An Insight into the impact of Data Heterogeneity on the performance of the `FedAvg` algorithm

Akash Dhasade (314979) & Iris Kremer (337635) & Titouan Renard (272257)

Abstract *In this work, we present an experimental study of the behavior of Federated Averaging (`FedAvg`) algorithm on heterogeneous data. We investigate the validity and tightness of theoretical results derived in the convergence proof of the algorithm (for which we provide a walk-through in the appendix) through carefully constructing a synthetic data-set satisfying the theoretical assumptions. We then compare our results of the toy experiment to a benchmark federated dataset: the FEMNIST dataset.*

1 Introduction

Machine learning models are traditionally trained in data centers, in a setting which we refer to as centralized. This requires having the entire training dataset in one place. However, there are situations in which centrally collecting all data in is not possible, e.g. due to rising privacy concerns or huge volume of data. The federated learning (FL) setting provides an attractive alternative solution that addresses this issue, in which the training of a machine learning model is distributed over multiple clients, that each possess part of the data [5]. Each client optimizes the model on its local data and shares its gradients or updated weights with a central server, which aggregates the results from the different clients. Federated learning was introduced by McMahan et al., [1] where the authors also proposed the now popular and default algorithm of choice – the `FedAvg` algorithm.

Apart from possible privacy concerns, one of the main difficulties encountered in federated learning is dealing with the distribution of the training data across clients. Indeed, it is most often the case that the data is not independent and identically distributed (iid) across clients. This causes each client to converge towards its own local minima during training, which might drift from the global minima, depending upon the degree of data heterogeneity. In this setting, convergence is therefore more difficult to attain and good performance cannot be directly guaranteed.

The goal of this work is fourfold: (1) To study the theoretical convergence rate of the `FedAvg` algorithm extending upon the convergence proofs seen in the class (2) empirically validate and analyse the convergence result by simulating federated training on a curated toy dataset (3) compare the performance of `FedAvg` against the `FedProx` algorithm, an algorithm designed to alleviate the impact of data heterogeneity and (4) measure the performance gap of `FedAvg` on the iid and non-iid distributions of the real world FEMNIST dataset.

2 FedAVG and Convergence

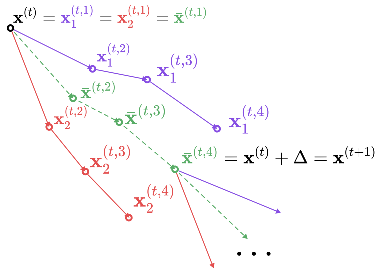
$\mathbf{x}^{(t)} = \mathbf{x}_1^{(t,1)} = \mathbf{x}_2^{(t,1)} = \bar{\mathbf{x}}^{(t,1)}$
$\mathbf{x}_1^{(t,2)}$
$\mathbf{x}_1^{(t,3)}$
$\bar{\mathbf{x}}^{(t,2)}$
$\mathbf{x}_2^{(t,2)}$
$\bar{\mathbf{x}}^{(t,3)}$
$\mathbf{x}_1^{(t,4)}$
$\mathbf{x}_2^{(t,3)}$
$\bar{\mathbf{x}}^{(t,4)} = \mathbf{x}^{(t)} + \Delta = \mathbf{x}^{(t+1)}$
$\mathbf{x}_2^{(t,4)}$
$\cdots$

Fig. 1: An illustrative diagram of `FedAvg` . The models $\bar{\mathbf{x}}^{(t,k)}$ correspond to shadow sequences of clients local models.

The trivial approach to federated optimization (introduced as a baseline in McMahan et. al., [1]) is the `FedSGD` algorithm which is the direct implementation of SGD in a distributed setting. Each client under `FedSGD` performs one training step on its local data and sends the gradients to the server, which updates the model by averaging gradients across clients before sending the model back to the clients for the next step. This approach leads to the total training time being dominated by the communication delay. `FedAvg` is introduced as a response to this problem, it lets the clients perform τ local update steps before aggregating the model, which provides a way around the communication bottleneck. As data in the federated setting may be non-iid across clients, `FedAvg` can fail when too many local steps are performed before aggregating. Still, under certain assumptions (in a convex setting), the algorithm provably converges [6] , with the following bound on the convergence rate:

$$\mathbb{E}\left[\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right] \leq \frac{D^2}{2\eta\tau T} + \frac{\eta\sigma^2}{M} + 4\tau\eta^2 L\sigma^2 + 18\tau^2\eta^2 L\zeta^2 \tag{1}$$

A walk-through the proof of this bound can be found in the appendix of this report. However, for the purpose of the experiments, the only relevant aspect is that this bound gives us a rate of convergence and an upper bound on the loss.

The righthand side of Eqn.(1) is composed of four terms, but in the case of some of our experiments, only the last one matters. Indeed, the first term approaches zero when the time T, i.e. the total number of rounds, goes to infinity, and the second and third ones are zero for our synthetic experiments, because we will only use full-batch gradient descent (more details in Section 3), meaning our variance bound σ^2 for stochastic gradients is zero. So in our experiments, as T becomes sufficiently large, we expect that the error goes to:

$$\mathbb{E}\left[\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right] \leq 18\tau^2\eta^2 L\zeta^2 \tag{2}$$

where $t \in [0, T)$ is the global round index, τ is the number of client local steps with local step index $k \in [0, \tau)$, η is the learning rate and L is the smoothness constant of our loss function.

