Federated Learning with Client Batching and Heterogeneous Data

Alexandria Pawlik

August 31, 2020

Abstract

Federated learning is a method for training deep learning models on decentralized data that limits data storage and protects client privacy. Introducing client batching is a common practice and can help to reduce overall communication and computation costs, but isn't necessary in all scenarios. In this paper, we investigate whether the importance of cohort size depends on heterogeneity of client data. We present our experiments and give proof of a bounded convergence speed for the global average model. We use experimental results and convergence bounds to outline when cohort size is most important and why.

1 Introduction

Federated learning (FL) is a distributed way of training deep neural networks on parallel clients. Clients are part of a federation managed by a central server which periodically combines the trained client models, most simply with model averaging [3]. The primary benefits of FL are limiting storage of data on the server and increasing client privacy by sending trained model parameters between client and server rather than collecting client data on the server.

In this paper, we investigate the use of federated learning with non-convex optimization and non-IID clients. Specifically, we study whether the importance of cohort size depends on heterogeneity of data. First, we present our experimental setup and results from model training with Google's TensorFlow Federated. We then give an in-depth proof of model averaging convergence efficiency from Yu et al. 2018 [4], followed by an extension where we generalize model averaging to the federated learning scenario and look again at convergence efficiency. We finish with an interpretation of cohort size and heterogeneity of data in federated learning applications.

1.1 Effects of Heterogeneous Clients

Distributed optimization in the convex setting has been researched significantly, but often assumes that client data are IID from a single distribution and client sample sizes are all the same. This is rarely the case in situations where FL might be most effective. For example, when personal mobile devices are used as clients, data distributions are different for each client and sample size would need to be limited to guarantee that all clients use the same number of data.

Data distributions have an important effect on FL because of their hierarchical structure. In practice, each client in the federation will have a different local data distribution. However, the server is optimizing the global model for a unified distribution which is a combination of all individual client distributions. This creates the case where the model will not converge during non-convex optimization. Because clients each run multiple local epochs, they have the potential to locally trend towards different stationary points before model averaging. This can produce global models that just aren't useful, which is why we chose to take a look at the effects heterogeneity of client data and the different ways for client data to be somewhat IID.

1.2 Effects of Cohort Size

In practice, we expect to have little-to-no control over how homogeneous client data is. This is where client batching comes into play. We predict that the importance of selecting a cohort size increases as data heterogeneity increases. We explain with two extreme examples.

Consider a large federation of clients with identical data. After each client runs gradient descent for some number of local epochs, they will end up with identical models. In this situation, any cohort size will produce the same average model and therefore cohort size is irrelevant.

Now consider the opposite case where a large federation of clients have very different data distributions. After each client runs a few epochs locally, their models will be more appropriate for their data distribution, but not necessarily for other clients' distributions. Using a cohort of all clients would have an unrealistic communication cost, and using a single client would produce a global model that's useless to all other clients. In this situation, cohort size would need to be selected very carefully to allow model averaging to remove the inherent bias of using sub-distributions to estimate a global distribution without compromising communication cost.

Looking at these edge cases, we see how the number of clients selected might depend on the distribution of client data. Next, we explain the layout of our experiments built to test this relationship.

2 Experimental Setup

To test the relationship between heterogeneity of data, cohort size, and model convergence, we used experiments that simulate federated learning using Google's TensorFlow Federated. Our experimental setup emulated that used in McMahan et al. 2016 [3], utilizing the same convolutional neural network (CNN) structure and MNIST dataset.

Our CNN model was sequential and converted pixel arrays to a digit 0-9. Layers were as follows:

- 5x5 2D Convolution (ReLU activation with 32 channels)
- 2x2 Max Pooling
- 5x5 2D Convolution (ReLU activation with 64 channels)
- 2x2 Max Pooling
- Flatten
- Densely-Connected (ReLU activation with 512 units)
- Densely-Connected (Softmax activation with 10 units)

There were 1,663,370 total model parameters.

For all trials, we used 100 clients and gave each 600 observations and labels. Each client trained for some number of local epochs, where each epoch was completed when the client had trained on all 600 data. Clients used stochastic gradient descent (SGD) with a varying learning rate, calculating loss as sparse categorical cross-entropy and accuracy as sparse categorical accuracy. The central server utilized Google's Federated Averaging process where selected client models were averaged at the end of each global round, which occurred after all selected clients finished running their local epochs in parallel. Other hyperparameters were changed periodically during experimentation and are further described in our results section.

Our experiments address the effects of heterogeneity of data by utilizing a few flexible data partitioning schemes. Various cohort sizes were implemented with random selection before each global training round. We did our best to maintain fairness between trials as determined by a few different measures of fairness.

2.1 TensorFlow Federated

Our model was built with Python and the Keras package, wrapped for use with federated learning, and trained and tested with TensorFlow Federated (TFF), Google's open-source library for machine learning. TFF simulates federated learning by creating local clients that train simultaneously, acting as a distributed client network. Basic datasets are available through both TensorFlow (TF) and TFF, already wrapped for their corresponding uses. We worked with TF's MNIST dataset (60,000 observations) to give us more control over data partitioning before we wrapped the data into a TFF-compatible dataset.

Before training, TFF requires some data preprocessing, though this is not entirely clear through their documentation. Data can be batched, shuffled, repeated, and more before being passed to the training process. Implementing cohorts in TFF is simple, as clients are represented simply by their datasets and dataset objects can be easily batched. We trained the model by passing the current global model parameters and selected client datasets to an iterative federated learning method that handles the implementation of parallel distributed training over each passed dataset. An updated model and training metrics were returned. Each global round was followed by a centralized testing round that used an independent TF MNIST test set (10,000 new observations). TF functions handled this testing for us because TFF implements testing as a compiled metric on distributed testing and we wanted to eliminate this source of randomness in our results.

2.2 Data Partitioning

To investigate heterogeneity of data, we came up with a few data partitioning methods to test. The implemented ones are discussed below. Note that these all sort by MNIST labels first.

2.2.1 Partially IID Data

For this scheme, we chose to control the percent of each client's data that would be IID from the overall MNIST dataset. Without a way to measure heterogeneity of data directly, we chose to use "IID-ness" of clients as a proxy for how similar client data distributions were. The percent of client data that was non-IID consisted of MNIST data from a limited number of labels such that its distribution was different from the population.

A fraction of the data from each of the 10 labels was pulled aside and mixed together as an IID data pool. Each client randomly selected their IID data from the pool, with overlap between clients allowed but minimal - data multiplicities were counted to confirm this. Next, each client randomly selected 2 labels and pulled half of their non-IID data from each, with overlap between clients allowed but again confirmed minimal. The only parameter required by this scheme was the percent of each client's data that should be IID.

Two special cases of this scheme exist, where the first is all clients fully IID and the other is all clients fully non-IID. In practice, only the fully IID case produced reasonable results, and a different partitioning method was needed to make all clients fully non-IID.

2.2.2 Sharding

Sharding the data by label [3] gave more consistent results for testing on clients with non-IID data. For each label, data was randomized then split into 20 shards of size 300. Shards (200 total) are then shuffled and 2 are distributed to each client.

2.3 Fairness of Trials

To address global fairness of trials, we concern ourselves primarily with constant communication and computation costs. Using the algorithm presented in McMahan et al. 2016 [3], an estimate of each of these values can be computed and held constant between runs.

2.3.1 Constant Number of Local Gradients Calculated (Per Global Round)

The most expensive computation done is gradient calculation, which we expect to occur once per data point per local batch.

The expected number of gradients per round (for all clients) is given by

$$g = CEN \tag{1}$$

where C is the number of clients per cohort, E is the number of local epochs, and N is the number of data points per client (600 for our MNIST dataset and 100 clients). The number of data points per client multiplied by the number of local epochs gives the number of data points processed each local training round. Note that we represent number of data points per client as a single variable, as opposed to dividing total data points by total number of clients, to allow for data overlap between clients.

To prioritize keeping this g value constant, we must manipulate one of C, E, N each time another is changed. For example, if we double the number of local epochs, we halve the number of data points per client or the cohort size. We may choose to keep the total number data points per client constant, since in practice we have little-to-no control over this value.

2.3.2 Constant Total Number of Local Gradients Calculated

To predict the total number of gradients calculated over the entire trial, we simply factor in R, the total number of communication rounds, such that

$$g_R = Ru = RCEN \tag{2}$$

where all variables are as above.

