Large Stepsizes Federated Learning on Logistic Regression with Linearly Separable Data: The Case of Heterogeneous Devices

Hok Fong Wong, Hoi-To Wai, and Chung-Yiu Yau

***Abstract*— This paper revisits the distributed learning problem for training a multinomial logistic regression model with the Federated Averaging (`FedAvg`) algorithm. We concentrate on a scenario with arbitrarily large stepsizes and heterogeneous update rules where the devices may perform a different number of local updates in each round. We show that, with linearly separable data, `FedAvg` is stable with any stepsizes and the objective values converge to zero at the rate of $\mathcal{O}(1/R)$, where R is the number of communication rounds. Our result also demonstrates that the effects of device heterogeneity vanish asymptotically. For sufficiently large R, the objective values decrease monotonically and is bounded by $\mathcal{O}(1/(RT_{\rm avg}))$, where $T_{\rm avg}$ is the average number of local update steps per communication round across devices. Numerical experiments support our findings.**

I. Introduction

Distributed optimization has been a major workhorse in multi-agent control and learning systems [1]–[3], where devices (a.k.a. agents) collaboratively optimize a global objective function by interleaving between local computation and communication. In recent years, federated learning (FL) has gained popularity as a mainstream distribtuted optimization paradigm where it accounts for heterogeneous devices operating in an asynchronous manner [4], [5].

The challenges of heterogeneous devices have set the study of FL apart from traditional distributed optimization by workers performing local updates before global aggregation of their local solutions, introducing *local drift* error that results from data and computational heterogeneity. The issue of local drift in FL has inspired the studies of algorithms such as `FedAvg` [6]–[8], `SCAFFOLD` [9], `FedPD` [10], and `FedProx` [11]. For instance, `FedAvg` relies on a simple averaging of local updates, and `SCAFFOLD`, `FedPD`, `FedProx` utilize control variates or proximal updates to reduce local drift errors. Convergence is guaranteed together with sufficiently small stepsizes with respect to (w.r.t.) problem-specific parameters such as smoothness.

On the other hand, several recent researches found that gradient algorithms may still converge with large stepsizes [12]–[15]. These pioneering works showed that for logistic regression with linearly separable data, gradient descent with arbitrarily large stepsizes stops oscillating after a transient phase, and even exhibits an accelerated convergence rate. This marks a significant departure from the classical analysis that demands small stepsizes to ensure monotonic descent.

In light of the above developments, this paper investigates the impact of large stepsizes on FL. We revisit the convergence of `FedAvg` with large stepsizes for logistic regression with linearly separable data. We are particularly interested in the effects of device computational heterogeneity, i.e., when the number of local update steps varies across devices, on the convergence of `FedAvg` with large stepsizes. Notice that a closely related work is the recent paper [16] which studied a special case of `FedAvg` with only homogeneous devices taking the same number of local updates. We inquire:

Does the convergence of `FedAvg` *with large stepsizes still hold under heterogeneous local updates, and how does device heterogeneity affect the convergence bound?*

To address the above question, we notice that the introduction of *heterogeneous devices* in local updates in `FedAvg` has led to a nontrivial complex dynamics with large stepsizes. Our main contributions are summarized below as:

- Through utilizing a new proof technique, we show that `FedAvg` with large stepsizes and heterogeneous devices is stable and the objective values converge to zero at $\mathcal{O}(1/R)$, where R is the number of communication rounds.
- We demonstrate that the effects of device heterogeneity on the convergence rate vanish asymptotically. For $R \gg 1$, the convergence of optimality gap in objective values behaves as $\mathcal{O}(1/(RT_{\rm avg}))$, where $T_{\rm avg}$ is the average number of local update steps per communication round.
- Our numerical experiments demonstrate that the asymptotic convergence of `FedAvg` with large stepsizes is not sensitive to the degree of device heterogeneity, corroborating with our theoretical findings.

We also developed our analysis for the multinomial logistic regression setting, extending from prior works such as [12].

Notations. For a vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes its ℓ_2-norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_{\rm op}$ denote its Frobenius norm and operator norm, respectively. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ for any positive integer n. We use $\mathcal{O}(\cdot)$ to hide absolute constants independent of problem parameters.

II. Problem Setup and FedAvg Algorithm

Consider a distributed optimization setup with a central server and m devices. Each device i has a local dataset $\mathcal{D}_i = \{(\boldsymbol{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, where $\boldsymbol{x}_{i,j} \in \mathbb{R}^d$ is the feature vector and $y_{i,j} \in [K]$ is the corresponding label. Parameterizing the multinomial linear model by $\mathbf{W} = [\mathbf{w}_1^\top; \mathbf{w}_2^\top; \ldots; \mathbf{w}_K^\top] \in$

Hok Fong Wong is with Department of CSE, The Chinese University of Hong Kong, Hong Kong SAR `hfwong2@cse.cuhk.edu.hk`.

Hoi-To Wai is with Department of SEEM, The Chinese University of Hong Kong, Hong Kong SAR `htwai@se.cuhk.edu.hk`.

Chung-Yiu Yau is with Department of ECE, University of Minnesota, Minnesota, USA `cyau@umn.edu`.

Algorithm 1 `FedAvg` Algorithm

Input: Initial model $\mathbf{W}^{(0)}$, stepsizes $\eta_l > 0, \eta_g > 0$, maximum number of communication rounds R.
for $r = 1$ to R **do**
  Server broadcasts $\mathbf{W}^{(r-1)}$ to all devices.
  for each device i in parallel **do**
    Set $\mathbf{D}_{i,0}^{(r)} = \mathbf{W}^{(r-1)}$.
    for $t = 0$ to $T_i - 1$ **do**
      Update $\mathbf{D}_{i,t+1}^{(r)} = \mathbf{D}_{i,t}^{(r)} - \eta_l \nabla \mathcal{L}_i(\mathbf{D}_{i,t}^{(r)})$.
    end for
    Send $\Delta \mathbf{D}_i^{(r)} = \mathbf{D}_{i,T_i}^{(r)} - \mathbf{W}^{(r-1)}$ to server.
  end for
  Server updates $\mathbf{W}^{(r)} = \mathbf{W}^{(r-1)} + \eta_g \sum_{i=1}^m \frac{n_i}{n} \Delta \mathbf{D}_i^{(r)}$.
end for

$\mathbb{R}^{K\times d}$, our goal is to tackle the following problem:

$$\inf_{\mathbf{W}\in\mathbb{R}^{K\times d}} \mathcal{L}(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^m n_i \mathcal{L}_i(\mathbf{W}), \tag{1}$$

where $n := \sum_{i=1}^m n_i$, $\mathcal{L}_i(\mathbf{W}) := \frac{1}{n_i}\sum_{j=1}^{n_i} \ell_{i,j}(\mathbf{W})$ with

$$\ell_{i,j}(\mathbf{W}) := \log\left(\frac{1}{\widehat{y}_{i,j|\mathbf{W}}}\right), \ \widehat{y}_{i,j|\mathbf{W}} := \frac{\exp(\mathbf{w}_{y_{i,j}}^\top \boldsymbol{x}_{i,j})}{\sum_{k=1}^K \exp(\mathbf{w}_k^\top \boldsymbol{x}_{i,j})}.$$

Notice that $\ell_{i,j}(\mathbf{W}) \geq 0$ is the cross-entropy loss for the j-th data point on device i and tackling (1) trains a K-classes linear classifier $\mathbf{W}$.

In this work, we concentrate on the federated averaging (`FedAvg`) algorithm [4] with heterogeneous local updates. As summarized in Algorithm 1, the algorithm mimicks the behavior of centralized gradient descent while allowing the devices to perform local update using only their own data. The difference matrix $\Delta \mathbf{D}_i^{(r)}$ is controlled by the local step size η_l and accumulates the local updates from device i in each round. This difference matrix will be sent to the server at round r, where the latter aggregates the updates with a global stepsize η_g. Depending on the discrepencies in hardware and network conditions, the heterogeneous devices may perform different number of local updates T_i in each communication round. We assume that $T_i \geq 1$ for all i. Notice that if $T_i = 1$ for all i, then the evolution of $\{\mathbf{W}^{(r)}\}_{r\geq 0}$ in `FedAvg` reduces to that of the (centralized) gradient descent algorithm on minimizing $\mathcal{L}(\mathbf{W})$.

The convergence of `FedAvg` has been widely studied, e.g., in [6]–[9]. A main technical challenge is to account for the device drift behavior when $T_i \geq 2$ and the resultant device heterogeneity that differentiate the algorithm from that of centralized gradient descent. As such, the majority of existing analysis focused on the small stepsizes regime with stochastic gradients and identical number of local steps $T_i = T$ for all i. For instance, [9, Theorem I] requires that $\eta_g \eta_l = \mathcal{O}(1/(L \max_i T_i))$, where L is the smoothness constant of $\mathcal{L}_i(\cdot)$, and [8] demonstrated that when $T_i \neq T_j$ for some $i \neq j$, `FedAvg` converges to a biased optimal solution depending on a reweighted objective function. On the other hand, the recent work [16] showed a new convergence result of `FedAvg` for arbitrarily large stepsizes applied to logistic regression with linearly separable data. However, a limitation of their analysis is that it only applies to homogeneous devices where $T_i = T$ for all i, as well as the binary classification setting with $K = 2$.

