k)$, we write for $k > 0$

$$f(\Theta^k, \Pi^k) + r^k(\Theta^k, \Pi^k) = g^k(\Theta^k, \Pi^k) \leq g^k(\Theta^{k-1}, \Pi^{k-1}) = f(\Theta^{k-1}, \Pi^{k-1}),$$

Thus,

$$r^k(\Theta^k, \Pi^k) \leq f(\Theta^{k-1}, \Pi^{k-1}) - f(\Theta^k, \Pi^k), \tag{390}$$

By summing over k then passing to the limit when $k \to +\infty$, we have

$$\sum_{k=1}^{\infty} r^k(\Theta^k, \Pi^k) \leq f(\Theta^0, \Pi^0) - f^\infty, \tag{391}$$

Finally since $r^k(\Theta^k, \Pi^k)$ is non negative for $k > 0$ (Prop. J.5), the sequence $\left(r^k(\Theta^k)\right)_{k\geq 0}$ necessarily converges to zero, i.e.,

$$\lim_{k\to\infty} r^k(\Theta^k, \Pi^k) = 0. \tag{392}$$

Proof of Eq. 386 Using Lemma. J.6, with $\Theta = \Theta^{k-1}$ and $\Pi = \Pi^{k-1}$, we write

$$\sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) = g^k\left(\Theta^{k-1}, \Pi^{k-1}\right) - g^k\left(\Theta^{k-1}, \Pi^k\right) \tag{393}$$

$$\leq g^k\left(\Theta^{k-1}, \Pi^{k-1}\right) - g^k\left(\Theta^k, \Pi^k\right) \tag{394}$$

Thus,

$$\sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k, \pi_t^{k-}\right) = f\left(\Theta^{k-1}, \Pi^{k-1}\right) - f\left(\Theta^k, \Pi^k\right) \tag{395}$$

Since $\mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right)$ is non-negative for $k > 0$ and $t \in [T]$, it follows that

$$\lim_{k\to+\infty} \sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) = 0 \tag{396}$$

Proof of Eq.387 Writing the L-smoothness of $\Theta \mapsto r^k\left(\Theta, \Pi^k\right)$ (Prop. J.5) we have

$$r^k\left(\Theta^k - \frac{1}{L}\nabla_\Theta r^k\left(\Theta^k, \Pi^k\right), \Pi^k\right) \leq r^k\left(\Theta^k, \Pi^k\right) - \frac{1}{2L}\left\|\nabla_\Theta r^k\left(\Theta^k, \Pi^k\right)\right\|_F^2 \tag{397}$$

Thus,

$$\left\|\nabla_\Theta r^k\left(\Theta^k, \Pi^k\right)\right\|_F^2 \leq 2L\left(r^k\left(\Theta^k, \Pi^k\right) - r^k\left(\Theta^k - \frac{1}{L}\nabla_\Theta r^k\left(\Theta^k, \Pi^k\right), \Pi^k\right)\right) \tag{398}$$

$$\leq 2L r^k\left(\Theta^k, \Pi^k\right) \tag{399}$$

because r^k is non-negative function (Prop. J.5). Finally, using Eq. 385, it follows that

$$\lim_{k\to\infty} \left\|\nabla r^k(\Theta^k, \Pi^k)\right\|_F^2 = 0. \tag{400}$$

Proof of Eq. 388 We suppose now that there exists $0 < \tilde{\alpha} < 1$ such that

$$\forall k > 0, \;\; g^k(\Theta^k, \Pi^k) - g^k(\Theta_*^k, \Pi^k) \leq \tilde{\alpha}\left(g^k(\Theta^{k-1}, \Pi^{k-1}) - g^k(\Theta_*^k, \Pi^k)\right), \tag{401}$$

It follows that,

$$g^k(\Theta^k, \Pi^k) - \tilde{\alpha} g^k(\Theta^{k-1}, \Pi^{k-1}) \leq (1 - \tilde{\alpha}) g^k(\Theta_*^k, \Pi^k), \tag{402}$$

then,

$$g^k(\Theta_*^k, \Pi^k) \geq \frac{1}{1-\tilde{\alpha}} \times \left[g^k(\Theta^k, \Pi^k) - \tilde{\alpha} \times g^k(\Theta^{k-1}, \Pi^{k-1})\right], \tag{403}$$

and by using the definition of g^k we have,

$$g^k(\Theta_*^k, \Pi^k) \geq \frac{1}{1-\tilde{\alpha}} \times \left[g^k(\Theta^k, \Pi^k) - \tilde{\alpha} \times f(\Theta^{k-1}, \Pi^{k-1})\right], \tag{404}$$

Since $\left\{\Theta_*^k, \Pi^k\right\}$ is a minimizer of g^k, we have,

$$g^k(\Theta_*^k, \Pi^k) \leq g^k(\Theta^{k-1}, \Pi^{k-1}) = f(\Theta^{k-1}, \Pi^{k-1}) \tag{405}$$

From Eq. 404 and Eq. 405, it follows that,

$$\frac{1}{1-\tilde{\alpha}} \times \left[g^k(\Theta^k, \Pi^k) - \tilde{\alpha} \times f(\Theta^{k-1}, \Pi^{k-1})\right] \leq g^k(\Theta_*^k, \Pi^k) \leq f(\Theta^{k-1}, \Pi^{k-1}), \tag{406}$$

Finally, since $f\left(\Theta^{k-1}, \Pi^{k-1}\right) \xrightarrow[k\to+\infty]{} f^\infty$ and $g^k\left(\Theta^k, \Pi^k\right) \xrightarrow[k\to+\infty]{} f^\infty$, it follows from Eq. 406 that,

$$\lim_{k\to\infty} g^k\left(\Theta_*^k, \Pi^k\right) = f^\infty. \tag{407}$$

Since g^k is μ-strongly convex in Θ (Assum. 10), we write

$$\frac{\mu}{2}\left\|\Theta^k - \Theta_*^k\right\|_F^2 \leq g^k\left(\Theta^k, \Pi^k\right) - g^k\left(\Theta_*^k, \Pi^k\right), \tag{408}$$