This may not be practical in our local trials that run until a certain target accuracy instead of for a fixed number of rounds. However, in practice we would be able to pick how many rounds we want to run.

2.3.3 Constant Number of Local Gradients Calculated (Per Global Round Per Client)

In order to measure computation on a per-client basis, we can make a slight modification to the previous value.

To calculate the expected number of updates per round per client in the cohort, we use

$$g' = \frac{g}{C} = EN \tag{3}$$

where all variables are as above.

To prioritize keeping this g' value constant, we must manipulate one of E, N each time another is changed. Again, we may choose to keep N constant. In this case, we are only looking at E, so we would keep it constant in order to keep g' constant.

2.3.4 Constant Total Number of Local Gradients Calculated (Per Client)

To predict the total number of gradients per client calculated over the entire trial, we include R and the ratio of cohort size to total number of clients to give the expected frequency of selection of a single client, such that

$$g'_R = \frac{g_R}{K} = \frac{RCg'}{K} = \frac{RCEN}{K} \tag{4}$$

where all variables are as above.

2.3.5 Constant Number of Local Updates (Per Global Round)

Another major computation cost is that of updating the model. Each time gradients are calculated for the points in a local batch, we make an update to the local model. This means that number of updates will essentially be represented by a count of local batches over all clients.

$$u = \frac{EN}{B} \tag{5}$$

Here B is the local batch size and all other variables are as above. Note that we do not factor cohort size into this u value, even though the above equation seems to represent number of local updates per client per round. This gives us a u value better equipped to predict the upper bound on how far we'll travel through gradient space each round.

To prioritize keeping this u value constant, we must manipulate one of E, N, B each time another is changed. For example, if we double the number of local epochs we can choose to either double the local batch size or halve the number of data points per client. Again, we may choose to keep N constant. In this case, we are only looking at the inversely proportional values of E and B, so we would apply the same multiplicative changes to them in order to keep u constant.

2.3.6 Constant Total Number of Local Updates

To predict the total number of updates made to local models over the entire trial, we simply factor in R, the total number of communication rounds, such that

$$u_R = Ru = \frac{REN}{B} \tag{6}$$

where all variables are as above.

2.3.7 Constant Number of Contacts with Server

To measure overall communication costs, we'll look at the number of times that clients contact the central server to send trained models. These server updates occur once per client per global round, so we'll represent the number of contacts with the central server as

$$c = RC \tag{7}$$

where C is the number of clients per cohort and R is the number of communication rounds.

To prioritize keeping this c value constant, we must manipulate R and C in an inversely proportional way.

2.3.8 Constant Number of Contacts with Server (Per Client)

If we care about communication costs for each client more than overall costs, we can make a slight modification to the previous value. The average number of contacts with the central server per client is given by

$$c' = \frac{c}{K} = \frac{RC}{K} \tag{8}$$

where K is the total number of clients and all other variables are as above. Note that this value is equivalent to the number of times we expect any given client to participate in a training round during a single trial.

In practice, we don't expect to have much control over the total number of clients, so keeping c' constant would be similar to keeping c constant.

2.3.9 In Practice

In order to use cohort size as our main predictor, we needed to determine a multiplicative relationship between cohort size and all other variables such that a maximum number of measures of fairness are kept constant between trials.

Each time we doubled cohort size, we halved the total number of rounds and the local batch size. This lead to constant values from equation 2 (total number of gradients calculated), 3 (number of gradients calculated per round per client in cohort), 4 (total number of gradients calculated per client), 6 (total number of updates to model), 7 (total number of contacts with central server), and 8 (total number of contacts with central server per client). This means that we would be able to keep constant all local communication, global communication, local computation, and all global computation not measured on a per-round basis.

C	R	B	g	g_R	g'	g'_R	u	u_R	c	c'
5	12	60	3000	36000	600	360	12	144	60	0.6
10	6	30	6000	36000	600	360	12	72	60	0.6
15	4	20	9000	36000	600	360	12	48	60	0.6
20	3	15	12000	36000	600	360	12	36	60	0.6
30	2	10	18000	36000	600	360	12	24	60	0.6

Table 1: Sample values for parameters and each measurement of fairness for each cohort size.

The maximum number of rounds allowed (here it's 12) was set after running a batch of tests that allowed the model to run until 90 percent accuracy. An upper limit was chosen based on simplicity of division in the above table.

2.4 Sources of Randomness

We examined all accessible sources of randomness in our code to ensure that they were controlled. We aimed to get the exact same results given a seed value and percent IID-ness. Each source of randomness is described below.

- The sample batch passed to the model constructor is used to set the initial parameter values of the model. This batch contains all of the data and is shuffled the same way every run, given a seed value.
- The dataset used to test the model is the same every run and is shuffled the same way every run, given a seed value.
- Client datasets are given a shuffle seed upon addition to the client array. The shuffle seed is the same every run and the datasets are set to reshuffle each local epoch. TensorFlow's API is unclear about if there is any way to shuffle datsets at the beginning of each global round, however it may not be very important to shuffle within datsets given the way SGD works.
- All shuffling during data partitioning is handled by a random number generator that's seeded the same way every run, given a seed value.
- Selection of clients for each cohort is done with a second random number generator that's seeded using the seed value and current round number.

In practice, our results showed identical data partitioning given a seed value and percent IID-ness, but not identical training and testing results. This indicates some hidden randomness in the TensorFlow Federated library, meaning that our trials were not perfectly reproducible.

3 Experimental Results

Our training process was structured such that we could test an array of values for any combination of random shuffler seeds, percent of client data that should be IID, cohort size, number of local epochs, and local batch size. For the first half of our experiments, runs lasted until the model reached some fixed target accuracy as per the experimental setup in McMahan et al. 2016 [3]. Halfway through, we developed our measures of a fair trial. One of the relevant variables in these equations is the total number of training rounds, so the second half of our experiments were restructured to run until reaching some round number limit.

3.1 Measuring Number of Rounds to Reach Target Accuracy

Our initial experiments focused around tuning our model. We tested our model on MNIST data in the federated learning scenario with only a single client to confirm that it was as efficient as the same model trained in a centralized learning scenario. This helped us to find the errors in our model and training process. We further tuned our model by training it on an array of learning rates and using the one that reached target accuracy the quickest. We found that different percentages of IID client data consistently required different learning rates. The GitHub for this project contains the code for all experiments, including a few others that we used to tune our model. Ultimately, we focused on running tests with an array of values for cohort size, percent IID-ness of client data, and shuffle seed (to increase our sample size).

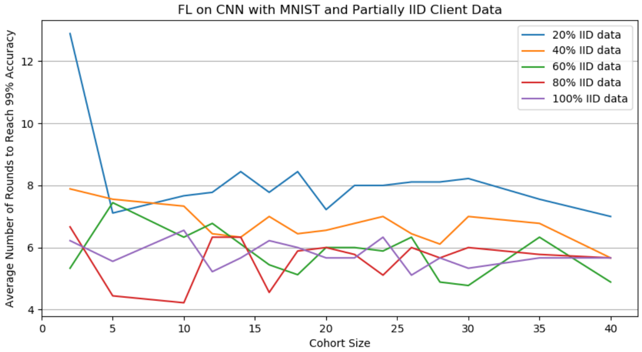

Figure 1: Number of communication rounds to reach 99% accuracy for varying cohort size, averaged over 9 shuffle seeds. Learning rate 0.1, batch size 10, and 10 local epochs. Client data are 20%, 40%, 60%, 80%, or 100% IID. Cohort sizes tested include 2, 5, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 35, and 40.

In the first half of our experiments, we let runs proceed until 90% and 99% model accuracy or until they hit some computation limit. For any given set of runs, all executed until the same target accuracy no matter the cohort size, percent IID-ness of data, or shuffle seed.

We found that runs with lower percent IID-ness of client data had somewhat higher round counts. You can see that this is vaguely the case in Figure 1 despite the noise. This aligned with our expectations because more IID client data leads to less biased client training, meaning that the federated average model is less noisy and more likely to converge quicker.

Though we expected number of rounds required to reach target accuracy to decrease with increase in cohort size (for any percent IID-ness), we got many noisy results before we started to see this trend. To eliminate the noise, we made small corrections to our sources of randomness and checked the data selection overlap between clients. At this point, we incorporated the previously defined measures of fairness.

3.2 Measuring Maximum Accuracy Reached within Round Limit

After we found that number of rounds was an important variable to control when managing fairness of trials, the second half of our experiments ran until some round limit. We restructured our training procedure, then determined round limits by cohort size (this relationship can be derived from the fairness equations presented previously and is shown in Table 1).