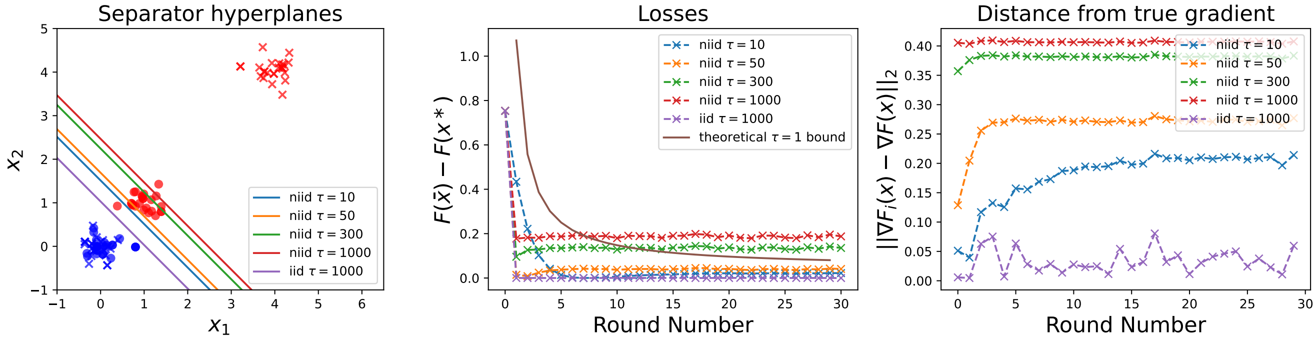

Fig. 2: Experiments with various τ performed on our toy data set. Left: separating hyperplanes. Center: loss curves. Right: dissimilarity between the local and global gradients.

Here, model $\bar{\boldsymbol{x}}^{(t,k)}$ is defined as the shadow sequence $\bar{\boldsymbol{x}}^{(t,k)} = (1/M)\sum_{i=1}^{M}\mathbf{x}_i^{(t,k)}$ with a total of M clients while $\boldsymbol{x}^*$ is the global minima as normally. An example shadow sequence is illustrated in Fig. 1. Finally, ζ is a bound on the difference between the local gradient of each client and the global gradient, defined as:

$$\max_i \sup_{\mathbf{x}} \|\nabla F_i(\mathbf{x}_i^{(t,k)}) - \nabla F(\mathbf{x}_i^{(t,k)})\| \leq \zeta \tag{3}$$

where $i \in [1, M]$ is the client index, $\nabla F_i(\mathbf{x})$ is the local gradient on client i's data and $\nabla F(\mathbf{x})$ is the global gradient. We refer the reader to appendix for an elaborate description of the parameters and assumptions, and the proof technique.

In this work, we are especially interested in the term ζ as it provides a metric on the data heterogeneity which increases when data is non-iid *i.e.* the gradient on the local data becomes a worse approximation of the gradient of the global data. In our experiments, we seek to demonstrate the slowdown in the convergence and worsening of the loss observed in correlation with increasing ζ.

Note that in real settings, there are a lot more issues than just non-iid data to be taken into account. For instance, not all clients have the same computing power, so some clients might be able to perform more local updates than others within a fixed time window. Then certain clients may have more data than others meaning that the amount of data is unbalanced across clients. Also selected clients may drop off during the training process, resulting in less model updates reaching the server. These are just a few examples of complications that arise in real applications. But in our simulations, we only consider ideal scenario *i.e.* all clients have the same amount of data, perform the same number of local steps and always return model updates to the server when selected for training.

3 Experimental Setup

We perform several experiments to investigate the effect of client-drift when using non-iid data to train a model with `FedAvg` . Some experiments are done on synthetic data, and others are performed on the more realistic FEMNIST dataset.

3.1 Synthetic data experiments

The toy setup consists of a binary classification task on linearly separable synthetic non-iid data. We generate this artificial data for the clients where each client has 50 data points each per class. The data of the first class is distributed as a gaussian of mean $(0, 0)$ and variance $(1/4, 1/4)$ for all the clients. For the second class, the first $M - 1$ clients have the data distributed as a gaussian of mean $(1, 1)$ with variance $(1/4, 1/4)$, while the last *i.e.* M-th client has the data distributed as a gaussian of mean $(4, 4)$ and variance $(1/4, 1/4)$. This last client forms our non-iid or outlier client that slows down convergence. Test data is generated the same way, but with 10 data points per class. We train a perceptron model using the binary cross entropy loss which notably yields a convex, L-smooth loss function satisfying the assumptions required to derive Eqn. (1).