The above motivated us to extend the analysis with large stepsize `FedAvg` to heterogeneous devices. To describe the setup, our analysis relies on the following standing assumptions inspired by [12], [16]. We require:

Assumption 1 (Bounded Feature Norms)**.** All data samples $\{\mathbf{x}_{i,j}, y_{i,j}\}$, where $i \in [n_j]$, satisfy $\|\mathbf{x}_{i,j}\| \leq 1$.

Assumption 2 (Global Linear Separability)**.** There exists $\gamma > 0$, $\mathbf{W}_\star = [\mathbf{w}_{\star,1}^\top; \mathbf{w}_{\star,2}^\top; \ldots; \mathbf{w}_{\star,K}^\top]$ with $\|\mathbf{W}_\star\|_F \leq 1$ such that for all data samples $(\mathbf{x}_{i,j}, y_{i,j})$ and $k \neq y_{i,j}$,

$$(\mathbf{w}_{\star,y_{i,j}} - \mathbf{w}_{\star,k})^\top \mathbf{x}_{i,j} \geq \gamma, \ \forall \ i, j.$$

Assumption 1 is a standard regularity condition to avoid potentially unbounded feature vectors. On the other hand, Assumption 2 requires that the (training) data samples are linearly separable by a global solution matrix $\mathbf{W}_\star$. Note that this implies $\inf_{\mathbf{W}} \mathcal{L}_i(\mathbf{W}) = 0$ for all i and $\inf_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = 0$, though the latter may only be approximated by a scaled version of $\mathbf{W}_\star$. We remark that the linearly separability condition is essential for large stepsize convergence analysis [12]–[14], [16]; see [15], [17] for recent developments on non-separable data.

III. Large Stepsize Convergence Analysis

This section presents the main convergence results for large stepsize `FedAvg` and provide details for their proofs.

Define the following constants on the number of local update steps taken by the devices

$$T_{\text{avg}} := \sum_{i=1}^m \frac{n_i T_i}{n}, \quad T_{\max} := \max_{i\in[m]} T_i, \quad T_{\min} := \min_{i\in[m]} T_i.$$

We also set $\bar{\eta} := \eta_g \eta_l$ as the product of local and global stepsizes. Note that $1 \leq T_{\min} \leq T_{\text{avg}} \leq T_{\max}$. Our first result is the following bounds on the average risk and magnitude of the decision variable, i.e.,

Theorem 3. *Under Assumptions 1, 2. For every $\eta_g, \eta_l > 0$ and $R \geq 1$, we have*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathcal{L}(\mathbf{W}^{(r)}) \leq \frac{\|\mathbf{W}^{(0)}\|_F^2}{\bar{\eta} R T_{\min}} + \frac{2(1 + \log^2(\gamma^2 \bar{\eta} R(K-1) T_{\text{avg}}) + \bar{\eta}^2 T_{\text{avg}}^2 + \eta_l^4 T_{\max}^4)}{\bar{\eta}\,\gamma^2 R T_{\min}}. \tag{2}$$

Moreover,

$$\|\mathbf{W}^{(R)}\|_F \leq \|\mathbf{W}^{(0)}\|_F + \frac{\sqrt{2} + \log(\gamma^2 \bar{\eta} R(K-1) T_{\text{avg}})}{\gamma} + \frac{\bar{\eta} T_{\text{avg}} + \eta_l^2 T_{\max}^2}{\gamma}. \tag{3}$$

Unlike the classical analysis of FedAvg [6], [8], [9], the bounds (2), (3) hold *for any* $\eta_g, \eta_l > 0$ and any distribution of $\{T_i\}_{i=1}^m$. It demonstrates that the average risk converges to zero at $\mathcal{O}(1/R)$ with heterogeneous devices. From (2), (3), we also notice that the effect of device heterogeneity is captured in the terms $\eta_l^4 \frac{T_{\max}^4}{T_{\min}}$, $\bar{\eta}^2 \frac{T_{\text{avg}}^2}{T_{\min}}$. Here, (2) shows that larger stepsizes $\eta_l, \bar{\eta}$ can amplify the effect of device heterogeneity. Interestingly, our result also indicates that choosing a smaller local stepsize η_l may improve the performance.

In comparison, the recent work [16, Theorem 4.1] showed a similar bound for FedAvg as (2) under the case of homogeneous devices with $T_i = T$ and binary classification. Similar to [12], they showed that

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathcal{L}(\mathbf{W}^{(r)}) = \mathcal{O}\left(\frac{\log^2(T(1+\bar{\eta}\gamma^2 R)) + \bar{\eta}^2 T^2}{\bar{\eta}\,\gamma^4 R}\right).$$

Considering the special case of (2) with $T_i = T$ and $K = 2$, we note that when $R \gg 1$, the above bound is worse than (2) by a factor of $1/\gamma^2$. As $\gamma < 1$, this suggests that Theorem 3 offers a tighter bound than [16, Theorem 4.1] even for the homogeneous devices case. We remark that the improvement of (2) is potentially due to the new proof structure which we shall develop in Section III-A.

We next show that under an additional condition on the degree of device heterogeneity, FedAvg exhibits a *stable phase* convergence behavior:

Theorem 4. *Under Assumptions 1 and 2, for every $\eta_g, \eta_l > 0$, suppose that there exists $s \geq 0$ such that $\mathcal{L}(\mathbf{W}^{(s)}) \leq \min\{\frac{1}{2n}, \frac{1}{4\bar{\eta}T_{avg}}, \frac{\gamma T_{\text{avg}}}{40n\eta_l T_{\max}^2\sqrt{K}}\}$, and the heterogeneity among clients satisfies $\max_i |T_i/T_{\text{avg}} - 1| \leq \gamma/(20\sqrt{K})$, then*

$$\mathcal{L}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r)}),\ \forall\ r \geq s. \tag{4}$$

The above observation shows that once $\mathcal{L}(\mathbf{W}^{(r)})$ is smaller than a threshold, the FedAvg algorithm enters a stable phase where the objective value decreases monotonically. Notice that (2) guarantees that as $R \gg 1$, there exists $s \in [R]$ such that $\mathcal{L}(\mathbf{W}^{(s)}) = \mathcal{O}(1/R)$ and thus the existence of such iteration index s. Theorem 4 further leads to an asymptotic convergence rate for FedAvg:

Corollary 5. *Under the same assumptions as Theorem 4,*

$$\mathcal{L}(\mathbf{W}^{(r)}) \leq \frac{16}{\bar{\eta}T_{\text{avg}}\gamma^2(r-s)},\ \forall\ r \geq s. \tag{5}$$

The above result strengthens the average risk bound in (2) and shows that the last iterate $\mathcal{L}(\mathbf{W}^{(r)})$ converges as $\mathcal{O}(1/(rT_{\text{avg}}))$ when $r \gg 1$. Importantly, the effects of device heterogeneity vanish asymptotically and the convergence rate is dominated by the average local update steps T_{avg}. Such estimation on the convergence rate is consistent with [8] which showed that FedAvg with small stepsizes converges at a similar rate dependent on T_{avg}. Note that unlike [8], FedAvg in our setting converges to an unbiased solution as $\{\mathcal{L}_i(\cdot)\}_{i=1}^m$ admit a common optimal solution under Assumption 2. We remark that the above results assume the distribution of $\{T_i\}_{i=1}^m$ to satisfy $\max_i |T_i/T_{\text{avg}} - 1| \leq \gamma/(20\sqrt{K})$, e.g., when T_{avg} is large or the variation of $\{T_i\}_{i=1}^m$ is small. As we show in Section IV, such stable phase convergence behavior can be observed in practice even without the above condition.

A. Proof of Theorem 3

Let $\mathbf{W}_\star$ be a solution matrix satisfying Assumption 2. We define an additional notation to facilitate our proof: for any $q \in [K]$ and $\mathbf{W} := [\mathbf{w}_1, \ldots, \mathbf{w}_K]$,

$$\widehat{y}_{i,j|\mathbf{W}}^{(q)} := \frac{\exp(\mathbf{w}_q^\top \boldsymbol{x}_{i,j})}{\sum_{k=1}^K \exp(\mathbf{w}_k^\top \boldsymbol{x}_{i,j})}.$$

Notice that $\sum_{k=1}^K \widehat{y}_{i,j|\mathbf{W}}^{(k)} = 1$. We first observe the following consequence of Assumption 2, summarized as the lemma:

Lemma 6. *Under Assumption 2. We observe the following bounds for any i, j,*

(i) with any $\alpha > 0$,

$$\ell_{i,j}(\alpha\mathbf{W}_\star) \leq (K-1)\exp(-\alpha\gamma). \tag{6}$$

(ii) with any $\mathbf{W}$,

$$\langle \nabla \ell_{i,j}(\mathbf{W}), \mathbf{W}_\star \rangle \leq -\gamma(1 - \widehat{y}_{i,j|\mathbf{W}}). \tag{7}$$

(iii) with any $\mathbf{W}$,

$$\|\nabla \ell_{i,j}(\mathbf{W})\|_F^2 \leq 2(1 - \widehat{y}_{i,j|\mathbf{W}}). \tag{8}$$

The above results are consequences of Assumption 2 and properties of the cross-entropy loss.