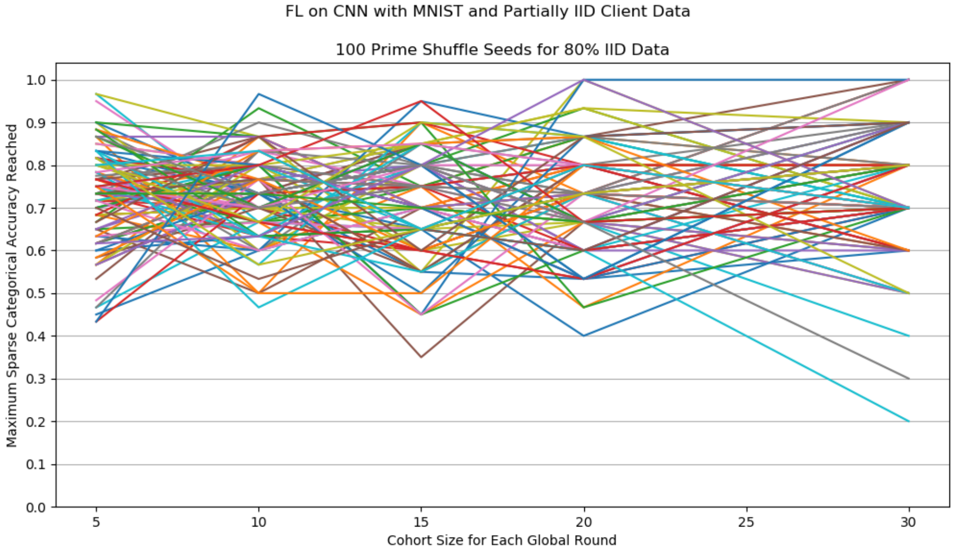

Figure 2: Maximum sparse categorical accuracy reached for varying cohort size for 100 different shuffle seeds. Learning rate 0.1 and 1 local epoch. Client data are 80% IID. Cohort sizes include 5, 10, 15, 20, and 30. Round limit and batch size are both inversely proportional to cohort size as per Table 1.

We continued to see noisy and incoherent trends, so we increased our sample size (see Figure 2). Using 100 trials presented us with a very problematic error in our data: it appeared as if for any given cohort size, there were limited values that maximum accuracy could take. We investigated this by individually changing our seeding procedure, changing our model compilation, and doubling all rounds limits. After finding the same problematic trend, we looked at the model's predictions compared to observations to see if our model was making biased predictions. We ran out of time before determining the source of this output grouping.

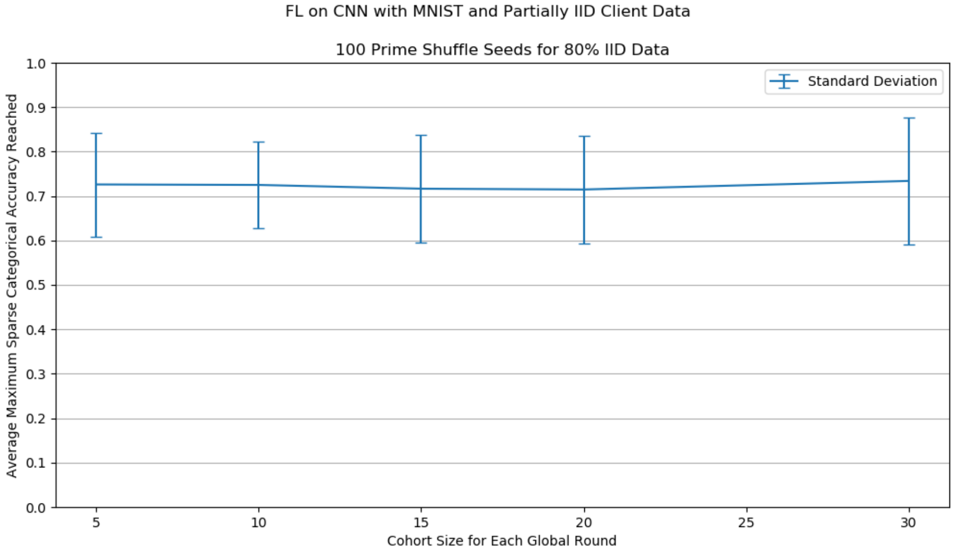

Figure 3: Data from Figure 2 with maximum sparse categorical accuracy reached averaged over 100 seeds.

Looking back on our final experiments, we find evidence suggesting that with all relevant computation and communication costs constant, cohort size doesn't matter. Compare Figure 2 to the averaged data in Figure 3. The average accuracy plot is relatively flat, which presents an interesting trend to investigate in the future. However, the large error bars are reason to further address the output grouping bug first as it may indicate an issue in the training procedure or with the distribution of the data itself.

4 Model Averaging

After our experiments, we also wanted to understand the efficiency of federated learning from a theoretical standpoint. The paper that we based our experiments on [3] didn't include a proof, so we utilize a different paper that looks at a similar situation. First we go in depth on the proof from [4] to help us understand it better. Then we extend it to make it more similar to the federated learning scenario, adding in our variable of interest: cohort size.

Model averaging, which periodically takes a global average of a series of individual models trained over N parallel workers for T iterations, is a common practice used for distributed training of deep neural networks. Compared with parallel mini-batch SGD, the communication overhead of model averaging is significantly reduced. Model averaging for convex optimization with IID data has been studied extensively, however here we look at the non-convex case.

We prove that model averaging can achieve a good rate for model convergence of $O(1/\sqrt{NT})$ as long as the global averaging interval I and learning rate γ are carefully controlled.

4.1 Equations

Consider the distributed training of deep neural networks over multiple parallel clients [1] where all clients can access all or partial training data and aim to find a common global model with the minimum average training loss. This can be modeled as the following distributed parallel non-convex optimization

$$\min_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) \tag{Problem 1}$$

where

$$f(\mathbf{x}) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}) \tag{9}$$

and N is the number of clients and each $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\zeta}_i}\left[F_i(\mathbf{x};\boldsymbol{\zeta}_i)\right]$ is a smooth non-convex function with $\boldsymbol{\zeta}_i \sim \mathcal{D}_i$ where $\mathcal{D}_i$ can possibly be different for different $i \in N$ clients.

Define $\boldsymbol{\zeta}^{[t-1]}$ as the randomness up to iteration t such that $\boldsymbol{\zeta}^{[t-1]} \triangleq [\boldsymbol{\zeta}_i^{\tau}]_{i\in\{1,2,\ldots,N\},\tau\in\{1,2,\ldots,t-1\}}$. This paper assumes that each worker can locally observe unbiased independent stochastic gradients (around the last iteration solution $\mathbf{x}_i^{t-1}$) given by

$$\mathbf{G}_i^t = \nabla F_i(\mathbf{x}_i^{t-1};\boldsymbol{\zeta}_i^t)$$

with

$$\mathbb{E}_{\boldsymbol{\zeta}_i^t}\left[\mathbf{G}_i^t \middle| \boldsymbol{\zeta}^{[t-1]}\right] = \nabla f_i(\mathbf{x}_i^{t-1}), \forall i$$

The distributed training process described above is detailed in the following algorithm.

Algorithm 1: Parallel Restarted SGD

```
Input : Initialize $\mathbf{x}_i^0 = \bar{\mathbf{y}} \in \mathbb{R}^m$. Set learning rate $\gamma > 0$ and node synchronization
        interval (integer) $I > 0$
Output: Averaged global model $\mathbf{x}^T$.
for $t = 1$ to $T$ do
    Each node $i$ observes an unbiased stochastic gradient $\mathbf{G}_i^t$ of $f_i(\cdot)$ at point $\mathbf{x}_i^{t-1}$
    if $t$ is a multiple of $I$, i.e., $tmodI = 0$ then
        Calculate node average $\bar{\mathbf{y}} \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{t-1}$
        Each node $i$ in parallel updates its local solution
            $\mathbf{x}_i^t = \bar{\mathbf{y}} - \gamma \mathbf{G}_i^t$        (10)
    else
        Each node $i$ in parallel updates its local solution
            $\mathbf{x}_i^t = \mathbf{x}_i^{t-1} - \gamma \mathbf{G}_i^t$        (11)
    end
end
```

Though this algorithm describes parallel restarted SGD, it also strong corresponds with FL in that the global average model is calculated periodically.

