Bias Mitigation for Federated Learning with Spatially Correlated Participation

Oussama Harrak, Malcolm Egan, Claire Goursaud, Marie Line Alberi Morel, Alberto Conte

Bias Mitigation for Federated Learning with Spatially Correlated Participation

Oussama Harrak, Malcolm Egan, Claire Goursaud, Marie Line Alberi Morel and Alberto Conte

Abstract

Federated learning is a key strategy to exploit data distributed throughout edge networks. However, clients may participate intermittently only when they observe relevant data. Moreover, in spatially-distributed sensing, the activity of nearby clients can be correlated. In this paper, we study the impact of correlation and variable-size active client sets for FedSGD. We analyze the convergence for smooth nonconvex learning objectives and show that bias due to unequal participation can be effectively mitigated via importance weighting. We validate our analysis with experiments using the MNIST dataset and show the impact of constraints on the availability of communication links.

Federated Learning, Partial Participation, Correlated Activity

I. Introduction

Data availability is a key constraint for learning algorithms in networks. When data is distributed among a large number of edge devices, reliable communication of all data to a centralized server is often infeasible. As the computational resources of edge devices improve, it is desirable to instead locally process data and only communicate local updates at each edge device to the server. This approach is known as federated learning (FL), and there is now strong evidence that FL can obtain models at the server with high accuracy and low communication requirements [1].

Due to memory constraints, edge devices may not have access to large historical datasets and must compute local models using streaming data collected online, and clients only compute updates when new data is available [2]. In the case where data collection only occurs when a rare event is observed, updates are sent intermittently to the server. This scenario arises in classification of rare events such as high pollution or water levels. As a consequence, only a subset of clients participate in each update of the server model, leading to partial participation. Moreover, the number of participating clients in each communication round can be highly variable.

In this paper, we study the impact of bias arising from partial participation with correlated and variable-size active client sets for smooth non-convex learning objectives. We first analyze the impact of this bias on the convergence of FedSGD (or minibatch SGD). Our analysis shows that the bias cannot be controlled only by reducing the step size.

O. Harrak, M. Egan and C. Goursaud are with Inria, INSA Lyon, CITI, UR3720, 69621 Villeurbanne. M. Alberi Morel and A. Conte are with Nokia Bell-Labs France, Massy, France.

To mitigate the bias, we propose a scheme where aggregated client gradient estimates are *weighted*, often known as importance weighting. In our scheme, the importance weights are estimated from samples of active client sets. We show that for large numbers of communication rounds, the estimation error is not a dominant factor affecting the convergence.

We validate our theoretical analysis with numerical experiments on the MNIST dataset. Our experiments demonstrate that the importance weighting scheme significantly improves the average loss. We also observe a significant performance degradation with strict constraints on the number of active clients that can reliably communicate. This suggests that resource allocation is critical with correlated client activity.

A. Related Work

This work focuses on mitigating bias arising from partial participation, which naturally arises in edge networks due to computation or communication constraints. A key challenge is to cope with heterogeneity introduced by only a subset of client participating. The work in [3] considered partial participation for i.i.d. active client sets with a fixed size. Cyclic participation, where clients are organized into groups that participate in a structured manner, has been investigated in [4], [5], and general periodic participation processes have been studied in [6]. Analysis of general partial participation structures has been investigated in [7]. We highlight that this work assumes fixed active client set sizes. Varying size active client sets have been investigated in [8], [9] for asynchronous FL, however, correlation between client activities was not explicitly considered.

When clients participate at different frequencies, bias is a key problem. Bias mitigation in FL with Markovian participation has been studied in [10], [11]. Both of these schemes exploit importance weighting; however, they focus on equal size active client sets. In contrast, we account for both correlation in client activities and variable of size active client sets.

II. System Model

A. Federated Learning

Consider a FL system with M clients and a server which seeks to learn a model by solving

$$\min_{w\in\mathbb{R}^d} \frac{1}{M}\sum_{m=1}^{M} F_m(w) := F(w), \tag{1}$$

where

$$F_m(w) = \mathbb{E}_{\xi_m}[f_m(w,\xi_m)], \tag{2}$$

with $f_m : \mathbb{R}^d \times \Xi_m \to \mathbb{R}$ and ξ_m is a random variable on the set of data points available at client m denoted by Ξ_m. In general, $f_m(\cdot, s),\ s \in \Xi_m$ is a non-convex function. We impose the following standard assumptions on the objectives and data distributions in (1).

Assumption 1. *The objective F is β-smooth; i.e.,*

$$\|\nabla F(w_1) - \nabla F(w_2)\| \leq \beta\|w_1 - w_2\|,\ \forall w_1, w_2 \in \mathbb{R}^d. \tag{3}$$

Moreover, $\|\nabla F(w)\|^2 \leq D,\ \forall w \in \mathbb{R}^d$.