Our key step is to control the distance between $\mathbf{W}^{(r)}$ and $\mathbf{U} := \mathbf{U}_1 + \mathbf{U}_2 + \mathbf{U}_3$, defined as

$$\mathbf{U}_1 = \frac{\log((K-1)\gamma^2\bar{\eta}RT_{\text{avg}})}{\gamma}\mathbf{W}_\star,$$
$$\mathbf{U}_2 = \frac{\bar{\eta}T_{\text{avg}}}{\gamma}\mathbf{W}_\star, \mathbf{U}_3 = \frac{\eta_l^2 T_{\max}^2}{\gamma}\mathbf{W}_\star. \tag{9}$$

In particular, we notice that the server update in round r, $\Delta\mathbf{W}^{(r)} = \mathbf{W}^{(r+1)} - \mathbf{W}^{(r)}$, can be written as

$$\boldsymbol{\Delta}\mathbf{W}^{(r)} = -\frac{\bar{\eta}}{n}\sum_{i=1}^m \sum_{\tau=1}^{T_i} n_i \cdot \nabla \mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)}). \tag{10}$$

As such,

$$\|\mathbf{W}^{(r+1)} - \mathbf{U}\|_F^2 - \|\mathbf{W}^{(r)} - \mathbf{U}\|_F^2 = \|\boldsymbol{\Delta}\mathbf{W}^{(r)}\|_F^2 \tag{11}$$
$$+ \underbrace{2\langle \boldsymbol{\Delta}\mathbf{W}^{(r)}, \mathbf{W}^{(r)} - \mathbf{U}_1\rangle}_{=:\mathcal{A}_1} + \underbrace{2\langle -\boldsymbol{\Delta}\mathbf{W}^{(r)}, \mathbf{U}_2 + \mathbf{U}_3\rangle}_{=:\mathcal{A}_2}.$$

To bound $\mathcal{A}_1$, we have that as $\mathcal{L}_i(\cdot)$ is $1/2$-smooth (cf. Lemma 13) and convex,

$$\langle \nabla \mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)}), \mathbf{W}^{(r)} - \mathbf{U}_1\rangle$$
$$\geq \mathcal{L}_i(\mathbf{W}^{(r)}) - \mathcal{L}_i(\mathbf{U}_1) - \frac{1}{4}\|\mathbf{W}^{(r)} - \mathbf{D}_{i,\tau-1}^{(r)}\|_F^2. \tag{12}$$

Therefore,

$$\mathcal{A}_1 = -\frac{2\bar{\eta}}{n}\sum_{i=1}^m \sum_{\tau=1}^{T_i} n_i \langle \nabla \mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)}), \mathbf{W}^{(r)} - \mathbf{U}_1\rangle$$
$$\leq \frac{2\bar{\eta}}{n}\sum_{i=1}^m n_i T_i \left(\mathcal{L}_i(\mathbf{U}_1) - \mathcal{L}_i(\mathbf{W}^{(r)})\right) + \mathcal{E}^{(r)}$$

where

$$\mathcal{E}^{(r)} := \frac{\bar{\eta}}{2n}\sum_{i=1}^{m}\sum_{\tau=1}^{T_i} n_i \|\mathbf{W}^{(r)} - \mathbf{D}_{i,\tau-1}^{(r)}\|_F^2. \tag{13}$$

Applying Lemma 6-(i) with $\alpha = \log((K-1)\gamma^2\bar{\eta}RT_{\text{avg}})/\gamma$, it holds that $\mathcal{L}_i(\mathbf{U}_1) \leq 1/(\gamma^2\bar{\eta}RT_{\text{avg}})$ for all i. Therefore,

$$\frac{2\bar{\eta}}{n}\sum_{i=1}^{m} n_i T_i \mathcal{L}_i(\mathbf{U}_1) \leq \frac{2}{\gamma^2 R}. \tag{14}$$

On the other hand, with $\mathcal{L}_i(\mathbf{W}^{(r)}) \geq 0$, we have

$$\frac{2\bar{\eta}}{n}\sum_{i=1}^{m} n_i T_i \mathcal{L}_i(\mathbf{W}^{(r)}) \geq 2\bar{\eta}T_{\min}\mathcal{L}(\mathbf{W}^{(r)}). \tag{15}$$

This yields $\mathcal{A}_1 \leq \frac{2}{\gamma^2 R} - 2\bar{\eta}T_{\min}\mathcal{L}(\mathbf{W}^{(r)}) + \mathcal{E}^{(r)}$.

To bound $\mathcal{A}_2$, we can apply Lemma 6-(ii) and observe that $\mathcal{A}_2$ can be evaluated and bounded as

$$\begin{aligned}
&\frac{2\bar{\eta}}{n}\left(\frac{\bar{\eta}T_{\text{avg}}}{\gamma} + \frac{\eta_l^2 T_{\max}^2}{\gamma}\right)\sum_{i=1}^{m}\sum_{\tau=1}^{T_i} n_i \langle \nabla\mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)}), \mathbf{W}_\star\rangle \\
&\leq \frac{2\bar{\eta}}{n}\left(\bar{\eta}T_{\text{avg}} + \eta_l^2 T_{\max}^2\right)\sum_{i=1}^{m}\sum_{\tau=1}^{T_i}\sum_{j=1}^{n_i}\left(\widehat{y}_{i,j|\mathbf{D}_{i,\tau-1}^{(r)}} - 1\right).
\end{aligned} \tag{16}$$

Now, using Lemma 6-(iii), we notice that $\|\boldsymbol{\Delta}\mathbf{W}^{(r)}\|_F^2$ can be bounded as

$$\begin{aligned}
\|\boldsymbol{\Delta}\mathbf{W}^{(r)}\|_F^2 &\leq \left(\frac{\bar{\eta}^2}{n}\sum_{i=1}^{m}\sum_{\tau=1}^{T_i} n_i\right)\sum_{i=1}^{m}\sum_{\tau=1}^{T_i}\frac{n_i}{n}\|\nabla\mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)})\|_F^2 \\
&\leq \frac{2\bar{\eta}^2 T_{\text{avg}}}{n}\sum_{i=1}^{m}\sum_{\tau=1}^{T_i}\sum_{j=1}^{n_i}\left(1 - \widehat{y}_{i,j|\mathbf{D}_{i,\tau-1}^{(r)}}\right).
\end{aligned} \tag{17}$$

Lastly, using Lemma 6, the residual term $\mathcal{E}^{(r)}$ in (13) can be bounded as:

Lemma 7. *Under Assumption 2, for any $r \geq 0$, it holds*

$$\mathcal{E}^{(r)} \leq \frac{2\bar{\eta}\eta_l^2 T_{\max}^2}{n}\sum_{i=1}^{m}\sum_{t=1}^{T_i}\sum_{j=1}^{n_i}\left(1 - \widehat{y}_{i,j|\mathbf{D}_{i,t-1}^{(r)}}\right). \tag{18}$$

Combining (18), (16), and (17), we obtain $\mathcal{A}_2 + \mathcal{E}^{(r)} + \|\boldsymbol{\Delta}\mathbf{W}^{(r)}\|_F^2 \leq 0$. As such,

$$\begin{aligned}
&\|\mathbf{W}^{(r+1)} - \mathbf{U}\|_F^2 - \|\mathbf{W}^{(r)} - \mathbf{U}\|_F^2 \\
&\leq \frac{2}{\gamma^2 R} - 2\bar{\eta}T_{\min}\mathcal{L}(\mathbf{W}^{(r)}).
\end{aligned} \tag{19}$$

Re-arranging terms leads to

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathcal{L}(\mathbf{W}^{(r)}) \leq \frac{1}{2\bar{\eta}T_{\min}R}\left(\|\mathbf{W}^{(0)} - \mathbf{U}\|_F^2 + \frac{2}{\gamma^2}\right).$$

Using that

$$\|\mathbf{U}\|_F^2 \leq \frac{\log^2(\gamma^2\bar{\eta}R(K-1)T_{\text{avg}}) + \bar{\eta}^2 T_{\text{avg}}^2 + \eta_l^4 T_{\max}^4}{\gamma^2/2}, \tag{20}$$

we obtain (2) as a result. Iterating from (19) yields

$$\|\mathbf{W}^{(R)} - \mathbf{U}\|_F^2 \leq \|\mathbf{W}^{(0)} - \mathbf{U}\|_F^2 + 2/\gamma^2. \tag{21}$$

Taking square root on both sides and applying triangle inequality yields

$$\|\mathbf{W}^{(R)}\|_F \leq \|\mathbf{W}^{(0)}\|_F + \|\mathbf{U}\|_F + \sqrt{2}/\gamma. \tag{22}$$

Applying (20) yields (3) and conclude our proof.

Remark 8. Compared to the proof of [16, Theorem 4.1], our proof relies on a different choice of the reference point $\mathbf{U}$, together with a recursion for the round-by-round progress of `FedAvg`. These non-trivial modifications have enabled us to obtain a tighter bound than [16, Theorem 4.1] that also handles the heterogeneous devices case.