For each iteration t, define the average of local solutions $\mathbf{x}_i^t$ over all N clients as

$$\bar{\mathbf{x}}^t \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^t \tag{12}$$

It follows that

$$\bar{\mathbf{x}}^t = \bar{\mathbf{x}}^{t-1} - \gamma\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t \tag{13}$$

Theorem 1, which gives an upper bound on the average of the expectation of the norm squared of the expectation of the gradients each given the randomness up to the current iteration $\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right]$ over $T > 1$ iterations, shows that model averaging can, in expectation, achieve $O(1/\sqrt{NT})$ convergence to some stationary point for non-convex optimization by averaging only every $I = O(T^{1/4}/N^{3/4})$ iterations and setting learning rate $\gamma = \frac{\sqrt{N}}{L\sqrt{T}}$.

4.2 Assumption

We will assume that equation (9) satisfies the following assumptions.

1. Smoothness: *Each function $f_i(\mathbf{x})$ is Lipschitz continuous with modulus L.*

2. Bounded variances and second moments: *There exists constants $\sigma > 0$ and $G > 0$ such that*

$$\mathbb{E}_{\boldsymbol{\zeta}_i\sim\mathcal{D}_i}\left[\|\nabla F_i(\mathbf{x};\boldsymbol{\zeta}_i) - \nabla f_i(\mathbf{x})\|^2\right] \le \sigma^2, \forall \mathbf{x}, \forall i$$

$$\mathbb{E}_{\boldsymbol{\zeta}_i\sim\mathcal{D}_i}\left[\|\nabla F_i(\mathbf{x};\boldsymbol{\zeta}_i)\|^2\right] \le G^2, \forall \mathbf{x}, \forall i$$

4.3 Theorem

If $0 < \gamma \le \frac{1}{L}$ then for all $T \ge 1$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \le \frac{2}{\gamma T}(f(\bar{\mathbf{x}}^0) - f^*) + 4\gamma^2 I^2 G^2 L^2 + \frac{L}{N}\gamma\sigma^2$$

where f^* is the minimum value for training loss (9).

4.4 Proof

Fix $t \ge 1$. By the smoothness of f assumption and the property of Lipschitz continuous functions $f(y) \le f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$, we know that

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^t)\right] \le \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] + \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\rangle\right] + \frac{L}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\|^2\right] \tag{14}$$

First we will look at the third term on the right side of (14).

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\bar{\mathbf{x}}^t-\bar{\mathbf{x}}^{t-1}\|^2\right]$$

By equation (13), we know $\bar{\mathbf{x}}^t-\bar{\mathbf{x}}^{t-1}=-\gamma\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t$.

$$=\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\gamma\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\|^2\right]$$

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\|^2\right] \tag{15}$$

We can separate the expectation over $\boldsymbol{\zeta}^{[t]}$ into 2 expectations using the Law of Iterated Expectations.

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\|^2\middle|\boldsymbol{\zeta}^{[t-1]}\right]\right]$$

Applying the base inequality $\mathbb{E}\left[\|Z\|^2\right]=\mathbb{E}\left[\|Z-\mathbb{E}\left[Z\right]\|^2\right]+\|\mathbb{E}\left[Z\right]\|^2$ that holds for any vector Z, where $Z=\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t$, we get

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t-\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\middle|\boldsymbol{\zeta}^{[t-1]}\right]\|^2\middle|\boldsymbol{\zeta}^{[t-1]}\right]+\|\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\middle|\boldsymbol{\zeta}^{[t-1]}\right]\|^2\right]$$

By the definition of $\mathbf{G}$ where $\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\mathbf{G}_i^t\middle|\boldsymbol{\zeta}^{[t-1]}\right]=\nabla f_i(\mathbf{x}_i^{t-1})$ then

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\middle|\boldsymbol{\zeta}^{[t-1]}\right]+\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\|^2\middle|\boldsymbol{\zeta}^{[t-1]}\right]+\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

Distributing the outer expectation, we get

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\|^2\right]+\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

The first term contains a norm squared of a sum in the form $\|\frac{1}{N}\sum_{i=1}^{N}p_i\|^2$. We know that generally, $\|\sum_{i=1}^{N}p_i\|^2=\sum_{i=1}^{N}\|p_i\|^2+\sum_{i\neq j}p_i\cdot p_j$, therefore $\|\frac{1}{N}\sum_{i=1}^{N}(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\|^2=\frac{1}{N}\sum_{i=1}^{N}\|(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\|^2+\frac{1}{N}\sum_{i\neq j}(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\cdot(\mathbf{G}_{\mathbf{j}}^{\mathbf{t}}-\nabla f_j(\mathbf{x}_j^{t-1}))$. Substituting into the first term and holding over the second, we get

$$=\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\frac{1}{N}\sum_{i=1}^{N}\|(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\|^2+\frac{1}{N}\sum_{i\neq j}(\mathbf{G}_i^t-\nabla f_i(\mathbf{x}_i^{t-1}))\cdot(\mathbf{G}_{\mathbf{j}}^{\mathbf{t}}-\nabla f_j(\mathbf{x}_j^{t-1}))\right]$$
$$+\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

$$= \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right]$$

$$+ \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i \neq j} (\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})) \cdot (\mathbf{G}_{\mathbf{j}}^{\mathbf{t}} - \nabla f_j(\mathbf{x}_j^{t-1})) \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

Because each $\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})$ is independent across clients, an expectation of a dot product is equivalent to the dot product of expectations.

$$= \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right] + \gamma^2 \frac{1}{N} \sum_{i \neq j} \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})) \right]$$

$$\cdot \, \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[(\mathbf{G}_{\mathbf{j}}^{\mathbf{t}} - \nabla f_j(\mathbf{x}_j^{t-1})) \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

By the Law of Iterated Expectations

$$= \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right] + \gamma^2 \frac{1}{N} \sum_{i \neq j} \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\mathbb{E}_{\boldsymbol{\zeta}^t} \left[\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right] \right]$$

$$\cdot \, \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\mathbb{E}_{\boldsymbol{\zeta}^t} \left[\mathbf{G}_{\mathbf{j}}^{\mathbf{t}} - \nabla f_j(\mathbf{x}_j^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right] \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

Each $\mathbb{E}_{\boldsymbol{\zeta}^t} \left[\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right]$ in the second term does not rely on other clients' randomness, so it is equivalent to $\mathbb{E}_{\boldsymbol{\zeta}_i^t} \left[\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right]$.

$$= \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right] + \gamma^2 \frac{1}{N} \sum_{i \neq j} \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\mathbb{E}_{\boldsymbol{\zeta}_i^t} \left[\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right] \right]$$

$$\cdot \, \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\mathbb{E}_{\boldsymbol{\zeta}_j^t} \left[\mathbf{G}_{\mathbf{j}}^{\mathbf{t}} - \nabla f_j(\mathbf{x}_j^{t-1}) \middle| \boldsymbol{\zeta}^{[t-1]} \right] \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

By the definition of $\mathbf{G}$ where $\mathbb{E}_{\boldsymbol{\zeta}_i^t} \left[\mathbf{G}_i^t \middle| \boldsymbol{\zeta}^{[t-1]} \right] = \nabla f_i(\mathbf{x}_i^{t-1})$ then each $\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})$ has $\mathbf{0}$ mean in expectation, we can eliminate the entire second term.

$$= \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

$$= \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\|(\mathbf{G}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2 \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1}) \|^2 \right]$$

By the definition $\mathbf{G}_i^t = \nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t)$ we know

$$= \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\zeta}^{[t]}} \left[\|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

and by the Law of Iterated Expectations

$$= \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\mathbb{E}_{\boldsymbol{\zeta}^t} \left[\|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \middle| \boldsymbol{\zeta}^{[t-1]} \right] \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

We can set an upper bound on this sum to replace the expectation over $\boldsymbol{\zeta}^{[t-1]}$ using the fact that the mean will be smaller than the maximum case.

$$\leq \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \sup_{\boldsymbol{\zeta}^{[t-1]}} \mathbb{E}_{\boldsymbol{\zeta}^t} \left[\|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \middle| \boldsymbol{\zeta}^{[t-1]} \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

This is the same as $\sup_{\mathbf{x}_i^{t-1}}$ because all randomness in $\mathbf{x}_i^{t-1}$ is a result of the randomness in $\boldsymbol{\zeta}^{[t-1]}$.