It follows that,

$$\lim_{k\to+\infty} \left\|\Theta^k - \Theta_*^k\right\|_F^2 = 0 \tag{409}$$

□

Finally, we can prove the main result of this section by combining the previous lemmas

Proposition K.4. *Suppose that Assum. 1, 4, 5, 7, 10 and Assum. 9 hold, with $G^2 = 0$ and $\alpha \leq \frac{1}{\beta^2\kappa^5}$. Then the updates of federated surrogate optimization converges to a stationary point of f, i.e.,*

$$\lim_{k\to+\infty} \left\|\nabla_\Theta f(\Theta^k, \Pi^k)\right\|_F^2 = 0.$$

and,

$$\lim_{k\to+\infty} \sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) = 0.$$

Proof. For $k > 0$, we write

$$f(\Theta^k, \Pi^k) = g^k(\Theta^k, \Pi^k) - r^k(\Theta^k, \Pi^k) \tag{410}$$

Computing the gradient norm, we have,

$$\left\|\nabla f(\Theta^k, \Pi^k)\right\|_F = \left\|\nabla g^k(\Theta^k, \Pi^k) - \nabla r^k(\Theta^k, \Pi^k)\right\|_F \tag{411}$$

$$\leq \left\|\nabla g^k(\Theta^k, \Pi^k)\right\|_F + \left\|\nabla r^k(\Theta^k, \Pi^k)\right\|_F \tag{412}$$

Since g^k is L-smooth in Θ, we write

$$\left\|\nabla g^k(\Theta^k, \Pi^k)\right\|_F = \left\|\nabla g^k(\Theta^k, \Pi^k) - \nabla g^k(\Theta_*^k, \Pi^k)\right\|_F \tag{413}$$

$$\leq L\left\|\Theta^k - \Theta_*^k\right\|_F \tag{414}$$

Thus by replacing Eq. 414 in Eq. 412, we have

$$\left\|\nabla f(\Theta^k, \Pi^k)\right\|_F \leq L^2 \left\|\Theta^k - \Theta^k_*\right\|_F^2 + \left\|\nabla r^k(\Theta^k, \Pi^k)\right\|_F \tag{415}$$

Using Lemma K.2, there exists $0 < \tilde{\alpha} < 1$, such that

$$\left[g^k(\Theta^k, \Pi^k) - g^k(\Theta^k_*, \Pi^k)\right] \leq \tilde{\alpha} \times \left[g^k(\Theta^{k-1}, \Pi^{k-1}) - g^k(\Theta^k_*, \Pi^k)\right] \tag{416}$$

Thus, the conditions of Lemma K.3 hold in expectation, and we can use Eq. 387 and 388, i.e.

$$\left\|\nabla r^k(\Theta^k, \Pi^k)\right\|_F^2 \xrightarrow[k\to+\infty]{} 0 \tag{417}$$

$$\left\|\Theta^k - \Theta^k_*\right\|_F^2 \xrightarrow[k\to+\infty]{} 0 \tag{418}$$

Finally, combining this with Eq. 415, we get the final result

$$\lim_{k\to+\infty} \left\|\nabla f(\Theta^k, \Pi^k)\right\|_F = 0 \tag{419}$$

Moreover Eq. 386 leads to

$$\lim_{k\to+\infty} \sum_{t=1}^{T} \frac{n_t}{n} \mathcal{KL}\left(\pi_t^k, \pi_t^{k-1}\right) = 0. \tag{420}$$

□

Table 3. Datasets and models.

DATASET	TASK	CLIENTS	TOTAL SAMPLES	MODEL
FEMNIST	HANDWRITTEN CHARACTER RECOGNITION	359	98,761	2-LAYER CNN
EMNIST	HANDWRITTEN CHARACTER RECOGNITION	100	81,425	2-LAYER CNN
CIFAR10	IMAGE CLASSIFICATION	80	60,000	MOBILENET-V2
CIFAR100	IMAGE CLASSIFICATION	100	60,000	MOBILENET-V2
SHAKESPEARE	NEXT-CHARACTER PREDICTION	778	4,226,158	STACKED-LSTM
SYNTHETIC	BINARY CLASSIFICATION	300	1,570,507	LINEAR MODEL

L. Details on Experimental Setup

L.1. Datasets and Models

In this section we provide detailed description of the datasets and models used in our experiments. We used a synthetic dataset, verifying assumptions 1-3, and five "real" datasets (CIFAR-10/CIFAR-100 (Krizhevsky, 2009), sub part of EMNIST (Cohen et al., 2017), sub part of FEMNIST (Caldas et al., 2018; McMahan et al., 2017) and Shakespeare (Caldas et al., 2018; McMahan et al., 2017)) from which, two (FEMNIST and Shakespeare) has natural clients partitioning. Below, we give a detailed description of the datasets and the models / tasks considered for each of them. Table 3 summarizes datasets, models, and number of clients.

L.1.1. CIFAR-10 / CIFAR-100

The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They both share the same $60,000$ input images. CIFAR-100 has a finer labeling, with 100 unique labels, in comparison to CIFAR-10, having 10 unique label. We used Dirichlet allocation (Wang et al., 2020a), with parameter $\alpha = 0.4$ to partition CIFAR-10 among 80 clients. We used Pachinko allocation (Reddi et al., 2021) with parameters $\alpha = 0.4$ and $\beta = 10$ to partition CIFAR-100 on 100 clients. For both of them we train MobileNet-v2 (Sandler et al., 2018) architecture with an additional linear layer. We used TorchVision (Marcel & Rodriguez, 2010) implementation of MobileNet-v2.