Experiment 1 consists of running the `FedAvg` algorithm on the synthetic dataset with non-iid data. We use a client learning rate $\eta = 0.2$ obtained through several trials and compare the convergence for various local steps $\tau \in \{10, 50, 300, 1000\}$ values. We make the experiments easier to analyse by freezing the slope of the separator to a fixed value which we know by construction of the data-set is the optimal value, and only optimize the bias. As baseline comparison, we run an additional experiment with $\tau = 1000$ on the exact same data, but redistributed across clients in order have an iid distribution between them.

Experiment 2 aims to compare `FedAvg` with `FedProx` [4], in the same synthetic setting. `FedProx` is another federated learning algorithm in which each client's objective is regularized by a local "proximal term" μ, such that each client now optimizes:

$$\min_{\mathbf{x}} H_i(\mathbf{x}, \mathbf{x}^{(t,0)}) = F_i(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^{(t,0)}\|^2 \tag{4}$$

The introduction of the proximal term alleviates the negative impact of using non-iid data as it forces the local weights to remain closer to the global weights. We run both `FedAvg` and `FedProx` with $\tau = 3000$ local steps on our non-iid synthetic data to visualise the difference, and use $\mu = 0.1$ as the regularisation coefficient in `FedProx` .

In both experiments, `FedAvg` always samples two clients per round, one from the first $M - 1$ clients while the other always being the outlier client.

3.2 Experiment on FEMNIST

Although experimenting with synthetic data gives useful insights for understanding, it is difficult to emulate a real world task with such datasets. Therefore, it is essential to run experiments on a real dataset as well. FEMNIST was introduced as a benchmark dataset for federated learning in LEAF [2], and is the extended MNIST dataset partitioned in such a way that each client corresponds to a different writer of the character. This is a 62 class classification task where we use the sparse categorical cross-entropy as our loss. Unlike in the synthetic case, the loss function is non-convex.

Experiment 3 consists of running `FedAvg` on FEMNIST with iid partitioning, i.e. distributing the data uniformly randomly across clients, and on non-iid FEMNIST partitioned by the writer. The server samples 20 random clients per round and we set the client local steps $\tau = 25$ and $\eta = 0.02$ after tuning on a small grid of different values for these experiments. We train a simple CNN model constructed out of two convolutional layers (each followed by a max pooling layer) and two fully connected layers.

3.3 Implementation

We implemented both the algorithms in TensorFlow Federated (TFF) framework, version 0.19.0 [3] using Python 3.8. The source code

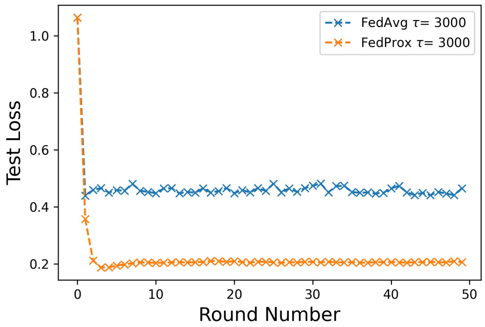

Fig. 3: Comparison of `FedAvg` and `FedProx` algorithm. Under non-iid data partitioning, the `FedProx` algorithm effectively alleviates client-drift by forcing clients to stay close to the server model.

is publicly available at `https://github.com/akash-07/optml-project`.

4 Results

4.1 Experiment 1: `FedAvg` on synthetic data

For this experiment, we plot the loss over the communication rounds on the test data for different values of τ (Fig 2. left), the final separator obtained each time (Fig. 2 center) and the maximum divergence measure between local gradients $\nabla_i F(x)$ and the global gradient $\nabla F(x)$ after each round (Fig 2. right). Observe that the bigger the τ, the faster the convergence in T as the total number of steps for a given T is higher. As τ increases we see a clear increase in the divergence of local vs. global gradients, we expect that this value goes to ζ for large τ. Furthermore, the final separators to which the algorithm converges are clearly worse when τ (and therefore client drift) gets larger. This correlates with local gradients diverging more from the global gradient. We also observe that the loss plateaus at higher values whenever ζ is higher. This is consistent with the bound of Eqn.(2), which confirms that the upper bound on the loss is increasing with ζ^2 and τ^2.

We also estimate the exact value of the bound in Eqn. (2) for each of our τ values and compare it with the experimental error results. As our synthetic experiments have $\sigma = 0$ the bound is given as a function of T by : $D^2/(2\eta\tau T) + 18\tau^2 + \eta^2 L\zeta^2$. We have set $\eta = 0.2$ and the value $\zeta = 0.407$ can be estimated empirically from the results shown in the third plot of Fig. 2. We are missing the smoothness constant L of the loss function, which we compute through numerically differentiating the loss function $L(x)$ twice and taking $L = \max \|\nabla^2 F\|_2 \approx 0.190$. D is computed by directly measuring the distance $\|\bar{x}^{(0,0)} - x^*\| \approx 2$.We observe that with these values, the practical implementation of the algorithm clearly outperforms the bound for large τ *i.e.* the loss converges to values much lower than Eqn. (2). We also include the bound as a function of the round for the $\tau = 1$ case in Fig. 2. To further illustrate the gap between theory and practice, we plot the difference between the performance $|F(\bar{x}) - F(x^*)|$ (where $\bar{x}$ is the converged loss) that we measure experimentally and the bound on converged performance $18\tau^2\eta^2L\zeta^2$ in Fig. 5. The experimental results we observe demonstrate the theoretical bound is not tight (by several order of magnitudes for large τ) and practical results are often much better. This is interesting since our synthetic data-set was built to match as much as possible the setting and assumptions required for the derivation of Eqn. (1) (see appendix).