$$\leq \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \sup_{\mathbf{x}_i^{t-1}} \mathbb{E}_{\boldsymbol{\zeta}^t} \left[\|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \middle| \boldsymbol{\zeta}^{[t-1]} \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

Because we're using $\sup_{\mathbf{x}_i^{t-1}}$ to replace the random variable $\mathbf{x}$ with the worst-case expectation, giving an upper bound on the entire statement, our conditioning on $\boldsymbol{\zeta}^{[t-1]}$ is made into a conditioning on a constant worst-case. Therefore this upper bound created is equivalent to

$$\leq \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\zeta}^t} \left[\sup_{\mathbf{x}_i^{t-1}} \|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

Each $\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})$ is not dependent on other clients' randomness at this point, so the expectation over $\boldsymbol{\zeta}^t$ is equivalent to an expectation over $\boldsymbol{\zeta}_i^t$.

$$\leq \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \sup_{\boldsymbol{\zeta}^{[t-1]}} \mathbb{E}_{\boldsymbol{\zeta}_i^t} \left[\|\nabla F_i(\mathbf{x}_i^{t-1}; \boldsymbol{\zeta}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right] + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

By Assumption 1, we know that $\mathbb{E}_{\boldsymbol{\zeta}_i} \left[\|\nabla F_i(\mathbf{x}; \boldsymbol{\zeta}_i) - \nabla f_i(\mathbf{x})\|^2 \right] \leq \sigma^2, \forall \mathbf{x}, \forall i$, therefore

$$\leq \gamma^2 \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2 + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

$$\leq \gamma^2 \sigma^2 \frac{1}{N} + \gamma^2 \mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}} \left[\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i^{t-1})\|^2 \right]$$

We already factored out the variation due to $\boldsymbol{\zeta}^t$, so it is equivalent to write the expectation to be over $\boldsymbol{\zeta}^{[t]}$.

$$\leq \gamma^2\sigma^2\frac{1}{N} + \gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] \tag{16}$$

Next we will look at the second term on the right side of (14).

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\rangle\right]$$

By equation (13), we know $\bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1} = -\gamma\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t$.

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\rangle\right]$$

By the Iterated Law of Expectations

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\rangle\Big|\boldsymbol{\zeta}^{[t-1]}\right]\right]$$

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\langle\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\nabla f(\bar{\mathbf{x}}^{t-1})\Big|\boldsymbol{\zeta}^{[t-1]}\right], \mathbb{E}_{\boldsymbol{\zeta}^t}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\Big|\boldsymbol{\zeta}^{[t-1]}\right]\rangle\right]$$

Because $\bar{\mathbf{x}}^{t-1}$ is determined by $\boldsymbol{\zeta}^{[t-1]} = [\boldsymbol{\zeta}^1, ..., \boldsymbol{\zeta}^{t-1}]$

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \mathbb{E}_{\boldsymbol{\zeta}^t}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\Big|\boldsymbol{\zeta}^{[t-1]}\right]\rangle\right]$$

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\mathbf{G}_i^t\Big|\boldsymbol{\zeta}^{[t-1]}\right]\rangle\right]$$

By the definition of $\mathbf{G}$ where $\mathbb{E}_{\boldsymbol{\zeta}^t}\left[\mathbf{G}_i^t\Big|\boldsymbol{\zeta}^{[t-1]}\right] = \nabla f_i(\mathbf{x}_i^{t-1})$ then

$$= -\gamma\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\rangle\right]$$

Using the basic identity $\langle\mathbf{z}_1, \mathbf{z}_2\rangle = \frac{1}{2}(\|\mathbf{z}_1\|^2 + \|\mathbf{z}_2\|^2 - \|\mathbf{z}_1 - \mathbf{z}_2\|^2)$ for any two vectors of the same length, we know that

$$= -\frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t-1]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2 + \|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2 - \|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

We already factored out the variation due to $\boldsymbol{\zeta}^t$, so it is equivalent to write the expectation to be over $\boldsymbol{\zeta}^{[t]}$.

$$= -\frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2 + \|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2 - \|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] \quad (17)$$

Substituting equations (16) and (17) back into equation (14), we get

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right] \leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] + \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\langle\nabla f(\bar{\mathbf{x}}^{t-1}), \bar{\mathbf{x}}^{t} - \bar{\mathbf{x}}^{t-1}\rangle\right] + \frac{L}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\bar{\mathbf{x}}^{t} - \bar{\mathbf{x}}^{t-1}\|^2\right]$$

$$\leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] + \frac{-\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] + \frac{-\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$
$$+ \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] + \frac{L}{2}\gamma^2\sigma^2\frac{1}{N} + \frac{L}{2}\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

$$\leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] + \frac{-\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] + \frac{L\gamma^2}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$
$$+ \frac{-\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] + \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] + \frac{L\gamma^2\sigma^2}{2N}$$

$$\begin{aligned}\leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \frac{\gamma - L\gamma^2}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] + \frac{-\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \\ + \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] + \frac{L\gamma^2\sigma^2}{2N}\end{aligned} \quad (18)$$

The second-to-last term in the above equation can be simplified further.

$$\frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

By equation (9), we know that

$$= \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{\mathbf{x}}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

$$= \frac{\gamma}{2N^2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\sum_{i=1}^{N}(\nabla f_i(\bar{\mathbf{x}}^{t-1}) - \nabla f_i(\mathbf{x}_i^{t-1}))\|^2\right]$$

Using the rule $\|\sum_{i=1}^{N}\mathbf{z}_i\|^2 \leq N\sum_{i=1}^{N}\|\mathbf{z}_i\|^2$ for any vectors

$$\leq \frac{\gamma}{2N}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\sum_{i=1}^{N}\|\nabla f_i(\bar{\mathbf{x}}^{t-1}) - \nabla f_i(\mathbf{x}_i^{t-1})\|^2\right]$$

By the smoothness of each f_i from Assumption 1 and by the Lipschitz equation $|f(x) - f(x')| \leq L|x - x'|$, we know that

$$\leq \frac{\gamma}{2N}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\sum_{i=1}^{N} L^2\|\bar{\mathbf{x}}^{t-1} - \mathbf{x}_i^{t-1}\|^2\right]$$

The proof of Lemma 1 is simple algebra but was not included in this paper for brevity. We use Lemma 1 for the equation $\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\bar{\mathbf{x}}^{t} - \mathbf{x}_i^{t}\|^2\right] \leq 4\gamma^2 I^2 G^2$, therefore

$$\leq \frac{\gamma}{2N}\sum_{i=1}^{N} 4L^2\gamma^2 I^2 G^2$$

$$\leq \frac{\gamma}{2N}4NL^2\gamma^2 I^2 G^2$$

$$\leq 2\gamma^3 L^2 I^2 G^2$$

When we plug this term back into equation (18), we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right] \leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \frac{\gamma - L\gamma^2}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{t-1})\|^2\right] \\ - \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] + 2\gamma^3 L^2 I^2 G^2 + \frac{L\gamma^2\sigma^2}{2N}\end{aligned} \tag{19}$$

Under Theorem 1 we know $0 < \gamma \leq \frac{1}{L}$. Due to the inequality, we can drop the second term to get

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right] \leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] + 2\gamma^3 L^2 I^2 G^2 + \frac{L\gamma^2\sigma^2}{2N} \tag{20}$$

$$\frac{\gamma}{2}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right] + 2\gamma^3 L^2 I^2 G^2 + \frac{L\gamma^2\sigma^2}{2N}$$

If we divide both sides by $\frac{\gamma}{2}$, we get

$$\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2}{\gamma}(\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right]) + 4\gamma^2 L^2 I^2 G^2 + \frac{L\gamma\sigma^2}{N} \tag{21}$$

Because we're looking at a non-convex loss function, smaller gradients indicate when we approach a local optimum. If we sum all gradients and find a reasonable value, then we can confirm that the model will converge in expectation. Summing both sides over $t \in \{1, \ldots, T\}$ and dividing by T yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{1}{T}\sum_{t=1}^{T}(\frac{2}{\gamma}(\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right]) + 4\gamma^2L^2I^2G^2 + \frac{L\gamma\sigma^2}{N})$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2}{T\gamma}\sum_{t=1}^{T}(\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t-1})\right] - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{t})\right]) + 4\gamma^2L^2I^2G^2 + \frac{L\gamma\sigma^2}{N}$$

The right side summation will mostly self-cancel, resulting in

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2}{T\gamma}(\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{0})\right] - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{T})\right]) + 4\gamma^2L^2I^2G^2 + \frac{L\gamma\sigma^2}{N}$$

Since $f(\bar{\mathbf{x}}^0)$ has no relationship to t, we can simplify to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2}{T\gamma}(f(\bar{\mathbf{x}}^{0}) - \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{T})\right]) + 4\gamma^2L^2I^2G^2 + \frac{L\gamma\sigma^2}{N}$$

Let f^* be the minimum value of equation (9), therefore $f^* \leq \mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[f(\bar{\mathbf{x}}^{T})\right]$ and we can substitute

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2}{T\gamma}(f(\bar{\mathbf{x}}^{0}) - f^*) + 4\gamma^2L^2I^2G^2 + \frac{L\gamma\sigma^2}{N}$$

4.5 Corollary

This corollary follows by substituting γ and I values into Theorem 1:

Consider equation (9) under Assumption 1. Let $T \geq N$.