L.1.2. EMNIST

EMNIST (Extended MNIST) is a 62-class image classification dataset, extending the classic MNIST dataset. In our experiments, we consider 10% of the EMNIST dataset, that we partition using Dirichlet allocation of parameter $\alpha = 0.4$ over 100 clients. We train the same convolutional network as in (Reddi et al., 2021). The network has two convolutional layers (with 3×3 kernels), max pooling, and dropout, followed by a 128 unit dense layer.

L.1.3. FEMNIST

FEMNIST (Federated Extended MNIST) is A 62-class image classification dataset built by partitioning the data of Extended MNIST based on the writer of the digits/characters. In our experiments, we used a subset with 15% of the total number of writers in FEMNIST. For each one of them we kept 80% of the data for training and we kept 20% for test. We train the same convolutional network as in (Reddi et al., 2021). The network has two convolutional layers (with 3×3 kernels), max pooling, and dropout, followed by a 128 unit dense layer.

L.1.4. SHAKESPEARE

This dataset is built from The Complete Works of William Shakespeare and is partitioned by the speaking roles (McMahan et al., 2017). In our experiments, we discarded roles with less then two sentences. We consider character-level based language modeling on this dataset. The model takes as input a sequence of 200 English characters and predicts the next character. The model embeds the 80 characters into a learnable 8 dimensional embedding space, and uses two stacked-LSTM layers with 256 hidden units, followed by a densely-connected layer. We also normalized each character by its frequency of appearance.

Table 4. Average computation time and used GPU for each dataset.

DATASET	GPU	SIMULATION TIME
SHAKESPEARE	QUADRO RTX 8000	4H42MIN
FEMNIST	QUADRO RTX 8000	1H14MIN
EMNIST	GEFORCE GTX 1080 TI	46MIN
CIFAR10	GEFORCE GTX 1080 TI	2H37MIN
CIFAR100	GEFORCE GTX 1080 TI	3H9MIN
SYNTHETIC	GEFORCE GTX 1080 TI	20MIN

L.2. Implementation Details

L.2.1. MACHINES

We ran the experiments on a CPU/GPU cluster, with different GPUs available (e.g., Nvidia Tesla V100, GeForce GTX 1080 Ti, Titan X, Quadro RTX 6000, and Quadro RTX 8000). Most experiments with CIFAR10/CIFAR-100 and EMNIST were run on GeForce GTX 1080 Ti cards, while most experiments with Shakespeare and FEMNIST were run on the Quadro RTX 8000 cards. For each dataset, we ran around 30 experiments (not counting the development/debugging time), Table 4 gives the average amount of time needed to run one simulation for each dataset. The time needed per simulation was extremely long for Shakespeare dataset, because we used a batch size of 128. We remarked that increasing the batch size beyond 128 caused the model to converge to poor local minima, where the model keeps predicting a white space as next character.

L.2.2. LIBRARIES

We used PyTorch (Paszke et al., 2019) to build and train our models. We also used Torchvision (Marcel & Rodriguez, 2010) implementation of MobileNet-v2 (Sandler et al., 2018), and for image datasets preprossessing. We used LEAF (Caldas et al., 2018) to build FEMNIST dataset and the federated version of Shakespeare dataset.

L.2.3. HYPERPARAMETERS

For each method and each task, the learning rate was set via grid search on the set $\{10^{-0.5}, 10^{-1}, 10^{-1.5}, 10^{-2}, 10^{-2.5}, 10^{-3}\}$. `FedProx` and `pFedMe`'s penalization parameter μ was tuned via grid search on $\{10^{1}, 10^{0}, 10^{-1}, 10^{-2}, 10^{-3}\}$. For clustered FL, we used the same values of tolerance as the ones used in its official implementation (Sattler et al., 2020). We found tuning $\texttt{tol}_1$ and $\texttt{tol}_2$ particularly hard: no empirical rule is provided in (Sattler et al., 2020), and the few random setting we tried did not show any improvement in comparison to the default ones.

Table 5. Test accuracy: average across clients.

DATASET	LOCAL	FEDAVG	FEDAVG+	CLUSTERED FL	PFEDME	FEDEM (OURS)	D-FEDEM (OURS)
FEMNIST	71.0	78.6	75.3	73.5	74.9	**79.9**	77.2
EMNIST	71.9	82.6	83.1	82.7	83.3	**83.5**	**83.5**
CIFAR10	70.2	78.2	82.3	78.6	81.7	**84.3**	77.0
CIFAR100	31.5	40.9	39.0	41.5	41.8	**44.1**	43.9
SHAKESPEARE	32.0	**46.7**	40.0	46.6	41.2	**46.7**	45.4
SYNTHETIC	65.7	68.2	68.9	69.1	69.2	**74.7**	73.8

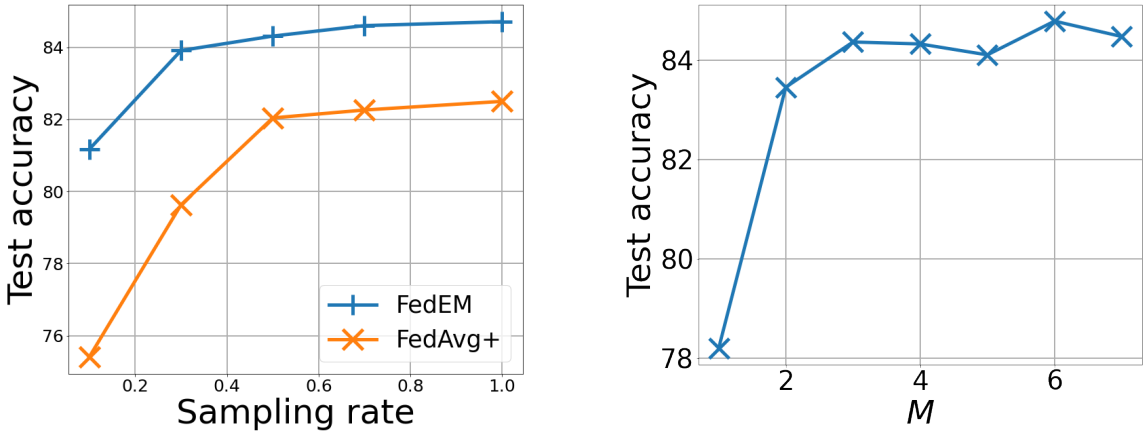

Figure 1. Effect of client sampling rate (left) and FedEM number of mixture components M (right) on the test accuracy for CIFAR10 (Krizhevsky, 2009).