FL aims to solve (1) via a distributed optimization method. In each communication round $t = 1, \dots, T$:

- A subset of clients $\mathcal{M}_t$ is active, where each client $m \in \mathcal{M}_t$ computes an update Δ_m.
- The updates $\Delta_m,\ m \in \mathcal{M}_t$ are communicated to the server, which then computes a new global model

$$w_t \leftarrow w_{t-1} - \frac{\eta}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} \Delta_m, \tag{4}$$

  where $\eta > 0$ is the step size.

A common algorithm to compute the updates Δ_m is FedSGD, where

$$\Delta_m = \frac{1}{K} \sum_{k=1}^{K} \nabla f_m(w_t, \xi_t^{(m,k-1)}), \tag{5}$$

where K is the number of samples available to client m and $(\xi_t^{(m,k-1)}, k \in [K])$ is the minibatch available to client m in communication round t.

B. Client Participation

A standard assumption in FL is that clients are activated in sets of fixed size. That is, $|\mathcal{M}_t|$ is constant for each communication round t. In contrast, we consider the scenario where client m participates at time t only if its data $(\xi_t^{(m,k-1)},\ k \in [K])$ is *relevant*. For example, a relevant data sample may correspond to extremes of pollution or water levels. In the case clients only rarely observe relevant data, they may participate only in a small number of communication rounds. Moreover, the number of clients that are active in each communication round may be highly variable.

To model the activity of clients, we define the vector $\mathbf{X}_t \in \{0,1\}^M$, where the element $X_{t,m}$ indicates the activity of client m. In particular, $X_{t,m} = 1$ if client m is active and $X_{t,m} = 0$ otherwise. The active client set in communication round t is then given by $\mathcal{M}_t = \{m \in [M] : X_{t,m} = 1\}$. We assume that the participation process $(\mathbf{X}_t,\ t \in [T])$ is i.i.d. in time.

On the other hand, the activity $X_{t,m}, X_{t,m'}$ of clients m and m', respectively, may be correlated; i.e., $\mathbb{E}[X_{t,m}X_{t,m'}] \neq \mathbb{E}[X_{t,m}]\mathbb{E}[X_{t,m'}]$. Correlation in client activity arises naturally when communication only occurs when a relevant event occurs. For example, two nearby clients observing temperature are often likely to both measure extreme temperature events. In general, $\mathbf{X}_t$ is therefore governed by a multivariate Bernoulli distribution.

In the remainder of this paper, we investigate the impact of rare relevant data and the correlation structure of client activities on the performance of FedSGD.

III. Convergence Analysis

In this section, we propose an analysis of the convergence of FedSGD with correlated client activities. Proofs are provided in the extended version [12].

Assumption 2. *Let*

$$c_m = \mathbb{E}\left[\frac{\mathbf{1}\{m \in \mathcal{M}_t\}}{|\mathcal{M}_t|}\right]. \tag{6}$$

Then, for all $w \in \mathbb{R}^d$, $t \in [T]$, and $m \in [M]$,

$$\mathbb{E}_t \left[\left\| \frac{\mathbf{1}\{m \in \mathcal{M}_t\}}{|\mathcal{M}_t| K} \sum_{k=1}^{K} \nabla f_m(w, \xi_t^{(m,k-1)}) - c_m \nabla F_m(w) \right\|^2 \right] \leq \frac{\sigma^2}{M}, \tag{7}$$

where $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on the history up to time t.

The quantity c_m plays a key role in our analysis, corresponding to the effective weight of client m. Note that correlation between clients implicitly impacts c_m via $|\mathcal{M}_t|$.

Theorem 1. *Suppose that the iterates (w_t) are generated by FedSGD and Assumptions 1 and 2 hold. If $\eta = \frac{1}{\sqrt{T}}$ and $T \geq 4\beta^2$, then*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[\|\nabla F(w_{t-1})\|^2 \right] \leq \frac{2\mathbb{E}[F(w_0) - F(w^*)]}{\sqrt{T}} + \frac{2M\beta\sigma^2}{\sqrt{T}} + MD \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2. \tag{8}$$

We highlight that the first term in Theorem 1 depends on the initial suboptimality gap and the second term depends on the variance of the stochastic gradient estimates. Note that the first and second terms diminish as the number of communication rounds increases. The third term is due to unequal participation of each client arising from biased stochastic gradient estimates. This term does not diminish with the number of communication rounds T and therefore prevents convergence to a stationary point as the number of communication rounds $T \to \infty$. In the following section, we apply importance weighting to reduce the impact of the bias.