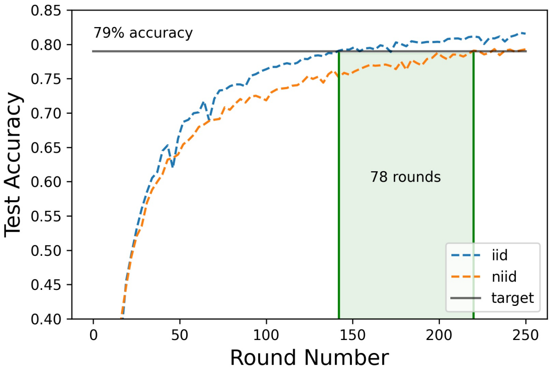

Fig. 4: Performance difference of `FedAvg` on the FEMNIST dataset under IID and non-IID data partitioning. Under heterogeneous data, `FedAvg` converges slowly while also reaching a lower convergence accuracy.

4.2 Experiment 2: `FedAvg` vs. `FedProx`

Fig. 3 shows the loss on test data over rounds obtained in the second experiment. We see that indeed, `FedProx` does not encounter the same convergence issue than `FedAvg` and reaches a much better loss thanks to the proximal term constraint on the local weights that keeps them close to the global weights. We also observe that `FedProx` takes a longer time to reach convergence (which is to be expected as the proximal term penalizes large updates).

4.3 Experiment 3: `FedAvg` on FEMNIST

The test accuracy for iid and non-iid partitioning of FEMNIST are shown in Fig. 4. Contrasting with our synthetic data results, here, we do observe a slower convergence with non-iid data as is indicated in the figure. The number of rounds to reach a target accuracy of 79% is higher by 78 rounds when using non-iid data (which amounts to a $\approx 55\%$ longer convergence time). This difference is quite significant for practical applications where each round entails significant communication costs.

Additionally, similar to the phenomenon observed in our first experiment, the model under non-iid converges to a lower final accuracy than what we observe with the iid data. The corresponds to the prediction made by the theory where non-iid settings result in a higher loss bound.

5 Conclusion

We have shown that the predictions made by the theory of `FedAvg` are matched by our experiments. We also observed that the bounds computed in the usual `FedAvg` convergence proof are not tight on simple examples. Furthermore we have shown that the behavior observed and predicted for simple smooth convex loss function translates to the realistic example of the FEMNIST dataset.

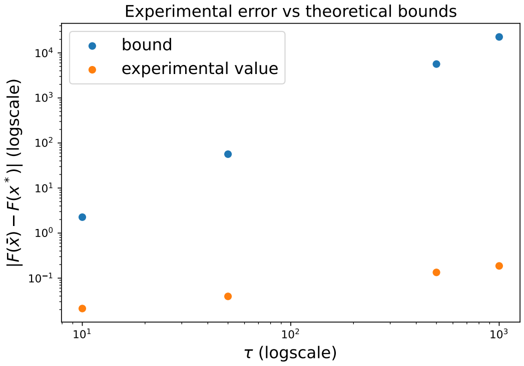

Fig. 5: Difference between the measured convergence performance of `FedAvg` and the expected performance according to theory.

References

[1]Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[2]Sebastian Caldas et al. “Leaf: A benchmark for federated settings”. In: *arXiv preprint arXiv:1812.01097* (2018).

[3]TensorFlow Federated. “Machine Learning on Decentralized Data”. In: *TensorFlow* (2019). https://www.tensorflow.org/federated.

[4]Tian Li et al. “Federated optimization in heterogeneous networks”. In: *Proceedings of Machine Learning and Systems* 2 (2020), pp. 429–450.

[5]Peter Kairouz et al. “Advances and open problems in federated learning”. In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210.

[6]Jianyu Wang et al. “A field guide to federated optimization”. In: *arXiv preprint arXiv:2107.06917* (2021).

Project report - CS-439

Appendix : A walk-through the convergence proof of the Federated Averaging (FedAVG) algorithm

Akash Dhasade (314979) & Iris Kremer (337635) & Titouan Renard (272257)

1 Setting and Algorithm

1.1 Problem definition

We consider a federated optimization problem of the form:

$$\min_x F(x),$$

where

$$F(x) = \frac{1}{M} \sum_{i=1}^{M} F_i(x).$$

The federated setting of the problem imposes that a given client i can only query the function F_i and that communications are allowed between the server and it's clients.

1.2 The Federated Averaging Algorithm

We consider the Federate Algorithm (*FedAVG*) which we formally describe below.