M. Additional Experimental Results

M.1. Fully Decentralized Federated Expectation-Maximization

D-FedEM considers the scenario where clients communicate directly in a peer-to-peer fashion instead of relying on the central server mediation. In order to simulate D-FedEM, we consider a binomial Erdős-Rényi graph (Erdös & Rényi, 1959) with parameter $p = 0.5$, and we set the mixing weight using *Fast Mixing Markov Chain* (Boyd et al., 2003) rule. We report the result of this experiment in Table 5, showing the average weighted accuracy with weight proportional to local dataset sizes. We observe that D-FedEM often performs better than other FL approaches and slightly worst than FedEM, except on CIFAR-10 where it has low performances.

M.2. Generalization to Unseen Clients

Table 4 shows that FedEM allows new clients to learn a personalized model at least as good as FedAvg's global one and always better than FedAvg+'s one. Unexpectedly, new clients achieve sometimes a significantly higher test accuracy than old clients (e.g., 47.5% against 44.1% on CIFAR100).

In order to better understand this difference, we looked at the distribution of FedEM personalized weights for the old clients and new ones. The average distribution entropy equals 0.27 and 0.92 for old and new clients, respectively. This difference shows that old clients tend to have more skewed distribution, suggesting that some components may be overfitting the local training dataset leading the old clients to give them a high weight.

M.3. Effect of M

A limitation of FedEM is that each client needs to update and transmit M components at each round, requiring roughly M times more computation and M times larger messages. Nevertheless, the number of components to consider in practice is

Table 6. Test and train accuracy comparison across different tasks. For each method, the best test accuracy is reported. For FedEM we run only $\frac{K}{M}$ rounds, where K is the total number of rounds for other methods–$K = 80$ for Shakespeare and $K = 200$ for all other datasets–and $M = 3$ is the number of components used in FedEM.

DATASET	LOCAL	FEDAVG	FEDPROX	FEDAVG+	CLUSTERED FL	PFEDME	FEDEM (OURS)
FEMNIST	71.0 (99.2)	**78.6** (79.5)	78.6 (79.6)	75.3 (86.0)	73.5 (74.3)	74.9 (91.9)	74.0 (80.9)
EMNIST	71.9 (99.9)	82.6 (86.5)	82.7 (86.6)	83.1 (93.5)	82.7 (86.6)	**83.3** (91.1)	82.7 (89.4)
CIFAR10	70.2 (99.9)	78.2 (96.8)	78.0 (96.7)	82.3 (98.9)	78.6 (96.8)	81.7 (99.8)	**82.5** (92.2)
CIFAR100	31.5 (99.9)	41.0 (78.5)	40.9 (78.6)	39.0 (76.7)	41.5 (78.9)	41.8 (99.6)	**42.0** (72.9)
SHAKESPEARE	32.0 (95.3)	**46.7** (48.7)	45.7 (47.3)	40.0 (93.1)	46.6 (48.7)	41.2 (42.1)	43.8 (44.6)
SYNTHETIC	65.7 (91.0)	68.2 (68.7)	68.2 (68.7)	68.9 (71.0)	69.1 (85.1)	69.2 (72.8)	**73.2** (74.7)

quite limited. We used $M = 3$ in our previous experiments, and Fig. 1 (right) shows that larger values do not yield much improvement and $M = 2$ already provides a significant level of personalization. In all experiments above, the number of communication rounds allowed all approaches to converge. As a consequence, even if other methods trained over $M = 3$ times more rounds—in order to have as much computation and communication as FedEM—the conclusions would not change. As a final experiment, we considered a time-constrained setting, where FedEM is limited to run one third ($= 1/M$) of the rounds (Table 6 in App. M.4). Even if FedEM does not reach its maximum accuracy, it still outperforms the other methods on 3 datasets.

M.4. Effect of M in Time-Constrained Setting

Recall that in FedEM, each client needs to update and transmit M components at each round, requiring roughly M times more computation and M times larger messages than the competitors in our study. In this experiment, we considered a challenging time-constrained setting, where FedEM is limited to run one third ($= 1/M$) of the rounds of the other methods. The results in Table 6 show that even if FedEM does not reach its maximum accuracy, it still outperforms the other methods on 3 datasets.

M.5. Full Results

Figures 2 to 7 show the evolution of average train loss, train accuracy, test loss, and test accuracy over time for each experiment.