IV. Bias Mitigation

In the convergence bound for FedSGD, there is a term due to biased gradient estimates which cannot be controlled by increasing the number of communication rounds T. As a consequence, FedSGD does not converge to a stationary point in the setting detailed in Sec. II; i.e.,

$$\mathbb{E}_t \left[\frac{1}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} \nabla F_m(w) \right] = \sum_{m=1}^{M} c_m \nabla F_m(w) \neq \frac{1}{M} \sum_{m=1}^{M} \nabla F_m(w) \tag{9}$$

Indeed, in the case of full client participation with independent activities, c_m in (6) reduces to $c_m = 1/M,\ m = 1, \ldots, M$, and the bias term in Theorem 1 vanishes. In contrast, under partial participation with correlated activities, we generally have

$$c_m = \mathbb{E}\left[\frac{\mathbf{1}\{m \in \mathcal{M}_t\}}{|\mathcal{M}_t|} \right] \neq \frac{1}{M},$$

so that the bias term does not disappear.

A common technique to remove bias in gradient estimates is importance weighting [10], [11], where the stochastic gradient estimate $\nabla f_m(w,\xi)$ of client m is replaced by $\frac{1}{Mc_m}\nabla f_m(w,\xi)$. However, importance weighting requires knowledge of $c_m,\ m=1,\ldots,M$, which must be estimated. As such, only an estimate $\hat{c}_m$ of c_m is available to the server to correct the bias, which still leads to a bias term in the case of FedSGD of

$$\text{Bias}_{\text{FedSGD}} = \frac{MD}{2}\sum_{m=1}^{M}\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_m}\right|^2. \tag{10}$$

We now investigate the choice of estimator for c_m. Suppose that in communication round t, the server has access to t samples $\{\mathbf{X}_1,\ldots,\mathbf{X}_t\}$ of client activities. A natural estimator of c_m is given by

$$\hat{c}_{m,t} = \begin{cases} \frac{1}{t}\sum_{i=1}^{t}\frac{\mathbf{1}\{X_{i,m}=1\}}{\sum_{j=1}^{M}X_{i,j}}, & \text{if } \frac{1}{t}\sum_{i=1}^{t}\frac{\mathbf{1}\{X_{i,m}=1\}}{\sum_{j=1}^{M}X_{i,j}} > \delta \\ \delta, & \text{otherwise,} \end{cases} \tag{11}$$

where $\delta>0$ is a fixed threshold. We establish the following convergence bounds for FedSGD and FedAvg with importance weighting using the estimator in (11).

Theorem 2. *Let $\delta>0$. Suppose the Assumptions in Theorem 1 hold, and $c_m>\delta,\ m\in[M]$. With the estimator in (11), there exists $0<R<\infty$, independent of T, such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(w_{t-1})\|^2] \le \frac{2\mathbb{E}[F(w_0)-F(w^*)]}{\sqrt{T}} + \frac{2\beta\sigma^2}{M\delta^2\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}D\sum_{m=1}^{M}\left(\left|1-\frac{c_m}{\delta}\right|^2\cdot 2\exp\left(-2t(c_m-\delta)^2\right)+\frac{R}{t\delta^2}\right), \tag{12}$$

In particular, $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(w_{t-1})\|^2]=\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$, where $\tilde{O}(\cdot)$ ignores logarithmic terms.

Theorem 2 shows that if clients have a non-zero activity probability, then the rate of convergence is not dominated by estimation error. However, when clients are rarely active, then the bias cannot be mitigated or there will be a penalty for small T due to a small value of δ. In the following section, we investigate the impact of biased gradient estimators via numerical experiments.

V. Numerical Results

We consider $M=10$ clients indexed from 1 to 10, which form three groups: $\mathcal{A}_0=\{1,2,3\}$, $\mathcal{A}_1=\{4,5,6,7\}$, and $\mathcal{A}_2=\{8,9,10\}$. The server aims to learn a classifier for a subset of the digits in the MNIST data set with images corresponding to $\{0,1,2\}$. In particular, the server seeks to train a feedforward neural network consisting of a fully connected input layer of size 784 (corresponding to 28×28 pixels), a hidden layer with 4 neurons and $\tanh$ activation, and a linear output layer for three-class classification.

Clients in the group $\mathcal{A}_i$, are only activated when they observe an image from the MNIST data set with label i. The probability that each client is activated is typically small. As such, an observation by client $m\in\mathcal{A}_i$ of an image with label i is viewed as a rare event.

To model correlation between client activities, client m in group $\mathcal{A}_i$ can only be activated when the indicator $\mathcal{E}_{i,t}\in\{0,1\}$ satisfies $\mathcal{E}_{i,t}=1$. If $\mathcal{E}_{i,t}=1$, the probability client m observes an image with label i is then given by $\Pr(X_{t,m}=1|\mathcal{E}_{i,t}=1)$.