1. If we choose $\gamma = \frac{\sqrt{N}}{L\sqrt{T}}$ in Algorithm 1 then we have $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \leq \frac{2L}{\sqrt{NT}}(f(\bar{\mathbf{x}}^0) - f^*) + \frac{4N}{T}I^2G^2 + \frac{1}{\sqrt{NT}}\sigma^2$.

2. If we further choose $I \leq \frac{T^{1/4}}{N^{3/4}}$, then $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \leq \frac{2L}{\sqrt{NT}}(f(\bar{\mathbf{x}}^0) - f^*) + \frac{4}{\sqrt{NT}}G^2 + \frac{1}{\sqrt{NT}}\sigma^2 = O(\frac{1}{\sqrt{NT}})$ where f^* is the minimum value of equation (9).

4.6 Interpretation

By this corollary, we are able to say that in expectation, model averaging can achieve $O(1/\sqrt{NT})$ convergence to some stationary point for non-convex optimization. Let's look closer at each of the three terms used to bound our sum of gradients in part 2 of this corollary.

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2\right] \leq \frac{2L}{\sqrt{NT}}(f(\bar{\mathbf{x}}^{0}) - f^*) + \frac{4}{\sqrt{NT}}G^2 + \frac{1}{\sqrt{NT}}\sigma^2$$

The first term is a multiple of the distance from our initial model to the model that minimizes global loss. We would expect this distance to play a part in how long a model takes to converge because in general, the further a model starts from a finishing point the longer it will take to get there.

The next term is a multiple of G^2, which by our assumption $\mathbb{E}_{\boldsymbol{\zeta}_i \sim \mathcal{D}_i}\left[\|\nabla F_i(\mathbf{x}; \boldsymbol{\zeta}_i)\|^2\right] \leq G^2, \forall \mathbf{x}, \forall i$ tells us that the term represents how big our gradients can be. We expect this to affect convergence because in general, larger gradients indicate that a model is further from a stationary point.

The final term is a multiple of σ^2, which by our assumption $\mathbb{E}_{\boldsymbol{\zeta}_i \sim \mathcal{D}_i}\left[\|\nabla F_i(\mathbf{x}; \boldsymbol{\zeta}_i) - \nabla f_i(\mathbf{x})\|^2\right] \leq \sigma^2, \forall \mathbf{x}, \forall i$ tells us that the term depends on how much variance exists in the gradients of individual clients. It is a function of both the noise in client data and variance due to SGD. It makes sense that this noise would affect convergence rate because local gradients with more variation produce global average gradients with more variation, which cause the model to take longer to settle at a stationary point.

5 Model Averaging with Cohorts

Model averaging differs from our federated learning process primarily in the use of client batching. We take the previous proof and introduce client batching with cohorts of size C for each averaging process. To choose cohorts, we define random variables $\mathbf{E}^{(s)}$ as binary vectors of size N containing C '1's and $N - C$ '0's. Our new training process uses the following algorithm.

For simplicity in our algorithm, we do not use the original proof's t round notation. Instead, let S be the total number of global training rounds and T be the number of local rounds per global round. Let $\mathbf{G}_i^{(s,t)}$ be the unbiased stochastic gradient of f_i observed at $\mathbf{x}_i^{(s,t-1)}$.

Algorithm 2: Parallel Restarted SGD with Cohorts

```
Input  : Initialize $\mathbf{x}^{(0)} = \bar{\mathbf{y}} \in \mathbb{R}^m$. Set learning rate $\gamma > 0$, node synchronization
         interval (integer) $T > 0$, and cohort size $C > 0$.
Output: Averaged global model $\mathbf{x}^{(S)}$.
for global round $s = 1$ to $S$ do
    For each client $i$:
        $\mathbf{x}_i^{(s,0)} = \bar{\mathbf{x}}^{(s-1)}$
    Draw binary vector $\mathbf{E}^{(s)}$ uniformly such that $\sum_{i=1}^{N} \mathbf{E}_i^{(s)} = C$
    for local round $t = 1$ to $T$ do
        For each client $i$:
            $\mathbf{x}_i^{(s,t)} = \mathbf{x}_i^{(s,t-1)} - \gamma \mathbf{G}_i^{(s,t)}$        (22)
    end
    Find averaged global model:
        $\bar{\mathbf{x}}^{(s)} = \frac{1}{C} \sum_{i=1}^{N} \mathbf{x}_i^{(s,T)} \mathbf{E}_i^{(s)}$
end
```

Note that in this algorithm, all clients are trained regardless of which were selected for averaging. This process differs from our experiments where only selected clients were trained. Both produce the same global model, however notation in our algorithm is made simpler by training all clients. We avoided this in practice to save computation costs.

5.1 Simplifying Assumptions

Now, our goal is to incorporate this new source of randomness $\mathbf{E}^{[s]}$ due to client batching into the proof to look at the effects on convergence. We want to start by looking at a scenario where $\mathbf{E}^{[s]}$ is the only source of randomness, so we'll eliminate $\boldsymbol{\zeta}^{[t]}$. To do this, we consider all clients to have infinite but not identical data and change SGD to gradient descent. All random variables $\boldsymbol{\zeta}^{[t]}$ are then constant and each gradient is the true population gradient. This makes our first analysis of $\mathbf{E}^{[s]}$ much easier and gives us a guide for future work. Additionally, we expect the relationships described in the proof to somewhat hold because all $\boldsymbol{\zeta}^{[t]}$ and $\mathbf{E}^{[s]}$ are independent.

5.2 Equations

Consider the distributed training of deep neural networks over multiple parallel clients [1] where all clients have infinite data so that each gradient is the true population gradient.

Define $\mathbf{E}^{[s-1]}$ as the randomness due to client selection up to global iteration s such that $\mathbf{E}^{[s-1]} \triangleq [\mathbf{E}^{\tau}]_{\tau \in \{1,2,\ldots,s-1\}}$.

Define $\mathbf{E}_i^{(s)}$ as a single binary value indicating whether client i is included in batch s.

For each global iteration s, define the average of local solutions $\bar{\mathbf{x}}^{(s)}$ over all C selected clients as

$$\bar{\mathbf{x}}^{(s)} = \bar{\mathbf{x}}^{(s-1)} - \gamma \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \sum_{t=1}^{T} \mathbf{G}_i^{(s,t)} \tag{23}$$

For simplicity of notation, also define $\mathbf{G}_i^{(s)}$ as the overall update to client i during global round s such that

$$\mathbf{G}_i^{(s)} = \sum_{t=1}^{T} \mathbf{G}_i^{(s,t)}$$

and $\bar{\mathbf{G}}^{(s)}$ as the average overall update to all N clients during global round s such that

$$\begin{aligned}\bar{\mathbf{G}}^{(s)} &= \mathbb{E}_{\mathbf{E}^{(s)}} \left[\frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} \middle| \mathbf{E}^{[s-1]} \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{G}_i^{(s)}\end{aligned}$$

By the definition of $\mathbf{E}_i^{(s)}$, we can also say that

$$\bar{\mathbf{G}}^{(s)} = \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \bar{\mathbf{G}}^{(s)}$$

5.3 Proof

By the smoothness of f assumption and the property of Lipschitz continuous functions $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$, we know that

$$\mathbb{E}_{\mathbf{E}^{[s]}} \left[f(\bar{\mathbf{x}}^{(s)}) \right] \leq \mathbb{E}_{\mathbf{E}^{[s]}} \left[f(\bar{\mathbf{x}}^{(s-1)}) \right] + \mathbb{E}_{\mathbf{E}^{[s]}} \left[\langle \nabla f_i(\bar{\mathbf{x}}^{(s-1)}), \bar{\mathbf{x}}^{(s)} - \bar{\mathbf{x}}^{(s-1)} \rangle \right] + \frac{L}{2} \mathbb{E}_{\mathbf{E}^{[s]}} \left[\|\bar{\mathbf{x}}^{(s)} - \bar{\mathbf{x}}^{(s-1)}\|^2 \right] \tag{24}$$