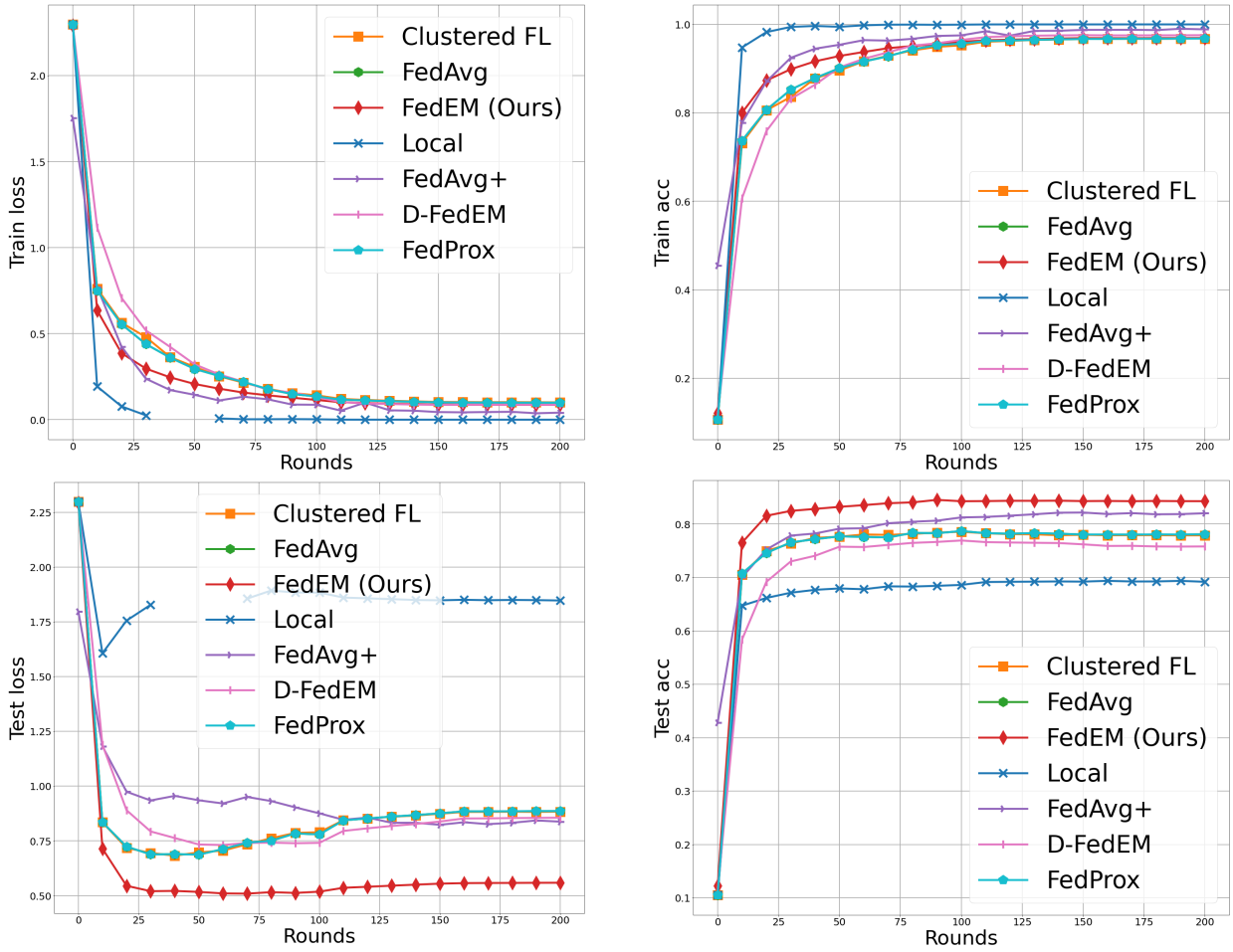

Figure 2. Train loss, train accuracy, test loss, and test accuracy for CIFAR10 (Krizhevsky, 2009). .