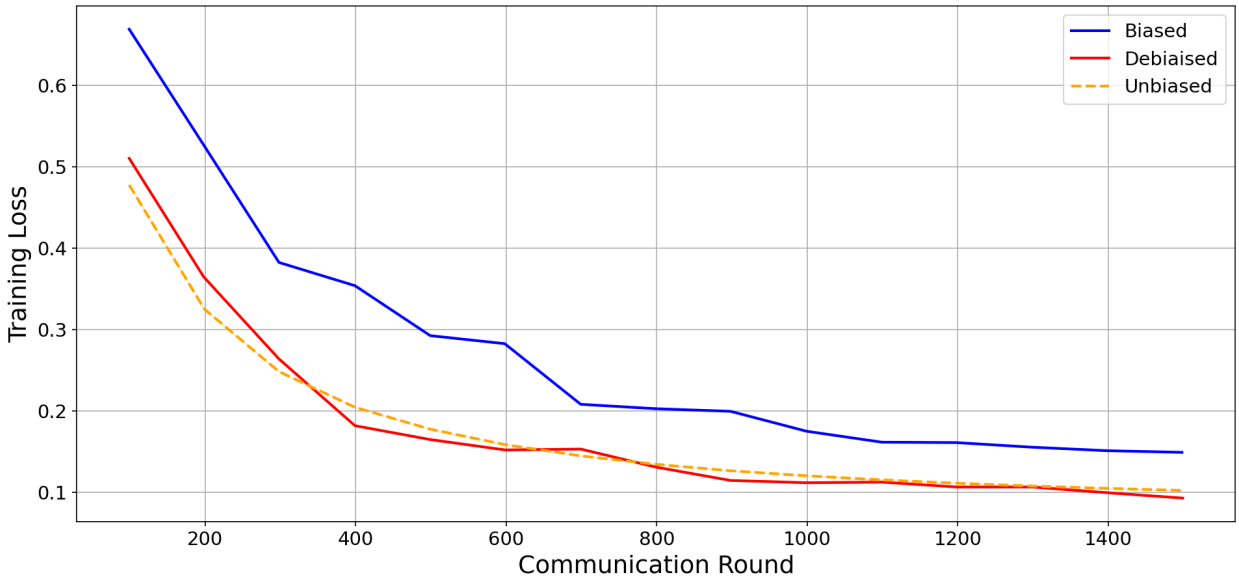

Fig. 1: Training loss averaged over 5 runs for unbiased, biased, and debiased via importance weighting schemes. Parameters: $T = 1500$, $\Pr(\mathcal{E}_0 = 1) = 0.3, \Pr(\mathcal{E}_1 = 1) = 0.6, \Pr(\mathcal{E}_2 = 1) = 0.3$. For client m in group i, $\Pr(X_{t,m} = 1|\mathcal{E}_i) = 0.95$ and $\Pr(X_{t,m} = 1|\mathcal{E}_i = 0) = 0$, $\eta = \frac{1}{\sqrt{T}} = 0.026$, $\delta = 0.01$.

As a consequence, the activities of clients in each group $\mathcal{A}_i$ are correlated due to dependence on the common variable $\mathcal{E}_{i,t}$. We also assume that the variables $\mathcal{E}_{i,t}$ and $\mathcal{E}_{i',t}$ are independent for $i \neq i'$, which implies that the activities of clients in distinct groups are independent. This leads to a joint probability mass function of the client activities given by

$$P_{\mathbf{X}_t}(\mathbf{x}) = \sum_{\mathbf{e}\in\{0,1\}^3} \prod_{i=0}^{2} \Pr(\mathcal{E}_{i,t} = e_i) \cdot \prod_{m\in\mathcal{A}_i} \Pr(X_{t,m} = x_m|\mathcal{E}_{i,t} = e_i). \tag{13}$$

Fig. 1 plots the training loss against the number of communication rounds for the client activity probabilities detailed in the caption. Results are averaged over 5 runs. Three curves are shown: *unbiased* corresponding to full participation; *biased* corresponding to partial participation (as in Sec. III); and *debiased* corresponding to partial participation with the importance weighting scheme in Sec. IV. Consistent with Theorems 1 and 2, the debiased curve converges to the unbiased curve while the biased curve does not achieve the same training loss, even for large numbers of communication rounds. The observation that the debiased curve converges to the unbiased curve suggests that the estimation error in the importance weights does not significantly impact performance for large T, consistent with Theorem 2.

In practice, communication constraints may limit the number of clients that can simultaneously send updates; e.g., due to limited available bandwidth. These communication constraints can have a significant impact when client activity is correlated, as clients are more likely to send updates at the same time than in the i.i.d. setting. To model communication constraints, we assume that clients in the group $\mathcal{A}_i$ can only successfully send updates to the server when a maximum of $M_{\max}$ clients in $\mathcal{A}_i$ are active. The constraint $M_{\max}$ can be viewed as the number of available orthogonal frequency bands allocated to each group of clients.