Algorithm 1: Federated Averaging

```
Input : initial model $\boldsymbol{x}^(0)$, learning rate $\eta$
for $t \in \{0, 1, ..., T-1\}$ do
    for $i \in \{1, ..., M\}$ clients in parallel do
        Initialize local model $\boldsymbol{x}_i^{(t,0)} \leftarrow \boldsymbol{x}^{(t)}$
        for $k \in \{0, ..., \tau\}$ do
            Compute local stochastic gradient $g_i(\boldsymbol{x}^{(t,k)})$
            Compute local step $\boldsymbol{x}_i^{(t,k+1)} \leftarrow \boldsymbol{x}_i^{(t,k)} - \eta g_i(\boldsymbol{x}^{(t,k)})$
        Compute local change over round $\Delta_i^{(t)} \leftarrow \boldsymbol{x}_i^{(t,\tau)} - \boldsymbol{x}_i^{(t,0)}$
    Average local updates $\Delta^{(t)} = \frac{1}{M} \cdot \sum_{i=1}^{M} \Delta_i^{(t)}$
    Update global model $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} + \Delta^{(t)}$
Return : $\boldsymbol{x}^{(t)}$
```

2 Analysis of the Algorithm

2.1 Setting and assumptions

As a preliminary step to the analysis of algorithm 1, we make the following 7 assumptions:

1. At any round t each *client* takes $\tau \in \mathbb{N}$ local SGD steps with constant learning rate η (which we denote as $\boldsymbol{x}_i^{(t,k+1)} \leftarrow \boldsymbol{x}_i^{(t,k)} - \eta g_i(\boldsymbol{x}_i^{(t,k)})$ with g_i is one draw of the stochastic gradient of F_i and $k \in [0, \tau)$.
2. The *server step* is computed as $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} + \Delta^{(t)}$.
3. There are (M) clients labelled $i \in \{0, 1, ..., M\}$ and each client contributes a uniform share of the global objective $F(\boldsymbol{x}) = \frac{1}{M} \sum_{i=1}^{M} F_i(\boldsymbol{x})$.
4. Each clients takes part in every round.
5. Each local objective F_i is convex and L-smooth.
6. Each client queries an unbiased stochastic gradient with σ^2-uniformly bounded variance in l_2 norm, i.e.

$$\mathbb{E}[g_i(\boldsymbol{x}_i^{(t,k)}) | \boldsymbol{x}_i^{(t,k)}] = \nabla F_i(\boldsymbol{x}_i^{(t,j)}), \tag{1}$$

$$\mathbb{E}[\| g_i(\boldsymbol{x}_i^{(t,k)}) - F_i(\boldsymbol{x}_i^{(t,j)}) \|^2 | \boldsymbol{x}_i^{(t,k)}] \leq \sigma^2. \tag{2}$$

7. The difference of local gradient $\nabla F_i(\boldsymbol{x})$ and the global gradient $\nabla F(\boldsymbol{x})$ is ζ-uniformly bounded in l_2 norm, i.e.

$$\max_i \sum_{\boldsymbol{x}} \| \nabla F_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x}) \| \leq \zeta. \tag{3}$$

First, let us define the shadow sequence which we will use to make the notation a bit more readable as we go through the proof:

Notation 2.1. *(Shadow sequence) We call the sequence described by $\bar{x}^{t,k} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{x}_i^{(t,k)}$ the shadow sequence.*

As we often do in the optimization literature we will try to show a result of the form:

$$\mathbb{E}\left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*) \right] = O\left(\frac{1}{\tau T} \right).$$

Which one can read as *as we progress, in expectation we are guaranteed to have an error that goes to some small constant*. Showing that this results is true amounts to finding a relevant upper bound decreasing in $\frac{1}{\tau T}$. We split our proving effort in two steps (which correspond to two lemmas):

1. We are making progress in each round $\mathbb{E}\left[\frac{1}{\tau} \sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*) \right]$ is bounded by some term decreasing when t increases.
2. All client iterates remain close to the global average (the shadow sequence), i.e. $\|\boldsymbol{x}_i^{(t,k)} - \hat{\boldsymbol{x}}^{(t,k)}\|_{l_2}$ is bounded in expectation.

Formally we will write our proof using one theorem that relies on two lemmas (showing both properties discussed above). The formal proofs are detailed in the next section.

2.2 Convergence of FedAVG

We will prove the following results.

Theorem 2.1. *(Convergence for Convex Functions) under the assumptions and assuming $\eta \leq \frac{1}{4L}$ one has:*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*) \right] &\leq \frac{D^2}{2\eta\tau T} + \frac{\eta\sigma^2}{M} + 4\tau\eta^2 L\sigma^2 + 18\tau^2\eta^2 L\zeta^2 \\ &= O\left(\frac{1}{\tau T} \right) \end{aligned} \tag{4}$$

To which we will get using two lemmas:

Lemma 2.2. *(Per round progress) Assuming $\eta \leq \frac{1}{4L}$, for one round t of the algorithm, one has:*