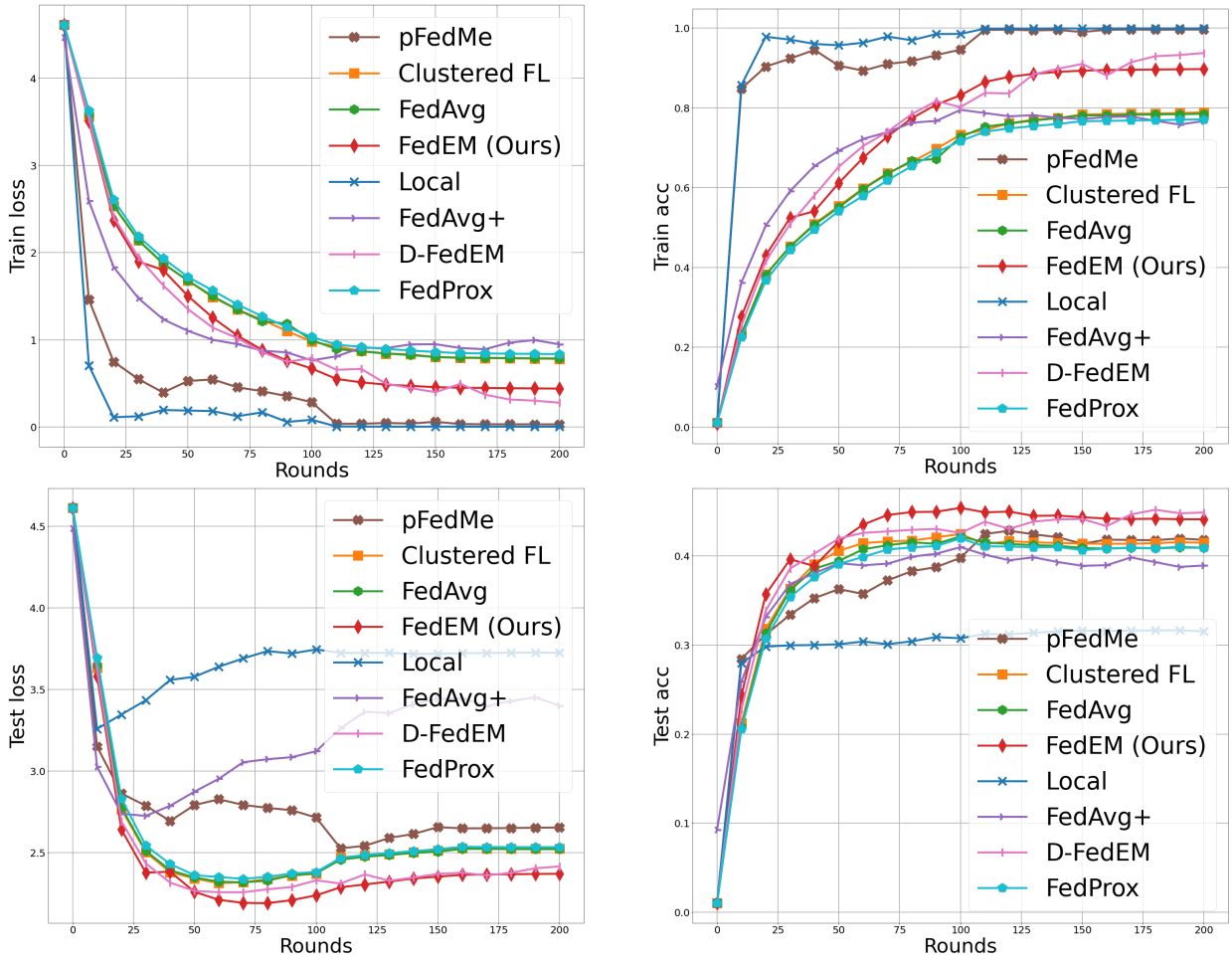

Figure 3. Train loss, train accuracy, test loss, and test accuracy for CIFAR100 (Krizhevsky, 2009).

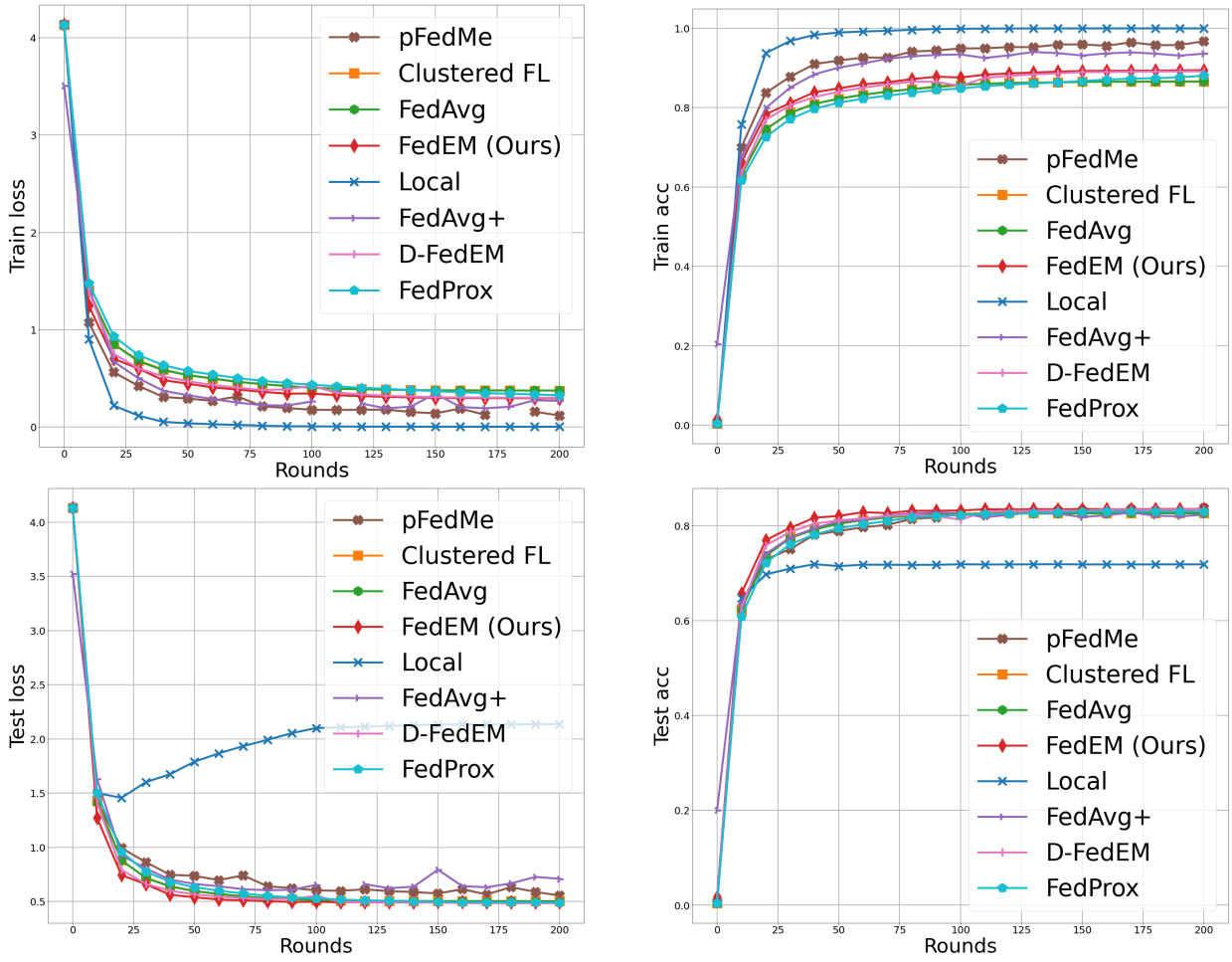

Figure 4. Train loss, train accuracy, test loss, and test accuracy for EMNIST (Cohen et al., 2017).