Fig. 2 shows the training loss for different values of $M_{\max}$, averaged over 5 runs, with the corresponding standard

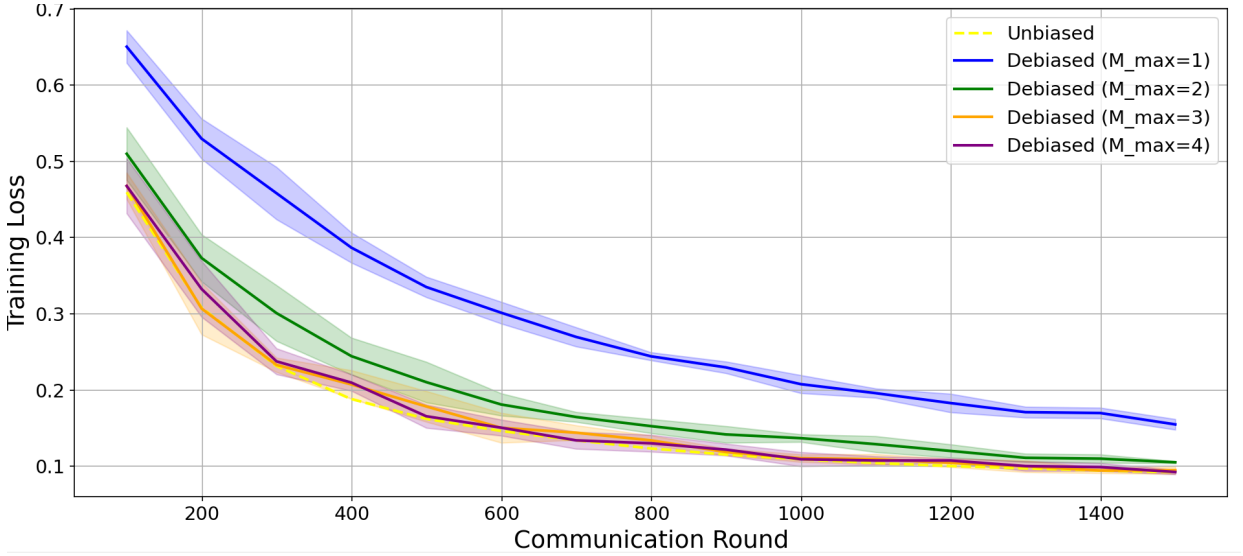

Fig. 2: Training loss averaged over 5 runs for unbiased, biased, and debiased via importance weighting schemes with constraints on the maximum number of active clients. Parameters : $T = 1500$, $\Pr(\mathcal{E}_0 = 1) = 0.3, \Pr(\mathcal{E}_1 = 1) = 0.6, \Pr(\mathcal{E}_2 = 1) = 0.3$. For client m in group i $\Pr(X_{t,m} = 1|\mathcal{E}_i) = 0.95$ and $\Pr(X_{t,m} = 1|\mathcal{E}_i = 0) = 0$, $\eta = \frac{1}{\sqrt{T}} = 0.026$, $\delta = 0.01$

deviations. Observe that debiasing via importance weighting still ensures convergence to the unbiased training loss. However, the convergence is slower than for the unbiased case. This is due to an increased variance of gradient estimates since the communication constraints prevent clients from participating, even when they observe data. This suggests the need for co-design of FL with the communication system to ensure that each active client can reliably communicate to the server.

VI. Conclusion

A key consequence of correlated participation with variable active set sizes is bias. We analyzed the bias and investigated a bias mitigation scheme known as importance weighting. Via numerical experiments, we validated our analysis and showed that correlation can lead to performance degradation when only a limited number of clients can be simultaneously active.

References

[1] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1-2, pp. 1–210, 2021.

[2] T. Huynh et al., "Streaming federated learning with Markovian data," *arXiv preprint arXiv:2503.18807*, 2025.

[3] S. Karimireddy et al., "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, 2020.

[4] Y. Cho et al., "On the convergence of federated averaging with cyclic client participation," in *International Conference on Machine Learning*, 2023.

[5] H. Eichner et al., "Semi-cyclic stochastic gradient descent," in *International Conference on Machine Learning*, 2019.

[6] M. Crawshaw and M. Liu, "Federated learning under periodic client participation and heterogeneous data: A new communication-efficient algorithm and analysis," in *Advances in Neural Information Processing Systems*, 2024.

[7] S. Wang and M. Ji, "A unified analysis of federated learning with arbitrary client participation," in *Advances in Neural Information Processing Systems*, 2022.