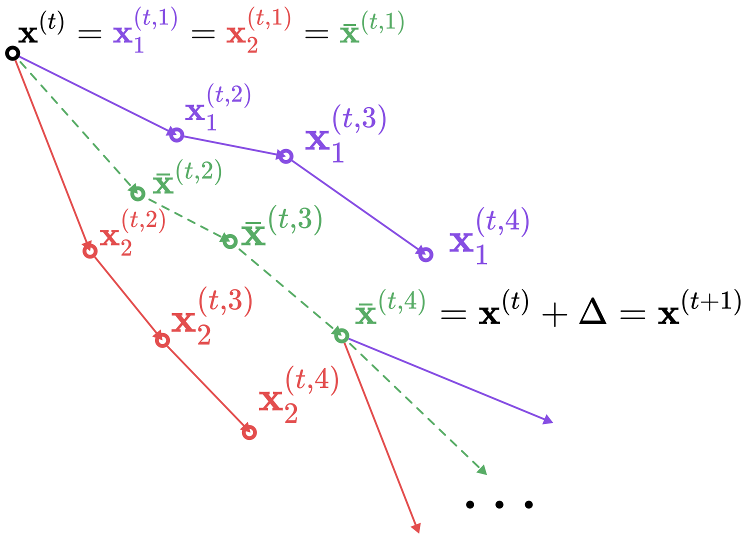
$\mathbf{x}^{(t)} = \mathbf{x}_1^{(t,1)} = \mathbf{x}_2^{(t,1)} = \bar{\mathbf{x}}^{(t,1)}$
$\mathbf{x}_1^{(t,2)}$
$\mathbf{x}_1^{(t,3)}$
$\bar{\mathbf{x}}^{(t,2)}$
$\mathbf{x}_2^{(t,2)}$
$\bar{\mathbf{x}}^{(t,3)}$
$\mathbf{x}_1^{(t,4)}$
$\mathbf{x}_2^{(t,3)}$
$\bar{\mathbf{x}}^{(t,4)} = \mathbf{x}^{(t)} + \Delta = \mathbf{x}^{(t+1)}$
$\mathbf{x}_2^{(t,4)}$
$\dots$

Fig. 1: Illustration of the progression of one 4 step round of algorithm 1 with the shadow sequence represented on green.

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{\tau}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right] \\ &\leq \frac{1}{2\eta\tau}\left(\|\bar{\boldsymbol{x}}^{(t,0)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,\tau)} - \boldsymbol{x}^*\|^2 | \mathcal{F}^{(t,0)}\right]\right) \\ &+ \frac{\eta\sigma^2}{M} + \overbrace{\frac{1}{M\tau}\sum_{i=1}^{M}\sum_{k=1}^{\tau-1}\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(0,t)}\right]}^{\textit{client drift}} \end{aligned} \tag{5}$$

Lemma 2.3. *(Bounded client drift) Assuming $\eta \leq \frac{1}{4L}$, for one round t of the algorithm, one has:*

$$\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(0,t)}\right] \leq 18\tau^2\eta^2\zeta^2 + 4\tau\eta^2\sigma^2 \tag{6}$$

*Proving Theorem 2.1 (convergence of FedAVG)*Most of the technical work will lie in proving the two lemmas, but first we will focus on proving theorem 2.1, while assuming that lemmas 2.2 and 2.3 are true.

Proof: (Of Theorem 2.1.) We want to find a bound for the quantity

$$\mathbb{E}\left[\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right],$$

to do so we will use the bound on $\mathbb{E}\left[\frac{1}{\tau}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right]$ which is given by lemma 2.2. First, let's write out the sum on which we will take the expectation and express it as a function of the per round client progress which we bounded in lemma 2.2:

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right], \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\overbrace{\mathbb{E}\left[\frac{1}{\tau}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right]}^{(\triangledown)}\right]. \end{aligned}$$

Observing that the term $(\triangledown)$ is the left side of the inequality (5) of lemma 2.2, we use the lemma to bound our expectation. Using linearity of expectation we split this expression in three different terms which we will then discuss separately.

$$\begin{aligned}
(\triangledown) \leq \mathbb{E}\Bigg[& \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{2\eta\tau}\left(\|\bar{\boldsymbol{x}}^{(t,0)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,\tau)} - \boldsymbol{x}^*\|^2 | \mathcal{F}^{(t,0)}\right]\right) \\
& + \frac{\eta\sigma^2}{M} + \frac{1}{M\tau}\sum_{i=1}^{M}\sum_{k=1}^{\tau-1}\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(0,t)}\right]\Bigg] \\
= & \overbrace{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2\eta\tau}\left(\|\bar{\boldsymbol{x}}^{(t,0)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,\tau)} - \boldsymbol{x}^*\|^2 | \mathcal{F}^{(t,0)}\right]\right)\right]}^{(\bigstar)} \\
& + \overbrace{\frac{\eta\sigma^2}{M}}^{(\diamond)} + \overbrace{\frac{1}{M\tau}\sum_{i=1}^{M}\sum_{k=1}^{\tau-1}\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(0,t)}\right]}^{(\dagger)}
\end{aligned}$$

Let us now consider the three terms. Terms $(\diamond)$ and $(\dagger)$ gives a bound on individual client drift (i.e. how far do the clients get from the shadow sequence), term $(\bigstar)$ gives a bound on the global progression. Here our goal is to show that $(\diamond)$ and $(\dagger)$ can be arbitrarily bounded as a function of the algorithm's parameters and that $(\bigstar)$ goes to 0 with $T \cdot \tau$. We now discuss bounds for every single term.