B. Proof of Theorem 4

We preface the proof by defining the gradient potentials:

$$\mathcal{G}(\mathbf{W}) := \sum_{i=1}^{m}\frac{n_i}{n}\mathcal{G}_i(\mathbf{W}),\ \ \mathcal{G}_i(\mathbf{W}) := \sum_{j=1}^{n_i}\frac{1 - \widehat{y}_{i,j|\mathbf{W}}}{n_i}.$$

Note that $\mathcal{G}_i(\mathbf{W}) \leq \mathcal{L}_i(\mathbf{W})$ for all i. We shall prove (4) through showing by induction that for any $r \geq s$, it holds

$$\mathcal{L}(\mathbf{W}^{(r)}) \leq \min\left\{\frac{1}{2n}, \frac{1}{4\bar{\eta}T_{\text{avg}}}, \frac{\gamma T_{\text{avg}}}{40n\eta_l T_{\max}^2\sqrt{K}}\right\},$$

$$\mathcal{G}(\mathbf{W}^{(r)}) \leq 1/(2n), \tag{23a}$$

$$\mathcal{L}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r)}). \tag{23b}$$

For the base case with $r = s$, we note that (23a) holds by the assumption of Theorem 4 and $\mathcal{G}(\mathbf{W}^{(r)}) \leq \mathcal{L}(\mathbf{W}^{(r)})$. To establish (23b), the key step is to analyze the one-step progress from $\mathbf{W}^{(r)}$ to $\mathbf{W}^{(r+1)}$ and to bound $\mathcal{L}(\mathbf{W}^{(r+1)})$ via a descent lemma type argument on $\log\mathcal{L}(\mathbf{W}^{(r)})$. To this end, the server-side update can be written as:

$$\mathbf{W}^{(r+1)} - \mathbf{W}^{(r)} := -\bar{\eta}T_{\text{avg}}(\nabla\mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}),$$

$$\mathbf{E}^{(r)} = \textstyle\sum_{i=1}^{m}\sum_{\tau=1}^{T_i}\frac{n_i}{nT_{\text{avg}}}\left(\nabla\mathcal{L}_i(\mathbf{D}_{i,\tau-1}^{(r)}) - \nabla\mathcal{L}_i(\mathbf{W}^{(r)})\right),$$

$$\tilde{\mathbf{E}}^{(r)} = \textstyle\sum_{i=1}^{m}\sum_{\tau=1}^{T_i}\frac{n_i}{nT_{\text{avg}}}\left(\nabla\mathcal{L}_i(\mathbf{W}^{(r)}) - \nabla\mathcal{L}(\mathbf{W}^{(r)})\right).$$

In Appendix III, we show that under the assumptions of Theorem 4 and $\mathcal{G}(\mathbf{W}^{(r)}) \leq 1/(2n)$, it holds that

$$\max\{\|\mathbf{E}^{(r)}\|_F, \|\tilde{\mathbf{E}}^{(r)}\|_F\} \leq \frac{1}{10}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F. \tag{24}$$

The above bound yields

$$\|\nabla\mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F \geq 4\|\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F. \tag{25}$$

Finally, using the 2-smoothness of $\log \mathcal{L}(\mathbf{W})$ (cf. Lemma 9),

$$\begin{aligned}
&\log \mathcal{L}(\mathbf{W}^{(r+1)}) - \log \mathcal{L}(\mathbf{W}^{(r)}) \\
&\leq \left\langle \frac{\nabla \mathcal{L}(\mathbf{W}^{(r)})}{\mathcal{L}(\mathbf{W}^{(r)})}, \mathbf{W}^{(r+1)} - \mathbf{W}^{(r)} \right\rangle \\
&\quad + \|\mathbf{W}^{(r+1)} - \mathbf{W}^{(r)}\|_F^2 \\
&= -\frac{\bar{\eta} T_{\text{avg}}}{\mathcal{L}(\mathbf{W}^{(r)})} \left\| \nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)} \right\|_F^2 \\
&\quad + \frac{\bar{\eta} T_{\text{avg}}}{\mathcal{L}(\mathbf{W}^{(r)})} \left\langle \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}, \nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)} \right\rangle \\
&\quad + \bar{\eta}^2 T_{\text{avg}}^2 \|\nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2. \qquad (26)
\end{aligned}$$

Subsequently, applying Cauchy-Schwarz and (25) yields

$$\begin{aligned}
&\log \mathcal{L}(\mathbf{W}^{(r+1)}) - \log \mathcal{L}(\mathbf{W}^{(r)}) \\
&\leq -\frac{\bar{\eta} T_{\text{avg}}}{\mathcal{L}(\mathbf{W}^{(r)})} \left\| \nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)} \right\|_F^2 \\
&\quad + \frac{\bar{\eta} T_{\text{avg}}}{4\mathcal{L}(\mathbf{W}^{(r)})} \left\| \nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)} \right\|_F^2 \\
&\quad + \bar{\eta}^2 T_{\text{avg}}^2 \|\nabla \mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2. \qquad (27)
\end{aligned}$$

As $\mathcal{L}(\mathbf{W}^{(s)}) \leq 1/(4\bar{\eta} T_{\text{avg}})$, the right hand side of the above is upper bounded by 0, implying $\log \mathcal{L}(\mathbf{W}^{(r+1)}) \leq \log \mathcal{L}(\mathbf{W}^{(r)})$.

For the induction step, assuming that (23) holds up to certain $r \geq s$. Then, we note that (23a) for $r+1$ can be verified by $\mathcal{L}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r)})$ and $\mathcal{G}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r+1)})$. Finally, the same argument for establishing (23b) at the base case can be applied to show that $\mathcal{L}(\mathbf{W}^{(r+2)}) \leq \mathcal{L}(\mathbf{W}^{(r+1)})$. This concludes the proof of Theorem 4.

C. Proof of Corollary 5

We shall prove (5) through showing that for any $r \geq s$,

$$\frac{1}{\mathcal{L}(\mathbf{W}^{(r+1)})} - \frac{1}{\mathcal{L}(\mathbf{W}^{(r)})} \geq \frac{\bar{\eta} T_{\text{avg}} \gamma^2}{16}. \qquad (28)$$

We notice from the proof of Theorem 4 that (23) holds for any $r \geq s$. As we shall show in Appendix IV, the above conditions guarantee that

$$\mathcal{L}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r)}) - \frac{\bar{\eta} T_{\text{avg}} \gamma^2}{16} \left(\mathcal{L}(\mathbf{W}^{(r)}) \right)^2. \qquad (29)$$

Dividing both sides by $\mathcal{L}(\mathbf{W}^{(r)}) \mathcal{L}(\mathbf{W}^{(r+1)})$, we obtain

$$\begin{aligned}
\frac{1}{\mathcal{L}(\mathbf{W}^{(r+1)})} - \frac{1}{\mathcal{L}(\mathbf{W}^{(r)})} &= \frac{\mathcal{L}(\mathbf{W}^{(r)}) - \mathcal{L}(\mathbf{W}^{(r+1)})}{\mathcal{L}(\mathbf{W}^{(r)}) \mathcal{L}(\mathbf{W}^{(r+1)})} \\
&\geq \frac{\bar{\eta} T_{\text{avg}} \gamma^2 \mathcal{L}(\mathbf{W}^{(r)})}{16 \mathcal{L}(\mathbf{W}^{(r+1)})}. \qquad (30)
\end{aligned}$$

Since $\mathcal{L}(\mathbf{W}^{(r+1)}) \leq \mathcal{L}(\mathbf{W}^{(r)})$, we obtain (28). It remains to apply a telescoping argument which shows

$$\frac{1}{\mathcal{L}(\mathbf{W}^{(r)})} \geq \frac{1}{\mathcal{L}(\mathbf{W}^{(s)})} + \frac{\bar{\eta} T_{\text{avg}} \gamma^2}{16} (r - s). \qquad (31)$$

Using $\mathcal{L}(\mathbf{W}^{(s)}) > 0$ proves (5) and concludes our proof.

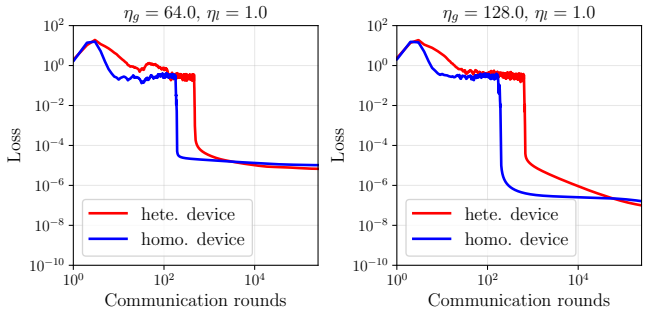


Fig. 1. Convergence of `FedAvg` with homogeneous ("homo. device": $T_i = 27$ for all devices) and heterogeneous ("hete. device": $T_i = 4$ for device 1-4, $T_i = 50$ for device 5-8) devices. Both configurations share the same average $T_{\text{avg}} = 27$ with global/local stepsizes: (Left) $\eta_g = 64, \eta_l = 1$, (Right) $\eta_g = 128, \eta_l = 1$.

IV. Numerical Experiments

This section reports the numerical experiments on a multi-class classification task using the large stepsize `FedAvg` algorithm. We consider a synthetic dataset with an input dimension $d = 12$, $K = 5$ classes, $n = 5,000$ samples. Each sample in the global dataset $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_m$ is constructed via a rejection sampling mechanism such that the linear classifier defined by the weight matrix $\mathbf{W}_\star \in \mathbb{R}^{K \times d}$ satisfies Assumptions 1 and 2 with $\gamma = 0.001$. Note that under this setup, as $\mathcal{L}(\mathbf{W})$ is $1/2$-smooth, traditional analysis of `FedAvg` requires a small stepsize as $\bar{\eta} = \eta_l \eta_g \leq 2$. We use a label-imbalanced data split: each client receives samples from at most two distinct classes.

A. Effect of Heterogeneous Local Epochs

We first examine the effect of heterogeneous devices on the convergence behavior. To this end, the dataset is partitioned among $m = 8$ clients; each client receives exactly 625 samples. In this experiment, we fix the local stepsize at $\eta_l = 1$, while the global stepsizes $\eta_g \in \{64, 128\}$ are examined. Two configurations for the number of local update steps, T_i, are considered: (i) *heterogeneous* local steps, where clients 1-4 perform $T_i = 4$ local updates per communication round, while clients 5-8 perform $T_i = 50$ updates, yielding $T_{\text{avg}} = 27$; (ii) *homogeneous* local steps, where all clients perform $T_i = 27$ local steps per round.