[8] D. Avdiukhin and S. Kasiviswanathan, "Federated learning under arbitrary communication patterns," in *International Conference on Machine Learning (ICML)*, 2021.

[9] Y. Yan et al., "Federated optimization under intermittent client availability," *INFORMS Journal on Computing*, vol. 36, no. 1, pp. 185–202, 2024.

[10] Z. Sun et al., "Debiasing federated learning with correlated client participation," in *International Conference on Machine Learning*, 2025.

[11] A. Rodio et al., "Federated learning under heterogeneous and correlated client availability," in *IEEE Conference on Computer Communications (INFOCOM)*, 2023.

[12] O. Harrak et al., "Bias mitigation for federated learning with spatially correlated participation," 2025, https://hal.science/hal-05265910.

APPENDIX

A. Preliminaries

Lemma 1. *Let $x, y \in \mathbb{R}^d$. Then,*

$$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle. \tag{14}$$

Lemma 2 (Relaxed Triangle Inequality)**.** *Let $x_m \in \mathbb{R}^d,\ m = 1, \dots, M$. Then,*

$$\left\|\sum_{m=1}^{M} x_m\right\|^2 \leq M \sum_{m=1}^{M} \|x_m\|^2. \tag{15}$$

B. Proof of Theorem 1

By Assumption 1,

$$\begin{aligned} F(w_t) &\leq F(w_{t-1}) - \eta \left\langle \nabla F(w_{t-1}), \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\rangle + \frac{\beta}{2}\|w_t - w_{t-1}\|^2 \\ &= F(w_{t-1}) - \eta \left\langle \nabla F(w_{t-1}), \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\rangle + \frac{\beta\eta^2}{2} \left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\|^2. \end{aligned} \tag{16}$$

Taking conditional expectations, we obtain

$$\begin{aligned} \mathbb{E}_t[F(w_t)] &\leq F(w_{t-1}) - \eta \left\langle \nabla F(w_{t-1}), \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\rangle + \frac{\beta\eta^2}{2}\mathbb{E}_t \left[\left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\|^2 \right] \\ &= F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 - \frac{\eta}{2}\left\| \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 + \frac{\eta}{2}\left\| \nabla F(w_{t-1}) - \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \\ &\quad + \frac{\beta\eta^2}{2}\mathbb{E}_t \left[\left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\|^2 \right] \\ &\overset{(i)}{\leq} F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 - \frac{\eta}{2}\left\| \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 + \frac{\eta}{2}\left\| \nabla F(w_{t-1}) - \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \\ &\quad + \beta\eta^2 \mathbb{E}_t \left[\left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) - \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \right] + \beta\eta^2 \left\| \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \\ &\overset{(ii)}{\leq} F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 + \frac{\eta}{2}\left\| \nabla F(w_{t-1}) - \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \\ &\quad + \beta\eta^2 \mathbb{E}_t \left[\left\| \sum_{m=1}^{M} \frac{\mathbf{1}\{m \in \mathcal{M}_t\}}{K|\mathcal{M}_t|} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) - \sum_{m=1}^{M} c_m \nabla F_m(w_{t-1}) \right\|^2 \right] \\ &\overset{(iii)}{\leq} F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 + \frac{\eta M D}{2} \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 \end{aligned}$$

$$+ M\beta\eta^2 \sum_{m=1}^{M} \mathbb{E}_t \left[\left\| \frac{\mathbf{1}\{m \in \mathcal{M}_t\}}{K|\mathcal{M}_t|} \sum_{k=1}^{K} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) - c_m \nabla F_m(w_{t-1}) \right\|^2 \right]$$

$$\overset{(iv)}{\leq} F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 + \frac{\eta M D}{2} \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 + M\beta\eta^2\sigma^2, \tag{17}$$

where (i) follows from the relaxed triangle inequality, (ii) follows from the step size condition $\frac{\eta}{2} \geq \beta\eta^2$, (iii) follows from the relaxed triangle inequality and the second inequality of Assumption 1 , and (iv) follows from Assumption 2.

Taking full expectations, we then have

$$\mathbb{E}[\|\nabla F(w_{t-1})\|^2] \leq \frac{2\mathbb{E}[F(w_{t-1}) - F(w_t)]}{\eta} + MD \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 + 2M\beta\eta\sigma^2 \tag{18}$$

Hence,

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla F(w_{t-1})\|^2] &\leq \frac{2\mathbb{E}[F(w_0) - F(w_T)]}{\eta T} + MD \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 + 2M\beta\eta\sigma^2 \\ &\leq \frac{2\mathbb{E}[F(w_0) - F(w^*)]}{\eta T} + MD \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 + 2M\beta\eta\sigma^2, \end{aligned} \tag{19}$$

where $w^* \in \arg\min_w F(w)$. Choosing $\eta = \frac{1}{\sqrt{T}}$ with $T \geq 4\beta^2$ (to ensure that $\frac{\eta}{2} \geq \beta\eta^2$),

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla F(w_{t-1})\|^2] \leq \frac{2\mathbb{E}[F(w_0) - F(w^*)]}{\sqrt{T}} + MD \sum_{m=1}^{M} \left| \frac{1}{M} - c_m \right|^2 + \frac{2M\beta\sigma^2}{\sqrt{T}}, \tag{20}$$

as required.