1. Term $(\diamond)$ is already a function of our algorithm's parameters, there is nothing to show here.
2. Now we consider term $(\dagger)$, this is the drift term:

$$\frac{1}{M\tau}\sum_{i=1}^{M}\sum_{k=1}^{\tau-1}\overbrace{\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(0,t)}\right]}^{(\spadesuit)}$$

it is a sum over term $(\spadesuit)$, which is the left side the inequality (6) of lemma 2.3. We plug the right side of (6) and as it is not a function of the sum variables we can drop the sums as well.

$$(\dagger) \leq 18\tau^2\eta^2\zeta^2 + 4\tau\eta^2\sigma^2$$

3. Finally we consider term $(\bigstar)$,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2\eta\tau}\left(\|\bar{\boldsymbol{x}}^{(t,0)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,\tau)} - \boldsymbol{x}^*\|^2 | \mathcal{F}^{(t,0)}\right]\right)\right],$$

here we have to use a few tricks. First using that expectation is linear we will separate our terms and using that $\mathbb{E}\left[\mathbb{E}[x]\right] = \mathbb{E}[x]$ we will drop the double expectation in the sum:

$$\frac{1}{2\eta\tau T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|\overbrace{\bar{\boldsymbol{x}}^{(t,0)}}^{\bar{\boldsymbol{x}}^{(t)}} - \boldsymbol{x}^*\|^2\right] - \mathbb{E}\left[\|\overbrace{\bar{\boldsymbol{x}}^{(t,\tau)}}^{\bar{\boldsymbol{x}}^{(t+1)}} - \boldsymbol{x}^*\|^2\right]\right).$$

At this point we first observe (as denoted above) that by definition of the algorithm we have $\bar{\boldsymbol{x}}^{(t+1)} = \bar{\boldsymbol{x}}^{(t+1,0)} = \bar{\boldsymbol{x}}^{(t,\tau)}$ and $\bar{\boldsymbol{x}}^{(t)} = \bar{\boldsymbol{x}}^{(t,0)}$, then for readability we make use of the notation $d(\boldsymbol{x}) = \mathbb{E}\left[\|\boldsymbol{x} - \boldsymbol{x}^*\|^2\right]$ and write out the sum over t:

$$\begin{aligned}
& \sum_{t=0}^{T-1}\left(d(\boldsymbol{x}^{(t)}) - d(\boldsymbol{x}^{(t+1)})\right) \\
& = d(\boldsymbol{x}^{(0)}) - d(\boldsymbol{x}^{(1)}) + d(\boldsymbol{x}^{(1)}) - d(\boldsymbol{x}^{(2)}) \\
& + \cdots + d(\boldsymbol{x}^{(T-2)}) - d(\boldsymbol{x}^{(T-1)}) + d(\boldsymbol{x}^{(T-1)}) - d(\boldsymbol{x}^{(T)}).
\end{aligned}$$

Observing that the terms cancel out we get the following expression:

$$\begin{aligned}
(\bigstar) = & \frac{1}{2\eta\tau T}\Big(d(\boldsymbol{x}^{(0)}) - d(\boldsymbol{x}^{(1)}) + d(\boldsymbol{x}^{(1)}) - d(\boldsymbol{x}^{(2)}) \\
& + \cdots + d(\boldsymbol{x}^{(T-2)}) - d(\boldsymbol{x}^{(T-1)}) + d(\boldsymbol{x}^{(T-1)}) - d(\boldsymbol{x}^{(T)})\Big) \\
= & \frac{1}{2\eta\tau T}\left(d(\boldsymbol{x}^{(0)}) - d(\boldsymbol{x}^{(T)})\right) \leq \frac{d(\boldsymbol{x}^{(0)})}{2\eta\tau T} = \frac{D^2}{2\eta\tau T}.
\end{aligned}$$

Where $d(\boldsymbol{x}^{(0)}) = \mathbb{E}[\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2] = \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2 = D^2$ with D the diameter to the global opt at the beginning of the gradient descent (this is a constant).

Putting all bounds back together we get the following convergence bound (which concludes the proof):

$$\mathbb{E}\left[\frac{1}{\tau T}\sum_{t=0}^{T-1}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right] \leq \overbrace{\frac{D^2}{2\eta\tau T}}^{(\bigstar)} + \overbrace{\frac{\eta\sigma^2}{M}}^{(\diamond)} + \overbrace{4\tau\eta^2 L\sigma^2 + 18\tau^2\eta^2 L\zeta^2}^{(\dagger)}$$

□

We now get in the (somewhat) more technical parts of the proof, proving that both the lemmas are correct.