The `FedAvg` algorithm is executed for $R = 250,000$ communication rounds. Fig. 1 shows the trajectory of cross-entropy loss on the full dataset for each communication round. As observed, in both global stepsize configurations, the heterogeneous device case exhibits a longer transient with larger fluctuations than the homogeneous device case before entering a stable monotonic regime. This behavior is consistent with Theorem 3 which predicts a slower convergence for heterogeneous devices. However, we observe that after the transient phase as $R \gg 1$, the convergence behavior of both heterogeneous and homogeneous cases become indistinguishable. This corroborates with Corollary 5 which shows that the asymptotic rate is dominated by

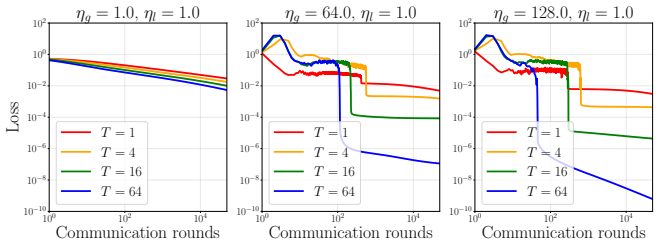


Fig. 2. Effects of the number of local steps T. All devices perform the same number of local epochs $T \in \{1, 4, 16, 64\}$, with global stepsizes $\eta_g \in \{1, 64, 128\}$ and local stepsize $\eta_l = 1$.

$\mathcal{O}(1/(T_{\text{avg}}R))$, independent of the device heterogeneity[1].

B. Effect of Large Local Steps under Label Imbalance

We next examine the effect of increasing the common number of local steps $T = T_i$ for each i. All clients perform the same number of local epochs $T \in \{1, 4, 16, 64\}$, with the stepsizes as $\eta_g \in \{1, 64, 128\}$ and $\eta_l = 1$.

The `FedAvg` algorithm is executed for $R = 50,000$ communication rounds. Fig. 2 plots the cross-entropy loss value against communication rounds with varying T. We first observe that for sufficiently large R, the convergence rates of all configurations are close to $\mathcal{O}(1/(TR))$ as predicted by Corollary 5, where increasing T leads to a faster convergence. Furthermore, we note that the large stepsizes settings $\eta_g \in \{64, 128\}$ significantly accelerate the convergence compared to the baseline case with $\eta_g = 1$, where the convergence is significantly slower.

V. Conclusion

This paper has extended the analysis of large stepsize `FedAvg` to the case of heterogeneous devices under a linear separability condition. We showed that the effect of device heterogeneity vanishes asymptotically, where the convergence rate is dominated by the average number of local steps across devices and validated the findings by numerical experiments. For future work, it will be interesting to explore non-separable data and regularized losses.

References

[1] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 2003.

[2] A. Nedić and J. Liu, "Distributed optimization for control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 77–103, 2018.

[3] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, pp. 1273–1282, PMLR, 2017.

[5] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Foundations and trends in machine learning*, vol. 14, no. 1-2, pp. 1–210, 2021.

[6] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2019.

[1]Note that the distribution of $\{T_i\}_{i=1}^m$ may violates the condition required by Theorem 4, yet we observe that the predicted behavior continues to hold.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[8] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "A novel framework for the analysis and design of heterogeneous federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5234–5249, 2021.

[9] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*, pp. 5132–5143, PMLR, 2020.

[10] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.

[11] X. Yuan and P. Li, "On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10752–10765, 2022.

[12] J. Wu, P. L. Bartlett, M. Telgarsky, and B. Yu, "Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency," in *The Thirty Seventh Annual Conference on Learning Theory*, pp. 5019–5073, PMLR, 2024.

[13] J. Wu, P. Marion, and P. Bartlett, "Large stepsizes accelerate gradient descent for regularized logistic regression," *arXiv preprint arXiv:2506.02336*, 2025.

[14] A. Tyurin, "From logistic regression to the perceptron algorithm: Exploring gradient descent with large step sizes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 20938–20946, 2025.

[15] S. Y. Meng, A. Orvieto, D. Y. Cao, and C. De Sa, "Gradient descent on logistic regression with non-separable data and large step sizes," *arXiv preprint arXiv:2406.05033*, 2024.

[16] M. Crawshaw, B. Woodworth, and M. Liu, "Constant stepsize local gd for logistic regression: Acceleration by instability," in *International Conference on Machine Learning*, pp. 11465–11492, PMLR, 2025.

[17] S. Y. Meng, B. Goujaud, A. Orvieto, and C. De Sa, "Gradient descent on logistic regression: Do large step-sizes work with data on the sphere?," *arXiv preprint arXiv:2507.11228*, 2025.

Appendix I
Proof of Lemma 6

To prove (6), notice that for any $k \neq y_{i,j}$, Assumption 2 guarantees that $(\mathbf{w}_{\star,y_{i,j}} - \mathbf{w}_{\star,k})^\top \mathbf{x}_{i,j} \geq \gamma$. Therefore,

$$\begin{aligned}
&\ell_{i,j}(\alpha \mathbf{W}_\star) \qquad (32)\\
&= \log\left(1 + \textstyle\sum_{k \neq y_{i,j}} \exp(\alpha \mathbf{w}_{\star,k}^\top \mathbf{x}_{i,j} - \alpha \mathbf{w}_{\star,y_{i,j}}^\top \mathbf{x}_{i,j})\right)\\
&\leq \log\left(1 + (K-1)\exp\left(\alpha \max_{k \neq y_{i,j}} (\mathbf{w}_{\star,k} - \mathbf{w}_{\star,y_{i,j}})^\top \mathbf{x}_{i,j}\right)\right)\\
&\leq \log\left(1 + (K-1)\exp\left(-\alpha\gamma\right)\right) \leq (K-1)\exp\left(-\alpha\gamma\right).
\end{aligned}$$

To prove (7), we first notice that the inner product is

$$\langle \nabla \ell_{i,j}(\mathbf{W}), \mathbf{W}_\star \rangle = \sum_{k=1}^K \left(\widehat{y}_{i,j|\mathbf{W}}^{(k)} - \mathbb{1}(y_{i,j} = k)\right) \mathbf{x}_{i,j}^\top \mathbf{w}_{\star,k},$$

where $\mathbb{1}(\cdot)$ is the indicator function with $\mathbb{1}(y_{i,j} = k) = 1$ for $k = y_{i,j}$ and 0 otherwise. Under Assumption 2, we have

$$\begin{aligned}
&\textstyle\sum_{k=1}^K \left(\widehat{y}_{i,j|\mathbf{W}}^{(k)} - \mathbb{1}(y_{i,j} = k)\right) \mathbf{x}_{i,j}^\top \mathbf{w}_{\star,k}\\
&= (\widehat{y}_{i,j|\mathbf{W}} - 1)\mathbf{x}_{i,j}^\top \mathbf{w}_{\star,y_{i,j}} + \textstyle\sum_{k \neq y_{i,j}} \widehat{y}_{i,j|\mathbf{W}}^{(k)} \mathbf{x}_{i,j}^\top \mathbf{w}_{\star,k}\\
&\leq (\widehat{y}_{i,j|\mathbf{W}} - 1)\mathbf{x}_{i,j}^\top \mathbf{w}_{\star,y_{i,j}} + \textstyle\sum_{k \neq y_{i,j}} \widehat{y}_{i,j|\mathbf{W}}^{(k)} \left(\mathbf{x}_{i,j}^\top \mathbf{w}_{\star,y_{i,j}} - \gamma\right)\\
&= \left(\textstyle\sum_{k=1}^K \widehat{y}_{i,j|\mathbf{W}}^{(k)} - 1\right) \mathbf{x}_{i,j}^\top \mathbf{w}_{\star,y_{i,j}} - \gamma \textstyle\sum_{k \neq y_{i,j}} \widehat{y}_{i,j|\mathbf{W}}^{(k)}\\
&= -\gamma(1 - \widehat{y}_{i,j|\mathbf{W}}). \qquad (33)
\end{aligned}$$

To prove (8), we observe that

$$\|\nabla \ell_{i,j}(\mathbf{W})\|_F^2 = \sum_{k=1}^K \left\| \left(\widehat{y}^{(k)}_{i,j|\mathbf{W}} - \mathbb{1}(y_{i,j}=k) \right) \mathbf{x}_{i,j} \right\|^2$$
$$= \|\mathbf{x}_{i,j}\|^2 \Big((1-\widehat{y}_{i,j|\mathbf{W}})^2 + \sum_{k\neq y_{i,j}} (\widehat{y}^{(k)}_{i,j|\mathbf{W}})^2 \Big). \quad (34)$$

Since $(1-\widehat{y}_{i,j|\mathbf{W}})^2 \leq 1-\widehat{y}_{i,j|\mathbf{W}}$ and

$$\sum_{k\neq y_{i,j}} (\widehat{y}^{(k)}_{i,j|\mathbf{W}})^2 \leq \sum_{k\neq y_{i,j}} \widehat{y}^{(k)}_{i,j|\mathbf{W}} = 1-\widehat{y}_{i,j|\mathbf{W}},$$

we have

$$\|\nabla \ell_{i,j}(\mathbf{W})\|_F^2 \leq 2(1-\widehat{y}_{i,j|\mathbf{W}}). \quad (35)$$