First we will look at the third term on the right side of (24).

$$\mathbb{E}_{\mathbf{E}^{[s]}} \left[\|\bar{\mathbf{x}}^{(s)} - \bar{\mathbf{x}}^{(s-1)}\|^2 \right]$$

By equation (23), we know $\bar{\mathbf{x}}^{(s)} - \bar{\mathbf{x}}^{(s-1)} = -\gamma \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \sum_{t=1}^{T} \mathbf{G}_i^{(s,t)}$.

$$= \mathbb{E}_{\mathbf{E}^{[s]}} \left[\|-\gamma \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \sum_{t=1}^{T} \mathbf{G}_i^{(s,t)}\|^2 \right]$$

$$= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s]}} \left[\|\frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \sum_{t=1}^{T} \mathbf{G}_i^{(s,t)}\|^2 \right]$$

Using the definition of $\mathbf{G}_i^{(s)}$,

$$= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s]}} \left[\| \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} \|^2 \right]$$

We can separate the expectation over $\mathbf{E}^{[s]}$ into 2 expectations using the Law of Iterated Expectations.

$$= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\| \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} \|^2 \middle| \mathbf{E}^{[s-1]} \right] \right]$$

Applying the base inequality $\mathbb{E}\left[\|Z\|^2\right] = \mathbb{E}\left[\|Z - \mathbb{E}\left[Z\right]\|^2\right] + \|\mathbb{E}\left[Z\right]\|^2$ that holds for any vector Z, where $Z = \frac{1}{C}\sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)}$, we get

$$\begin{aligned} &= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\| \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} - \mathbb{E}_{\mathbf{E}^{(s)}} \left[\frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} \middle| \mathbf{E}^{[s-1]} \right] \|^2 \middle| \mathbf{E}^{[s-1]} \right] \right] \\ &\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \mathbb{E}_{\mathbf{E}^{(s)}} \left[\frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} \middle| \mathbf{E}^{[s-1]} \right] \|^2 \right] \end{aligned}$$

Using the definition of $\bar{\mathbf{G}}^{(s)}$,

$$= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\| \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} - \frac{1}{C} \sum_{i=1}^{N} \mathbf{E}_i^{(s)} \bar{\mathbf{G}}^{(s)} \|^2 \middle| \mathbf{E}^{[s-1]} \right] \right] + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]$$

$$= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\| \frac{1}{C} \sum_{i=1}^{N} (\mathbf{E}_i^{(s)} \mathbf{G}_i^{(s)} - \mathbf{E}_i^{(s)} \bar{\mathbf{G}}^{(s)}) \|^2 \middle| \mathbf{E}^{[s-1]} \right] \right] + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]$$

The first term contains a norm squared of a sum in the form $\|\frac{1}{N}\sum_{i=1}^{N} p_i\|^2$. We know that generally, $\|\sum_{i=1}^{N} p_i\|^2 = \sum_{i=1}^{N} \|p_i\|^2 + \sum_{i\neq j} p_i \cdot p_j$, therefore

$$\begin{aligned} &= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\frac{1}{C^2} \sum_{i=1}^{N} \| \mathbf{E}_i^{(s)} (\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}) \|^2 \middle| \mathbf{E}^{[s-1]} \right] \right] \\ &\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\mathbb{E}_{\mathbf{E}^{(s)}} \left[\frac{1}{C^2} \sum_{i\neq j} (\mathbf{E}_i^{(s)} \mathbf{E}_j^{(s)}) (\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot (\mathbf{G}_j^{(s)} - \bar{\mathbf{G}}^{(s)}) \middle| \mathbf{E}^{[s-1]} \right] \right] \\ &\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right] \end{aligned}$$

Because each $\mathbf{G}_i^{(s)}$ is dependent on all $\mathbf{E}^{[s-1]}$ but not $\mathbf{E}^{(s)}$, then both $\mathbf{G}_i^{(s)}$ and $\bar{\mathbf{G}}^{(s)}$ are constant given the conditioning on $\mathbf{E}^{[s-1]}$.

$$
\begin{aligned}
&= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\frac{1}{C^2} \sum_{i=1}^{N} \| \mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)} \|^2 \cdot \mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \middle| \mathbf{E}^{[s-1]} \right] \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\frac{1}{C^2} \sum_{i \neq j} \left((\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot (\mathbf{G}_j^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot \mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \mathbf{E}_j^{(s)} \middle| \mathbf{E}^{[s-1]} \right] \right) \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]
\end{aligned}
$$

Each $\mathbf{E}^{(s)}$ is drawn uniformly, so conditioning on $\mathbf{E}^{[s-1]}$ will not change the expectation of any $\mathbf{E}_i^{(s)}$.

$$
\begin{aligned}
&= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\frac{1}{C^2} \sum_{i=1}^{N} \| \mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)} \|^2 \cdot \mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \right] \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\frac{1}{C^2} \sum_{i \neq j} \left((\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot (\mathbf{G}_j^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot \mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \mathbf{E}_j^{(s)} \right] \right) \right] + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]
\end{aligned}
$$

By the definition of $\mathbf{E}^{(s)}$, we know that $\mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \right] = \frac{C}{N}$. Using basic probability, we can calculate that $\mathbb{E}_{\mathbf{E}^{(s)}} \left[\mathbf{E}_i^{(s)} \mathbf{E}_j^{(s)} \right] = \frac{C(C-1)}{N(N-1)}$.

$$
\begin{aligned}
&= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{C}{N} \right) \frac{1}{C^2} \sum_{i=1}^{N} \| \mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)} \|^2 \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{C(C-1)}{N(N-1)} \right) \frac{1}{C^2} \sum_{i \neq j} \left((\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}) \cdot (\mathbf{G}_j^{(s)} - \bar{\mathbf{G}}^{(s)}) \right) \right] + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]
\end{aligned}
$$

Next, we will multiply out the second term.

$$
\begin{aligned}
&= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{1}{CN} \right) \sum_{i=1}^{N} \| \mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)} \|^2 \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{C-1}{CN(N-1)} \right) \left(\sum_{i \neq j} \mathbf{G}_i^{(s)} \cdot \mathbf{G}_j^{(s)} - 2(N-1) \bar{\mathbf{G}}^{(s)} \sum_{i=1}^{N} \mathbf{G}_i^{(s)} + N(N-1) \| \bar{\mathbf{G}}^{(s)} \|^2 \right) \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]
\end{aligned}
$$

By the definition of $\bar{\mathbf{G}}^{(s)}$, we know that $\sum_{i=1}^{N} \mathbf{G}_i^{(s)} = N \bar{\mathbf{G}}^{(s)}$.

$$
\begin{aligned}
&= \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{1}{CN} \right) \sum_{i=1}^{N} \| \mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)} \|^2 \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\left(\frac{C-1}{CN(N-1)} \right) \left(\sum_{i \neq j} \mathbf{G}_i^{(s)} \cdot \mathbf{G}_j^{(s)} - 2N(N-1) \| \bar{\mathbf{G}}^{(s)} \|^2 + N(N-1) \| \bar{\mathbf{G}}^{(s)} \|^2 \right) \right] \\
&\quad + \gamma^2 \mathbb{E}_{\mathbf{E}^{[s-1]}} \left[\| \bar{\mathbf{G}}^{(s)} \|^2 \right]
\end{aligned}
$$

$$= \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}\|^2\right]$$
$$+ \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(\frac{C-1}{N-1}\Big)\Big(\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)} - N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2\Big)\right] + \gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

Next, we'll re-write $\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)} = \sum_{i=1}^{N}(\mathbf{G}_i^{(s)}\cdot\sum_{j=1,i\neq j}\mathbf{G}_j^{(s)}) = \sum_{i=1}^{N}(\mathbf{G}_i^{(s)}\cdot N(\bar{\mathbf{G}}^{(s)} - \mathbf{G}_i^{(s)}))$ for clarity.