C. Proof of Theorem 2

Using Assumption 1 and taking conditional expectations yields

$$\begin{aligned} \mathbb{E}_t[F(w_t)] &\leq F(w_{t-1}) - \eta \left\langle \nabla F(w_{t-1}), \sum_{m=1}^{M} \frac{c_m}{M\hat{c}_{m,t}} \nabla F_m(w_{t-1}) \right\rangle \\ &\quad + \frac{\beta\eta^2}{2}\mathbb{E}_t \left[\left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \frac{1}{M\hat{c}_{m,t}} \nabla f_m(w_{t-1}, \xi_t^{(m,k-1)}) \right\|^2 \right]. \\ &\leq F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 - \frac{\eta}{2} \left\| \sum_{m=1}^{M} \frac{c_m}{M\hat{c}_{m,t}} \nabla F_m(w_{t-1}) \right\|^2 + \\ &\frac{\eta}{2} \left\| \nabla F(w_{t-1}) - \sum_{m=1}^{M} \frac{c_m}{M\hat{c}_{m,t}} \nabla F_m(w_{t-1}) \right\|^2 \\ &\quad + \frac{\beta\eta^2}{2}\mathbb{E}_t \left[\left\| \frac{1}{K|\mathcal{M}_t|} \sum_{m\in\mathcal{M}_t} \sum_{k=1}^{K} \frac{\nabla f_m(w_{t-1}, \xi_t^{(m,k-1)})}{M\hat{c}_{m,t}} \right\|^2 \right] \\ &\overset{(i)}{\leq} F(w_{t-1}) - \frac{\eta}{2}\|\nabla F(w_{t-1})\|^2 - \frac{\eta}{2} \left\| \sum_{m=1}^{M} \frac{c_m}{M\hat{c}_{m,t}} \nabla F_m(w_{t-1}) \right\|^2 + \\ &\frac{\eta}{2} \left\| \nabla F(w_{t-1}) - \sum_{m=1}^{M} \frac{c_m}{M\hat{c}_{m,t}} \nabla F_m(w_{t-1}) \right\|^2 \end{aligned}$$

$$
\begin{aligned}
&+\beta\eta^2\mathbb{E}_t\left[\left\|\frac{1}{K|\mathcal{M}_t|}\sum_{m\in\mathcal{M}_t}\sum_{k=1}^{K}\frac{\nabla f_m(w_{t-1},\xi_t^{(m,k-1)})}{M\hat{c}_{m,t}}-\sum_{m=1}^{M}\frac{c_m}{M\hat{c}_{m,t}}\nabla F_m(w_{t-1})\right\|^2\right]+\\
&\beta\eta^2\left\|\sum_{m=1}^{M}\frac{c_m}{M\hat{c}_{m,t}}\nabla F_m(w_{t-1})\right\|^2\\
&\overset{(ii)}{\leq} F(w_{t-1})-\frac{\eta}{2}\|\nabla F(w_{t-1})\|^2+\frac{\eta}{2}\left\|\nabla F(w_{t-1})-\sum_{m=1}^{M}\frac{c_m}{M\hat{c}_{m,t}}\nabla F_m(w_{t-1})\right\|^2\\
&\quad+\beta\eta^2\mathbb{E}_t\left[\left\|\sum_{m=1}^{M}\frac{\mathbf{1}\{m\in\mathcal{M}_t\}}{K|\mathcal{M}_t|}\sum_{k=1}^{K}\frac{\nabla f_m(w_{t-1},\xi_t^{(m,k-1)})}{M\hat{c}_{m,t}}-\sum_{m=1}^{M}\frac{c_m}{M\hat{c}_{m,t}}\nabla F_m(w_{t-1})\right\|^2\right]\\
&\overset{(iii)}{\leq} F(w_{t-1})-\frac{\eta}{2}\|\nabla F(w_{t-1})\|^2+\frac{\eta MD}{2}\sum_{m=1}^{M}\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\\
&\quad+\frac{\beta\eta^2}{M\delta^2}\sum_{m=1}^{M}\mathbb{E}_t\left[\left\|\frac{\mathbf{1}\{m\in\mathcal{M}_t\}}{K|\mathcal{M}_t|}\sum_{k=1}^{K}\nabla f_m(w_{t-1},\xi_t^{(m,k-1)})-c_m\nabla F_m(w_{t-1})\right\|^2\right]\\
&\overset{(iv)}{\leq} F(w_{t-1})-\frac{\eta}{2}\|\nabla F(w_{t-1})\|^2+\frac{\eta MD}{2}\sum_{m=1}^{M}\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2+\frac{\beta\eta^2\sigma^2}{M\delta^2}
\end{aligned}
\tag{21}
$$