Appendix II
Proof of Lemma 7

We concentrate on the following term for each i,

$$\sum_{\tau=1}^{T_i} n_i \|\mathbf{W}^{(r)} - \mathbf{D}^{(r)}_{i,\tau-1}\|_F^2. \quad (36)$$

If $T_i = 1$, then the above term is zero. For $T_i \geq 2$, from local update rule $\mathbf{D}^{(r)}_{i,\tau} = \mathbf{D}^{(r)}_{i,\tau-1} - \eta_l \nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1})$, we have

$$\|\mathbf{D}^{(r)}_{i,\tau} - \mathbf{W}^{(r)}\|_F^2 = \|\mathbf{D}^{(r)}_{i,\tau-1} - \eta_l \nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1}) - \mathbf{W}^{(r)}\|_F^2$$
$$\leq \frac{T_i}{T_i-1} \|\mathbf{D}^{(r)}_{i,\tau-1} - \mathbf{W}^{(r)}\|_F^2 + \eta_l^2 T_i \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1})\|_F^2$$
$$\leq \eta_l^2 T_i \textstyle\sum_{t=1}^{\tau-1} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2 \left(\frac{T_i}{T_i-1}\right)^{\tau-t}. \quad (37)$$

We now proceed to bound

$$\mathcal{E}^{(r)} = \frac{\bar{\eta}}{n} \sum_{i=1}^m \sum_{\tau=2}^{T_i} n_i \|\mathbf{D}^{(r)}_{i,\tau-1} - \mathbf{W}^{(r)}\|_F^2$$
$$\leq \frac{\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m n_i T_i \sum_{\tau=2}^{T_i} \sum_{t=1}^{\tau-1} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2 \left(1+\frac{1}{T_i-1}\right)^{\tau-t}$$
$$= \frac{\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m n_i T_i \sum_{t=1}^{T_i-1} \sum_{\tau=1}^{T_i-t} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2 \left(1+\frac{1}{T_i-1}\right)^{\tau}.$$

The above can be further bounded by:

$$\leq \frac{\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m n_i T_i \sum_{t=1}^{T_i-1} \sum_{\tau=1}^{T_i-1} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2 \left(1+\frac{1}{T_i-1}\right)^{\tau}$$
$$\leq \frac{\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m n_i T_i \sum_{t=1}^{T_i-1} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2 \cdot (e-1) T_i$$
$$\leq \frac{2\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m n_i T_i^2 \sum_{t=1}^{T_i-1} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F^2$$
$$\leq \frac{4\bar{\eta}\eta_l^2}{n} \sum_{i=1}^m T_i^2 \sum_{t=1}^{T_i-1} \sum_{j=1}^{n_i} \left(1-\widehat{y}_{i,j|\mathbf{D}^{(r)}_{i,t-1}}\right). \quad (38)$$

Appendix III
Missing Proofs of Theorem 4

We preface the proof by a few additional lemmas, whose proofs can be found in Appendix VI of `https://www1.se.cuhk.edu.hk/~htwai/pdf/cdc26-largestep.pdf`.

Lemma 9. *Under Assumption 1. For every $i \in [m]$, the function $\log \mathcal{L}_i(\mathbf{W})$ is 2-smooth.*

Lemma 10. *Under Assumptions 1 and 2. Suppose that there exists $s \geq 0$ with $\mathcal{L}_i(\mathbf{D}^{(r)}_{i,s}) \leq 1/\eta_l$, then for every $\eta_l > 0$,*

$$\mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau+1}) \leq \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau}), \ \forall\ \tau \geq s, \ \forall\ i \in [m]. \quad (39)$$

Lemma 11. *Under Assumption 1, for any $\mathbf{W}$, we have*

$$\|\nabla \mathcal{L}(\mathbf{W})\|_F \leq \sqrt{K}\mathcal{L}(\mathbf{W}), \quad \|\nabla^2 \mathcal{L}(\mathbf{W})\|_{\text{op}} \leq 2\mathcal{L}(\mathbf{W}). \quad (40)$$

Lemma 12. *Under Assumptions 1 & 2, for every $\eta_g, \eta_l > 0$,*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathcal{G}(\mathbf{W}^{(r)}) \leq \frac{\sqrt{2}+\|\mathbf{W}^{(0)}\|_F}{\bar{\eta}\gamma^2 R}$$
$$+ \frac{\log(\gamma^2\bar{\eta}R(K-1)T_{\text{avg}}) + \bar{\eta}T_{\text{avg}} + \eta_l^2 T_{\max}^2}{\bar{\eta}\gamma^2 R/2}. \quad (41)$$

Furthermore, if $\mathcal{G}(\mathbf{W}) \leq 1/(2n)$, then

$$\|\nabla \mathcal{L}(\mathbf{W})\|_F \geq (\gamma/2)\mathcal{L}(\mathbf{W}).$$

Proof of (25). Our main focus is to prove (25). Let $C = \{(1-t)\mathbf{W}^{(r)} + t\mathbf{D}^{(r)}_{i,\tau-1} : t \in (0,1)\}$, notice that

$$\|\mathbf{E}^{(r)}\|_F \leq \sum_{i=1}^m \sum_{\tau=1}^{T_i} \frac{n_i}{nT_{\text{avg}}} \|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1}) - \nabla \mathcal{L}_i(\mathbf{W}^{(r)})\|_F$$
$$\leq \sum_{i=1}^m \sum_{\tau=1}^{T_i} \frac{n_i}{nT_{\text{avg}}} \max_{\mathbf{W}\in C} \|\nabla^2 \mathcal{L}_i(\mathbf{W})\|_{\text{op}} \left\|\mathbf{D}^{(r)}_{i,\tau-1} - \mathbf{W}^{(r)}\right\|_F$$
$$\leq \sum_{i=1}^m \sum_{\tau=1}^{T_i} \frac{2n_i}{nT_{\text{avg}}} \max_{\mathbf{W}\in C} \mathcal{L}_i(\mathbf{W}) \left\|\mathbf{D}^{(r)}_{i,\tau-1} - \mathbf{W}^{(r)}\right\|_F,$$

where the second last inequality is due to Lemma 11. Using Lemma 10, $\mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1}) \leq \mathcal{L}_i(\mathbf{W}^{(r)})$ for all $\tau \geq 1$, we have

$$\|\mathbf{E}^{(r)}\|_F \leq \frac{2}{nT_{\text{avg}}} \sum_{i=1}^m \sum_{\tau=1}^{T_i} n_i \mathcal{L}_i(\mathbf{W}^{(r)}) \left\|\mathbf{D}^{(r)}_{i,\tau-1} - \mathbf{W}^{(r)}\right\|_F$$
$$\leq \frac{2}{nT_{\text{avg}}} \sum_{i=1}^m \sum_{\tau=1}^{T_i-1} n_i \mathcal{L}_i(\mathbf{W}^{(r)}) \left\|\sum_{t=1}^{\tau} \eta_l \nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\right\|_F$$
$$\leq \frac{2\eta_l}{nT_{\text{avg}}} \sum_{i=1}^m \sum_{\tau=1}^{T_i-1} \sum_{t=1}^{\tau} n_i \mathcal{L}_i(\mathbf{W}^{(r)}) \left\|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\right\|_F. \quad (42)$$

Lemma 11 shows $\|\nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})\|_F \leq \sqrt{K}\mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1})$, and hence we can bound $\|\mathbf{E}^{(r)}\|_F$ by

$$\tfrac{2\eta_l}{nT_{\text{avg}}} \textstyle\sum_{i=1}^m \sum_{\tau=1}^{T_i-1} \sum_{t=1}^{\tau} n_i \mathcal{L}_i(\mathbf{W}^{(r)})(\sqrt{K}\mathcal{L}_i(\mathbf{D}^{(r)}_{i,t-1}))$$
$$\leq \tfrac{2\eta_l\sqrt{K}}{nT_{\text{avg}}} \textstyle\sum_{i=1}^m (n_i T_i \mathcal{L}_i(\mathbf{W}^{(r)}))^2$$
$$\leq \tfrac{2\eta_l\sqrt{K}T_{\max}^2}{nT_{\text{avg}}} \textstyle\sum_{i=1}^m (n_i \mathcal{L}_i(\mathbf{W}^{(r)}))^2$$
$$\leq \frac{2n\eta_l\sqrt{K}T_{\max}^2}{T_{\text{avg}}} \left(\mathcal{L}(\mathbf{W}^{(r)})\right)^2. \quad (43)$$

Using $\mathcal{L}(\mathbf{W}^{(r)}) \leq \gamma T_{\text{avg}}/(40n\eta_l T_{\max}^2\sqrt{K})$ and Lemma 12, we have that $\|\mathbf{E}^{(r)}\|_F \leq \frac{\gamma}{20}\mathcal{L}(\mathbf{W}^{(r)}) \leq \frac{1}{10}\|\nabla \mathcal{L}(\mathbf{W}^{(r)})\|_F$.