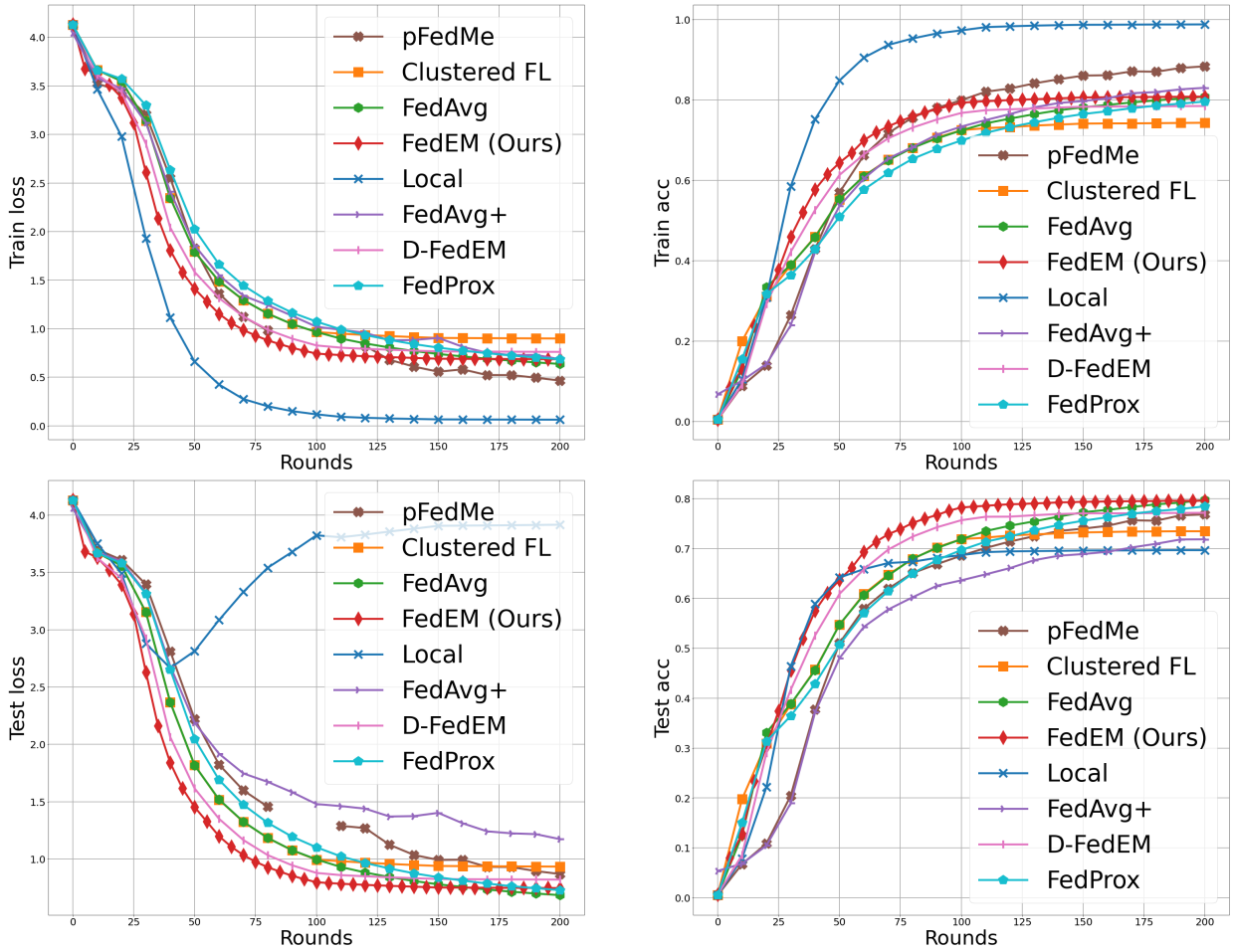

Figure 5. Train loss, train accuracy, test loss, and test accuracy for FEMNIST (Caldas et al., 2018; McMahan et al., 2017).