$$= \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}\|^2\right]$$
$$+\gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(\frac{C-1}{N-1}\Big)\Big(\sum_{i=1}^{N}(\mathbf{G}_i^{(s)}\cdot N(\bar{\mathbf{G}}^{(s)})) - \sum_{i=1}^{N}(\mathbf{G}_i^{(s)}\cdot\mathbf{G}_i^{(s)}) - N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2\Big)\right]$$
$$+ \gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

$$= \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}\|^2\right]$$
$$+ \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(\frac{C-1}{N-1}\Big)\Big(N\bar{\mathbf{G}}^{(s)}(\sum_{i=1}^{N}\mathbf{G}_i^{(s)}) - \sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2 - N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2\Big)\right]$$
$$+ \gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

By the definition of $\bar{\mathbf{G}}^{(s)}$, we know that $\sum_{i=1}^{N}\mathbf{G}_i^{(s)} = N\bar{\mathbf{G}}^{(s)}$.

$$= \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}\|^2\right]$$
$$+ \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(\frac{C-1}{N-1}\Big)\Big(N^2\|\bar{\mathbf{G}}^{(s)}\|^2 - \sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2 - N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2\Big)\right]$$
$$+ \gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

$$= \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)} - \bar{\mathbf{G}}^{(s)}\|^2\right]$$
$$+ \gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(\frac{C-1}{N-1}\Big)\Big(N\|\bar{\mathbf{G}}^{(s)}\|^2 - \sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2\Big)\right] + \gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

We will then combine the first and second terms by mutating the first to be a multiple of the second.

$$\begin{aligned}&\sum_{i\ =1}^{N}\|\mathbf{G}_i^{(s)}-\bar{\mathbf{G}}^{(s)}\|^2\\&=\sum_{i=1}^{N}(\|\mathbf{G}_i^{(s)}\|^2-2\mathbf{G}_i^{(s)}\bar{\mathbf{G}}^{(s)}+\|\bar{\mathbf{G}}^{(s)}\|^2)\\&=N\|\mathbf{G}_i^{(s)}\|^2+\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-2\bar{\mathbf{G}}^{(s)}\sum_{i=1}^{N}\mathbf{G}_i^{(s)}\\&=N\|\mathbf{G}_i^{(s)}\|^2+\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-2N\|\bar{\mathbf{G}}^{(s)}\|^2\end{aligned}$$

$$\sum_{i\ =1}^{N}\|\mathbf{G}_i^{(s)}-\bar{\mathbf{G}}^{(s)}\|^2=\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-N\|\bar{\mathbf{G}}^{(s)}\|^2$$

When we substitute this value back in, we get

$$=\gamma^2\Big(\frac{1}{CN}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\Big(1-\frac{C-1}{N-1}\Big)\Big(\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-N\|\bar{\mathbf{G}}^{(s)}\|^2\Big)\right]+\gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$$

$$=\gamma^2\Big(\frac{1}{N-1}\Big)\Big(\frac{1}{C}-\frac{1}{N}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-N\|\bar{\mathbf{G}}^{(s)}\|^2\right]+\gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]\tag{25}$$

Using the previous calculation $\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)}-N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2=N\|\bar{\mathbf{G}}^{(s)}\|^2-\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2$, we can also express this as

$$=\gamma^2\Big(\frac{1}{N-1}\Big)\Big(\frac{1}{C}-\frac{1}{N}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2-\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)}\right]+\gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]\tag{26}$$

Substituting equation (26) back into equation (24), we get

$$\begin{aligned}\mathbb{E}_{\mathbf{E}^{[s]}}\left[f(\bar{\mathbf{x}}^{(s)})\right]\leq{}&\mathbb{E}_{\mathbf{E}^{[s]}}\left[f(\bar{\mathbf{x}}^{(s-1)})\right]+\mathbb{E}_{\mathbf{E}^{[s]}}\left[\langle\nabla f_i(\bar{\mathbf{x}}^{(s-1)}),\bar{\mathbf{x}}^{(s)}-\bar{\mathbf{x}}^{(s-1)}\rangle\right]\\&+\frac{L}{2}\gamma^2\Big(\frac{1}{N-1}\Big)\Big(\frac{1}{C}-\frac{1}{N}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2-\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)}\right]\\&+\frac{L}{2}\gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]\end{aligned}$$

We ran out of time to complete this proof extension, but we expect the other two terms to look similar to their corresponding terms in the original proof. Looking at the three original terms, we can say that cohorts will have no effect on the distance from our initial model to the minimum model (first term) and some effect on gradient size due to model averaging (second term). However, we expect cohorts to affect gradient variance the most. The smaller we make our cohorts, the more we might find that gradients have a high variance, especially as different clients work towards possibly different stationary points in this non-convex problem.

5.4 Interpretation

If we compare (25) and (26) to our original variance term $\frac{L}{2}\gamma^2\mathbb{E}_{\boldsymbol{\zeta}^{[t]}}\left[\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i^t\|^2\right]$ in equation (15), we see that their second term $\frac{L}{2}\gamma^2\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\|\bar{\mathbf{G}}^{(s)}\|^2\right]$ is in the same format. Therefore, (25) and (26) are essentially equal to the original term plus some new term, which we single out as the excess noise due to client batching and express in two different forms.

$$\frac{L}{2}\gamma^2\Big(\frac{1}{N-1}\Big)\Big(\frac{1}{C}-\frac{1}{N}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[\sum_{i=1}^{N}\|\mathbf{G}_i^{(s)}\|^2-N\|\bar{\mathbf{G}}^{(s)}\|^2\right] \tag{27}$$

$$=\frac{L}{2}\gamma^2\Big(\frac{1}{N-1}\Big)\Big(\frac{1}{C}-\frac{1}{N}\Big)\mathbb{E}_{\mathbf{E}^{[s-1]}}\left[N(N-1)\|\bar{\mathbf{G}}^{(s)}\|^2-\sum_{i\neq j}\mathbf{G}_i^{(s)}\cdot\mathbf{G}_j^{(s)}\right] \tag{28}$$

As we would expect, this excess noise term is zeroed out when cohort size C is equal to number of clients N, making the original model averaging a special case of our model averaging with cohorts. Looking at (27), we see that this term is also zeroed out when all $\mathbf{G}_i^{(s)}$ values are equal, which will only happens when all clients have exactly the same data. Looking at (28), we see that the dot product causes the term to be zeroed out when all gradients are parallel to each other, which again will only happen when all clients have exactly the same data. Even if this term is not zero, we can still expect it to be minimal when most gradients point in similar directions. Close to stationary points we expect similar client data to pull the model in similar directions, therefore we can say that similarity of gradient directions is a good indicator of data heterogeneity between clients.

If given more time, we would have liked to continue this proof extension to completion, then factor back in the original source of randomness due to client data $\boldsymbol{\zeta}^{[t]}$. We would also address fairness of trials within the proof the way that we did in our experiments. If we completed the proof by keeping costs constant, we would hope to find constraints on C, S, and T that limit the sum of gradients.

6 Conclusion

In this paper, we used both experimental results and theoretical abstraction to indicate that heterogeneity of data likely plays a role in determining the importance of cohort size. Our experiments generally showed an increase in the importance of cohort size as client data become more IID, though there is a lot more to be done in investigating noise and finding a method of determining optimal cohort size. The proof extension gave promising evidence of a similar relationship between cohort size and heterogeneity, but more importantly proved that gradient magnitude and direction play a part in the way that noise due to client batching affects a model's convergence rate. This work has promise as a starting point for further research on the importance of using cohorts in federated learning and finding an optimal cohort size.

7 New Skills

- A solid foundation of deep learning from my preliminary research using the Deep Learning Book [2]
- Designing machine learning models and structuring an appropriate training process
- Better Python skills with Google's TensorFlow Federated and Matplotlib
- Structuring experiments and making periodic changes as things go wrong or results are contrary to expectations
- Submitting scripts to a remote work node such as Great Lakes and properly documenting the process
- How to approach reading an unfamiliar academic paper for the first time
- De-coding the math used in academic proofs - what's implied and what needs to be explicitly pointed out
- Taking expectations when there are many random variables present
- A better general knowledge of the field of optimization
- Use of Big O notation in statistics
- Writing a research summary paper and effectively communicating results that don't fully prove a hypothesis

References

[1] J. Dean et al. "Large Scale Distributed Deep Networks". In: *NIPS*. 2012.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[3] H Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *arXiv preprint arXiv:1602.05629* (2016).

[4] Hao Yu, Sen Yang, and Shenghuo Zhu. *Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning*. 2018. arXiv: `1807.06629 [math.OC]`.