Next, we bound $\|\tilde{\mathbf{E}}^{(r)}\|_F$. Observe that

$$\begin{aligned}\tilde{\mathbf{E}}^{(r)} &= \frac{1}{nT_{\text{avg}}}\sum_{i=1}^m n_i T_i \cdot \left(\nabla\mathcal{L}_i(\mathbf{W}^{(r)}) - \nabla\mathcal{L}(\mathbf{W}^{(r)})\right)\\ &= \frac{1}{nT_{\text{avg}}}\sum_{i=1}^m n_i T_i \nabla\mathcal{L}_i(\mathbf{W}^{(r)}) - \nabla\mathcal{L}(\mathbf{W}^{(r)})\\ &= \frac{1}{nT_{\text{avg}}}\sum_{i=1}^m n_i T_i \nabla\mathcal{L}_i(\mathbf{W}^{(r)}) - \frac{1}{n}\sum_{i=1}^m n_i \nabla\mathcal{L}_i(\mathbf{W}^{(r)})\\ &= \frac{1}{n}\sum_{i=1}^m n_i\left(\frac{T_i}{T_{\text{avg}}} - 1\right)\nabla\mathcal{L}_i(\mathbf{W}^{(r)}). \end{aligned} \tag{44}$$

By Lemma 11, $\|\nabla\mathcal{L}_i(\mathbf{W}^{(r)})\|_F \le \sqrt{K}\mathcal{L}_i(\mathbf{W}^{(r)})$. Further using the last inequality in Lemma 12, we have

$$\begin{aligned}\|\tilde{\mathbf{E}}^{(r)}\|_F &\le \tfrac{1}{n}\textstyle\sum_{i=1}^m n_i \left|\frac{T_i}{T_{\text{avg}}} - 1\right| \cdot \|\nabla\mathcal{L}_i(\mathbf{W}^{(r)})\|_F\\ &\le \tfrac{\sqrt{K}}{n}\textstyle\sum_{i=1}^m n_i \left|\frac{T_i}{T_{\text{avg}}} - 1\right| \cdot \mathcal{L}_i(\mathbf{W}^{(r)})\\ &\le \max_i\left|\tfrac{T_i}{T_{\text{avg}}} - 1\right| \sqrt{K}\tfrac{2}{\gamma}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F. \end{aligned} \tag{45}$$

Under the assumption that $\max_i |T_i/T_{\text{avg}} - 1| \le \gamma/(20\sqrt{K})$, we have that $\|\tilde{\mathbf{E}}^{(r)}\|_F \le \frac{1}{10}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F$.

A. Proof of Lemma 9

To verify that $\log\mathcal{L}_i(\mathbf{W})$ is 2-smooth, we observe that:

$$\begin{aligned}&\nabla^2\log\mathcal{L}_i(\mathbf{W})\\ &= \frac{1}{\mathcal{L}_i(\mathbf{W})^2}\left(\nabla^2\mathcal{L}_i(\mathbf{W})\cdot\mathcal{L}_i(\mathbf{W}) - (\nabla\mathcal{L}_i(\mathbf{W}))^{\otimes 2}\right)\\ &\preceq \frac{1}{\mathcal{L}_i(\mathbf{W})}\nabla^2\mathcal{L}_i(\mathbf{W}) \preceq \frac{1}{\mathcal{L}_i(\mathbf{W})}\mathcal{L}_i(\mathbf{W})\cdot 2\mathbf{I} = 2\mathbf{I}, \end{aligned} \tag{46}$$

where the second last inequality comes from Lemma 11.

B. Proof of Lemma 10

Using the 2-smoothness of $\log\mathcal{L}_i(\cdot)$, we get

$$\begin{aligned}&\log\mathcal{L}_i(\mathbf{D}_{i,\tau+1}^{(r)})\\ &\le \log\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)}) + \left\langle\tfrac{\nabla\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)})}{\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)})}, \mathbf{D}_{i,\tau+1}^{(r)} - \mathbf{D}_{i,\tau}^{(r)}\right\rangle\\ &\quad + \|\mathbf{D}_{i,\tau+1}^{(r)} - \mathbf{D}_{i,\tau}^{(r)}\|_F^2\\ &= \log\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)}) - \eta_l\left(\tfrac{1}{\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)})} - \eta_l\right)\|\nabla\mathcal{L}_i(\mathbf{D}_{i,\tau}^{(r)})\|_F^2. \end{aligned} \tag{47}$$

The above implies that if $\mathcal{L}(\mathbf{D}_{i,\tau}^{(r)}) \le \frac{1}{\eta_l}$, then

$$\mathcal{L}(\mathbf{D}_{i,\tau+1}^{(r)}) \le \mathcal{L}(\mathbf{D}_{i,\tau}^{(r)}). \tag{48}$$

Appendix IV
Missing Proof of Corollary 5

We shall prove (29) in this section. By letting $\mathbf{V} = \lambda\mathbf{W}^{(r)} + (1-\lambda)\mathbf{W}^{(r+1)}, \lambda\in(0,1)$, we have

$$\begin{aligned}&\mathcal{L}(\mathbf{W}^{(r+1)}) \le \mathcal{L}(\mathbf{W}^{(r)})\\ &- \bar{\eta}T_{\text{avg}}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2 - \bar{\eta}T_{\text{avg}}\langle\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}, \nabla\mathcal{L}(\mathbf{W}^{(r)})\rangle\\ &+ \frac{(\bar{\eta}T_{\text{avg}})^2}{2}\|\nabla^2\mathcal{L}(\mathbf{V})\|_{\text{op}}\|\nabla\mathcal{L}(\mathbf{W}^{(r)}) + \mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2. \end{aligned} \tag{49}$$

Using Lemma 11, we have $\|\nabla^2\mathcal{L}(\mathbf{V})\|_{\text{op}} \le 2\mathcal{L}(\mathbf{V}) \le 2(\lambda\mathcal{L}(\mathbf{W}^{(r)}) + (1-\lambda)\mathcal{L}(\mathbf{W}^{(r+1)})) \le 2\mathcal{L}(\mathbf{W}^{(r)})$, where we again used $\mathcal{L}(\mathbf{W}^{(r+1)}) \le \mathcal{L}(\mathbf{W}^{(r)})$. Thus,

$$\begin{aligned}&\mathcal{L}(\mathbf{W}^{(r+1)}) \le \mathcal{L}(\mathbf{W}^{(r)}) - \bar{\eta}T_{\text{avg}}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2\\ &- \bar{\eta}T_{\text{avg}}\langle\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}, \nabla\mathcal{L}(\mathbf{W}^{(r)})\rangle\\ &+ 2(\bar{\eta}T_{\text{avg}})^2\mathcal{L}(\mathbf{W}^{(r)})\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2\\ &+ 2(\bar{\eta}T_{\text{avg}})^2\mathcal{L}(\mathbf{W}^{(r)})\|\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2. \end{aligned} \tag{50}$$

Note we have $2(\bar{\eta}T_{\text{avg}})\mathcal{L}(\mathbf{W}^{(r)}) \le 1/2$ by the hypothesis in this corollary. It yields

$$\begin{aligned}&\mathcal{L}(\mathbf{W}^{(r+1)}) \le \mathcal{L}(\mathbf{W}^{(r)}) - (\bar{\eta}T_{\text{avg}}/2)\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2\\ &- \bar{\eta}T_{\text{avg}}\langle\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}, \nabla\mathcal{L}(\mathbf{W}^{(r)})\rangle\\ &+ 2(\bar{\eta}T_{\text{avg}})^2\mathcal{L}(\mathbf{W}^{(r)})\|\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2. \end{aligned} \tag{51}$$

By Cauchy-Schwarz inequality and Lemma 11, we further notice that

$$\begin{aligned}&- \bar{\eta}T_{\text{avg}}\langle\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}, \nabla\mathcal{L}(\mathbf{W}^{(r)})\rangle\\ &+ 2(\bar{\eta}T_{\text{avg}})^2\mathcal{L}(\mathbf{W}^{(r)})\|\mathbf{E}^{(r)} + \tilde{\mathbf{E}}^{(r)}\|_F^2\\ &\le \frac{\bar{\eta}T_{\text{avg}}}{5}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2 + \frac{\bar{\eta}T_{\text{avg}}}{50}\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F^2. \end{aligned} \tag{52}$$

Combining terms yields

$$\mathcal{L}(\mathbf{W}^{(r+1)}) \le \mathcal{L}(\mathbf{W}^{(r)}) - \frac{\bar{\eta}T_{\text{avg}}}{4}\|\mathcal{L}(\mathbf{W}^{(r)})\|_F^2. \tag{53}$$

Further using $\|\nabla\mathcal{L}(\mathbf{W}^{(r)})\|_F \ge \frac{\gamma}{2}\mathcal{L}(\mathbf{W}^{(r)})$ (cf. Lemma 12), we conclude the proof of (29).

Appendix V
Auxiliary Lemmas

Lemma 13. *Under Assumptions 1, 2, the loss function $\mathcal{L}(\mathbf{W})$ is $1/2$-smooth.*

The above lemma can be verified directly by evaluating the eigenvalues of $\nabla^2\mathcal{L}(\mathbf{W})$.

Lemma 14. *If F_m is twice continuously differentiable, then for any $\mathbf{u},\mathbf{v}$ in a convex set $C = \{(1-t)\mathbf{u} + t\mathbf{v} : t\in(0,1)\}$, we have*

$$\|\nabla F_m(\mathbf{u}) - \nabla F_m(\mathbf{v})\| \le \left(\max_{\mathbf{w}\in C}\|\nabla^2 F_m(\mathbf{w})\|\right)\|\mathbf{u} - \mathbf{v}\|.$$

The above lemma is a direct consequence of Taylor's theorem for vector-valued functions.