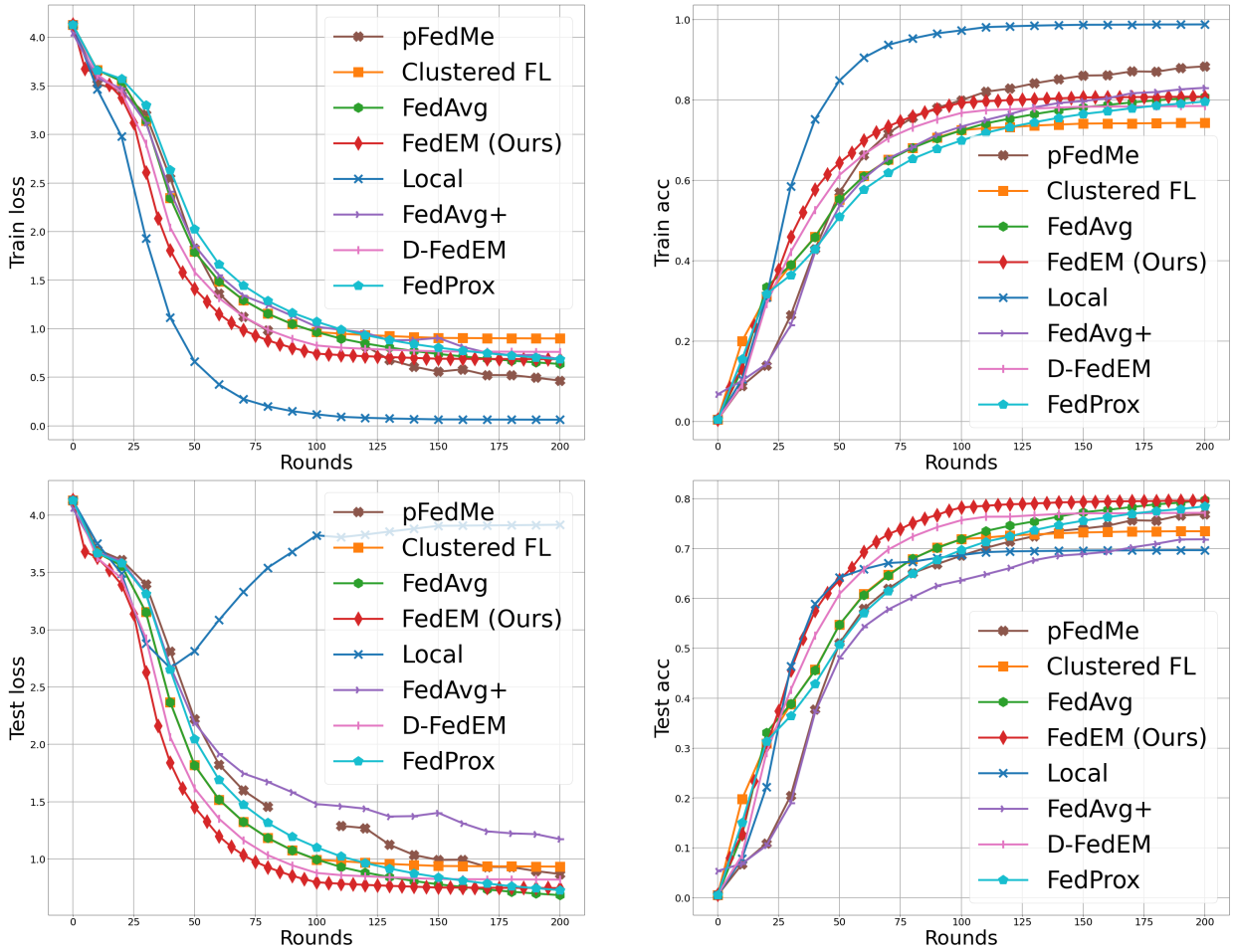

Figure 6. Train loss, train accuracy, test loss, and test accuracy for Shakespeare (Caldas et al., 2018; McMahan et al., 2017).

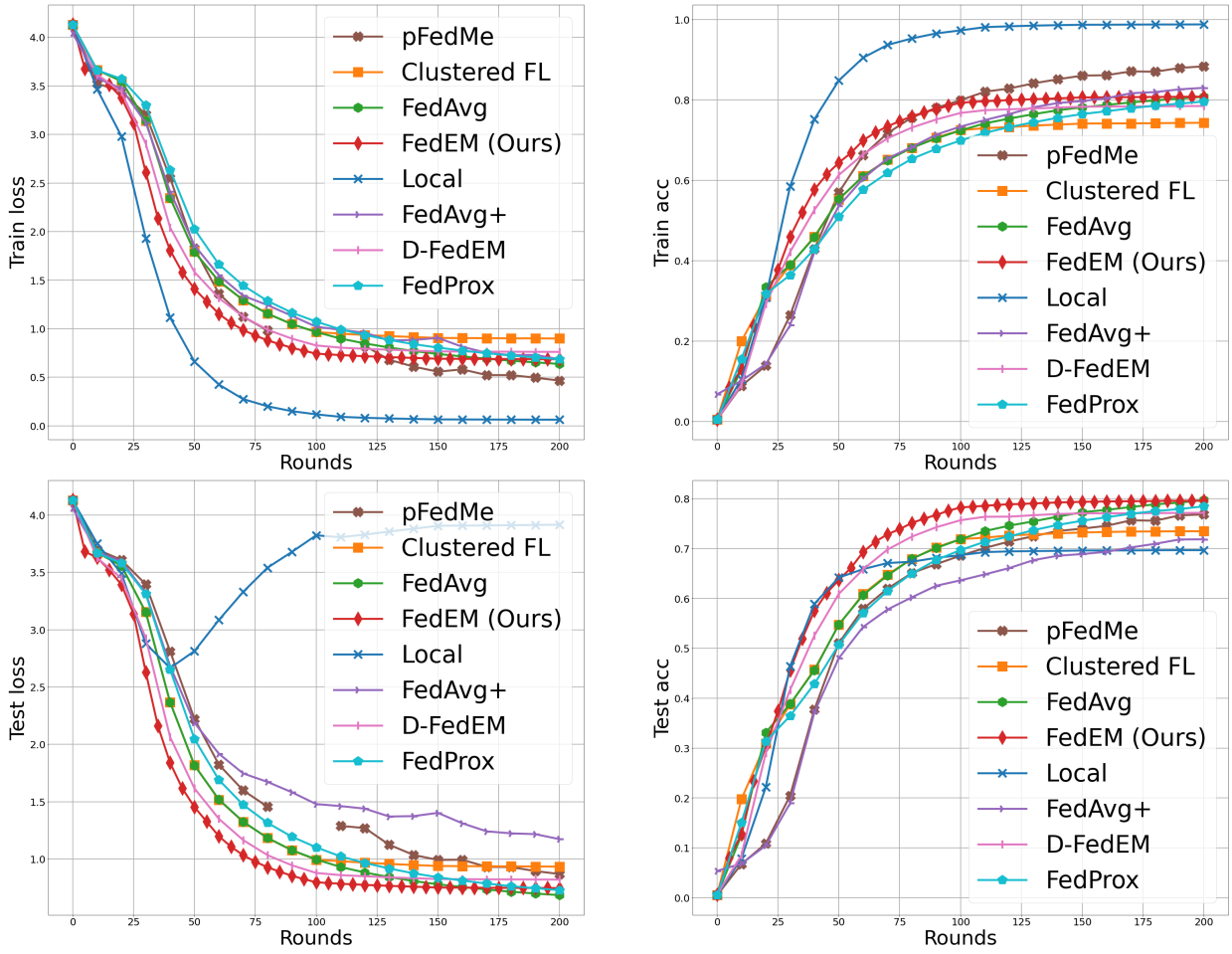

Figure 7. Train loss, train accuracy, test loss, and test accuracy for synthetic dataset.