where (i) follows from the relaxed triangle inequality, (ii) follows from the step size condition $\frac{\eta}{2}\geq\beta\eta^2$, (iii) follows from the relaxed triangle inequality, the second inequality of Assumption 1 and the fact that $\hat{c}_{m,t}\geq\delta$, and (iv) follows from Assumption 2. Taking the full expectation we then have :

$$
\mathbb{E}[\|\nabla F(w_{t-1})\|^2]\leq\frac{2\mathbb{E}[F(w_{t-1})-F(w_t)]}{\eta}+MD\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right]+\frac{2\beta\eta\sigma^2}{M\delta^2}
\tag{22}
$$

Therefore ,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(w_{t-1})\|^2]&\leq\frac{2\mathbb{E}[F(w_{t-1})-F(w_t)]}{\eta T}+\frac{1}{T}\sum_{t=1}^{T}MD\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right]+\frac{2\beta\eta\sigma^2}{M\delta^2}\\
&\leq\frac{2\mathbb{E}[F(w_{t-1})-F(w^*)]}{\eta T}+\frac{1}{T}\sum_{t=1}^{T}MD\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right]+\frac{2\beta\eta\sigma^2}{M\delta^2}
\end{aligned}
\tag{23}
$$

Hence, we need to bound

$$
\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right],
\tag{24}
$$

Define

$$
Z_{mi}=\frac{\mathbf{1}\{X_{i,m}=1\}}{\sum_{j=1}^{M}X_{i,j}},
\tag{25}
$$

and

$$
\overline{Z}_{mt}=\frac{1}{t}\sum_{i=1}^{t}Z_{mi}.
\tag{26}
$$

Then,

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right] &= \frac{1}{M}\sum_{m=1}^{M}\left(\left|1-\frac{c_m}{\delta}\right|^2\Pr(\overline{Z}_{mt}\le\delta)+\mathbb{E}\left[\left|1-\frac{c_m}{\overline{Z}_{mt}}\right|^2\middle|\overline{Z}_{mt}>\delta\right]\Pr\left(\overline{Z}_{mt}>\delta\right)\right)\\ &\le \frac{1}{M}\sum_{m=1}^{M}\left(\left|1-\frac{c_m}{\delta}\right|^2\Pr(\overline{Z}_{mt}\le\delta)+\mathbb{E}\left[\frac{(Z_{mt}-c_m)^2}{\delta^2}\middle|\overline{Z}_{mt}>\delta\right]\mathbb{P}(\overline{Z}_{mt}>\delta)\right)\\ &\le \frac{1}{M}\sum_{m=1}^{M}\left(\left|1-\frac{c_m}{\delta}\right|^2\Pr(\overline{Z}_{mt}\le\delta)+\frac{S}{t\delta^2}\mathrm{Var}(Z_{m1})\right),\end{aligned} \tag{27}$$

where

$$S=\sup_u\frac{\Pr(Z_{mt}=u|\overline{Z}_{mt}\ge\delta)}{\Pr(Z_{mt}=u)}. \tag{28}$$

By assumption $c_m>\delta$. Hence ,

$$\begin{aligned}\Pr(\overline{Z}_{mt}\le\delta) &= \Pr\left(tc_m-\sum_{i=1}^{t}Z_{mi}\ge t(c_m-\delta)\right)\\ &\le \Pr\left(\left|\sum_{i=1}^{t}Z_{mi}-tc_m\right|\ge t(c_m-\delta)\right)\\ &\le 2\exp\left(-2t(c_m-\delta)^2\right),\end{aligned} \tag{29}$$

which follows from the fact that $\mathbb{E}[Z_{mi}]=c_m$ and Hoeffding's inequality.

We then have

$$\sum_{m=1}^{M}\mathbb{E}\left[\left|\frac{1}{M}-\frac{c_m}{M\hat{c}_{m,t}}\right|^2\right]\le\frac{1}{M}\sum_{m=1}^{M}\left(\left|1-\frac{c_m}{\delta}\right|^2\cdot 2\exp\left(-2t(c_m-\delta)^2\right)+\frac{S}{t\delta^2}\mathrm{Var}(Z_{m1})\right) \tag{30}$$

setting $R=S\cdot\mathrm{Var}(Z_{m1})$ yields the desired result.