APPENDIX VI
PROOFS OF LEMMAS IN THEOREM 4

A. Proof of Lemma 11

We first need to bound the gradient norm with relation to the loss. Observe that

$$\begin{aligned}&\|\nabla \ell_{i,j}(\mathbf{W})\|_F^2 = \textstyle\sum_{k=1}^K \|\nabla_{\mathbf{w}_k} \ell_{i,j}(\mathbf{W})\|^2 \\ &\le \|\mathbf{x}_{i,j}\|^2 \Big((1-\widehat{y}_{i,j|\mathbf{W}})^2 + \textstyle\sum_{k\neq y_{i,j}} (\widehat{y}^{(k)}_{i,j|\mathbf{W}})^2 \Big) \\ &\le K(1-\widehat{y}_{i,j|\mathbf{W}})^2 \|\mathbf{x}_{i,j}\|^2 \le K(1-\widehat{y}_{i,j|\mathbf{W}})^2. \end{aligned} \tag{54}$$

Since $-\log z \ge 1-z$ for $z \in (0,1]$, we have

$$\ell_i(\mathbf{W}) = -\log \widehat{y}_{i,j|\mathbf{W}} \ge 1 - \widehat{y}_{i,j|\mathbf{W}}, \tag{55}$$

$$\|\nabla \ell_i(\mathbf{W})\|_F \le \sqrt{K}(1-\widehat{y}_{i,j|\mathbf{W}}) \le \sqrt{K}\ell_i(\mathbf{W}). \tag{56}$$

Henceforth,

$$\begin{aligned}&\|\nabla \mathcal{L}(\mathbf{W})\|_F \le \tfrac{1}{n} \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \|\nabla \ell_{i,j}(\mathbf{W})\|_F \\ &\le \sqrt{K} \cdot \tfrac{1}{n} \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \ell_{i,j}(\mathbf{W}) = \sqrt{K}\mathcal{L}(\mathbf{W}). \end{aligned} \tag{57}$$

Next, we bound the Hessian. Observe that for the classes $k, l \in [K]$, we have

$$\nabla^2_{\mathbf{w}_k,\mathbf{w}_l} \ell_{i,j}(\mathbf{W}) = \begin{cases} \widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}) \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top, & \text{if } k = l, \\ -\widehat{y}^{(k)}_{i,j|\mathbf{W}} \widehat{y}^{(l)}_{i,j|\mathbf{W}} \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top, & \text{if } k \neq l. \end{cases} \tag{58}$$

For the diagonal block of the Hessian, we have

$$\begin{aligned}\|\nabla^2_{\mathbf{w}_k,\mathbf{w}_k} \ell_{i,j}(\mathbf{W})\|_{\text{op}} &= \widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}) \|\mathbf{x}_{i,j}\|^2 \\ &\le \widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}). \end{aligned} \tag{59}$$

For the off-diagonal blocks on row k, we have

$$\begin{aligned}\textstyle\sum_{l\neq k} \|\nabla^2_{\mathbf{w}_k,\mathbf{w}_l} \ell_{i,j}(\mathbf{W})\|_{\text{op}} &\le \textstyle\sum_{l\neq k} \widehat{y}^{(k)}_{i,j|\mathbf{W}} \widehat{y}^{(l)}_{i,j|\mathbf{W}} \|\mathbf{x}_{i,j}\|^2 \\ &= \widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}). \end{aligned} \tag{60}$$

Now, we use the Gershgorin Circle Theorem to get

$$\begin{aligned}&\|\nabla^2 \ell_{i,j}(\mathbf{W})\|_{\text{op}} \\ &\le \max_k \Big(\|\nabla^2_{\mathbf{w}_k,\mathbf{w}_k} \ell_{i,j}(\mathbf{W})\|_{\text{op}} + \textstyle\sum_{l\neq k} \|\nabla^2_{\mathbf{w}_k,\mathbf{w}_l} \ell_{i,j}(\mathbf{W})\|_{\text{op}} \Big) \\ &\le \max_k 2\, \widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}). \end{aligned} \tag{61}$$

If $k = y_{i,j}$, then

$$\widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}) \le 1 - \widehat{y}_{i,j|\mathbf{W}} \le -\log \widehat{y}_{i,j|\mathbf{W}} = \ell_{i,j}(\mathbf{W}). \tag{62}$$

Otherwise, we have

$$\widehat{y}^{(k)}_{i,j|\mathbf{W}} \le 1 - \widehat{y}_{i,j|\mathbf{W}} \le \ell_{i,j}(\mathbf{W}), \tag{63}$$

$$\begin{aligned}\widehat{y}^{(k)}_{i,j|\mathbf{W}}(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}) &\le (1-\widehat{y}_{i,j|\mathbf{W}})(1-\widehat{y}^{(k)}_{i,j|\mathbf{W}}) \\ &\le 1 - \widehat{y}_{i,j|\mathbf{W}} \le \ell_{i,j}(\mathbf{W}). \end{aligned} \tag{64}$$

This leads to the conclusion:

$$\begin{aligned}&\|\nabla^2 \mathcal{L}(\mathbf{W})\|_{\text{op}} \le \tfrac{1}{n} \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \|\nabla^2 \ell_{i,j}(\mathbf{W})\|_{\text{op}} \\ &\le \tfrac{1}{n} \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} 2\ell_{i,j}(\mathbf{W}) = 2\mathcal{L}(\mathbf{W}). \end{aligned} \tag{65}$$

B. Proof of Lemma 12

By Lemma 6-(ii), we observe that

$$\langle \nabla \mathcal{L}_i(\mathbf{W}), \mathbf{W}_\star \rangle \le -\frac{\gamma}{n_i} \sum_{j=1}^{n_i} \big(1 - \widehat{y}_{i,j|\mathbf{W}}\big) = -\gamma \mathcal{G}_i(\mathbf{W}). \tag{66}$$

Since $\mathcal{G}_i(\mathbf{W}) \ge 0$, we proceed by

$$\begin{aligned}&\langle \mathbf{W}^{(r+1)}, \mathbf{W}_\star \rangle \\ &= \langle \mathbf{W}^{(r)}, \mathbf{W}_\star \rangle - \frac{\bar{\eta}}{n} \sum_{i=1}^m \sum_{\tau=1}^{T_i} n_i \langle \nabla \mathcal{L}_i(\mathbf{D}^{(r)}_{i,\tau-1}), \mathbf{W}_\star \rangle \\ &\ge \langle \mathbf{W}^{(r)}, \mathbf{W}_\star \rangle + \frac{\bar{\eta}\gamma}{n} \sum_{i=1}^m \sum_{\tau=1}^{T_i} n_i \mathcal{G}_i(\mathbf{D}^{(r)}_{i,\tau-1}) \\ &\ge \langle \mathbf{W}^{(r)}, \mathbf{W}_\star \rangle + \frac{\bar{\eta}\gamma}{n} \mathcal{G}(\mathbf{W}^{(r)}). \end{aligned} \tag{67}$$

This yields

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathcal{G}(\mathbf{W}^{(r)}) \le \frac{\langle \mathbf{W}^{(R)}, \mathbf{W}_\star \rangle_F - \langle \mathbf{W}^{(0)}, \mathbf{W}_\star \rangle_F}{\gamma \eta R}. \tag{68}$$

Eq. (41) is then obtained using (3) and $\|\mathbf{W}_\star\|_F \le 1$.

By Lemma 6-(ii), we also have

$$\langle \nabla \mathcal{L}(\mathbf{W}), \mathbf{W}_\star \rangle \le -\gamma \mathcal{G}(\mathbf{W}). \tag{69}$$

Since $\mathcal{G}(\mathbf{W}) \ge 0$, we have $|\langle \nabla \mathcal{L}(\mathbf{W}), \mathbf{W}_\star \rangle| \ge \gamma \mathcal{G}(\mathbf{W})$. By Cauchy-Schwarz inequality, this implies

$$\begin{aligned}\|\nabla \mathcal{L}(\mathbf{W})\|_F &\ge \|\nabla \mathcal{L}(\mathbf{W})\|_F \cdot \|\mathbf{W}_\star\|_F \\ &\ge |\langle \nabla \mathcal{L}(\mathbf{W}), \mathbf{W}_\star \rangle| \ge \gamma \mathcal{G}(\mathbf{W}). \end{aligned} \tag{70}$$

Suppose that in addition, $\mathcal{G}(\mathbf{W}) \le 1/(2n)$, then since each term $(1 - \widehat{y}_{i,j|\mathbf{W}})$ is non-negative, we must have

$$1 - \widehat{y}_{i,j|\mathbf{W}} \le n \cdot \mathcal{G}(\mathbf{W}) \le 1/2 \tag{71}$$

for all i, j. Thus, $\widehat{y}_{i,j|\mathbf{W}} \ge 1/2$ for all i, j. For any $z \in (0,1]$, we have the inequality $-\log(z) \le (1-z)/z$ and thus

$$\ell_{i_j}(\mathbf{W}) \le \frac{1 - \widehat{y}_{i,j|\mathbf{W}}}{\widehat{y}_{i,j|\mathbf{W}}} \le 2\left(1 - \widehat{y}_{i,j|\mathbf{W}}\right). \tag{72}$$

Henceforth, we can derive that

$$\mathcal{L}(\mathbf{W}) \le \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} 2\left(1 - \widehat{y}_{i,j|\mathbf{W}}\right) = 2\mathcal{G}(\mathbf{W}), \tag{73}$$

and when combined with (70), this implies that $\|\nabla \mathcal{L}(\mathbf{W})\|_F \ge \gamma \mathcal{G}(\mathbf{W}) \ge \frac{\gamma}{2} \mathcal{L}(\mathbf{W})$.