Proving Lemma 2.2 (per round progression)

Proof: (Of Lemma 2.2.)
Let's start with the per round progress lemma, we want to bound the quantity :

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{k=1}^{\tau} F(\bar{\boldsymbol{x}}^{(t,k)}) - F(\boldsymbol{x}^*)\right],$$

in $O(\frac{1}{\tau})$. Similarly to what we just did for the theorem (and to how most of these convergence proofs are computed), we will to try to bound a single term of the sum (in expectation) by the previous term and then we will telescope the sum to get a serviceable bound. In other words we are trying to bound the expectation:

$$(\dagger) = \mathbb{E}\Big[\overbrace{F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)}^{(\heartsuit)}|\mathcal{F}^{(t,k)}\Big], \tag{7}$$

In order to get a bound on the expectation $(\dagger)$, most of the work we will have to do will lie in understanding the $(\heartsuit)$ term on which we take the expectation. *This is where the proof starts getting technical*. Writing it out we can first split it between the separate client objective functions:

$$(\heartsuit) = F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*) = \frac{1}{M}\sum_{i=1}^{M}\Big(F_i(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)\Big)\,. \tag{8}$$

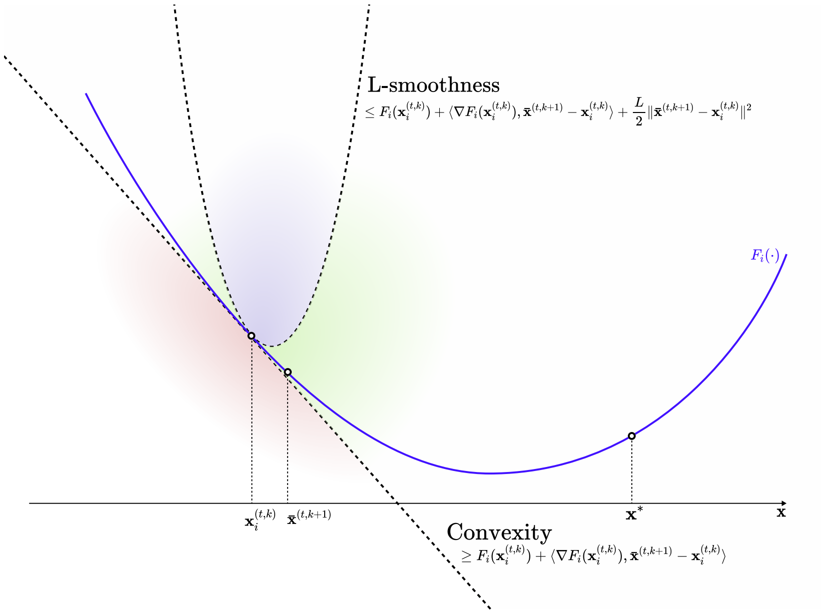
L-smoothness
$\leq F_i(\mathbf{x}_i^{(t,k)}) + \langle\nabla F_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)}\rangle + \frac{L}{2}\|\bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)}\|^2$
$F_i(\cdot)$
$\mathbf{x}_i^{(t,k)}$
$\bar{\mathbf{x}}^{(t,k+1)}$
$\mathbf{x}^*$
$\mathbf{x}$
Convexity
$\geq F_i(\mathbf{x}_i^{(t,k)}) + \langle\nabla F_i(\mathbf{x}_i^{(t,k)}), \bar{\mathbf{x}}^{(t,k+1)} - \mathbf{x}_i^{(t,k)}\rangle$

Fig. 2: Illustration of the L-smooth/convex properties of F_i. Convexity implies that the function is above the half space tangent to the graph at any point (in red) and that is is below the parabola with parameter L (in blue). In other words, the function being L-smooth and convex is necessarily in the green area of the plot. Furthermore observe that the global minimum $\boldsymbol{x}^*$ of F is not the minimum of F_i, so we cannot use that property directly.

Looking at the expression above, we clearly see that bounding ($\heartsuit$) will amount to bound $F_i(\bar{\boldsymbol{x}}^{(t,k+1)})$, which we will do using the smoothness (property 2.3) and convexity (property 2.2) properties from the assumptions made in section 2.1. First, we get an upper bound from the L-smoothness (property 2.3) of F_i:

$$F_i(\bar{\boldsymbol{x}}^{(t,k+1)}) \leq F_i(\boldsymbol{x}) + \langle \nabla F_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x} \rangle + \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x}.$$

This inequality is true for all $\boldsymbol{x} \in \boldsymbol{dom}(f)$ but we care about the algorithm's step and for each $F_i(\bar{\boldsymbol{x}}^{(t,k+1)})$ the step of the algorithm depends of $\boldsymbol{x}_i^{(t,k)}$ (the parameters of client i at the previous time step). Using observation 2.3 (which we can do as F_i is convex) we further bound the ($\spadesuit$) term as a function of $\boldsymbol{x}^*$ instead of $\boldsymbol{x}_i^{(t,k)}$ which we don't know.

$$\begin{aligned} F_i(\bar{\boldsymbol{x}}^{(t,k+1)}) &\leq \overbrace{F_i(\boldsymbol{x}_i^{(t,k)}) + \langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}_i^{(t,k)} \rangle}^{(\spadesuit)} + \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}_i^{(t,k)}\|^2 \\ &\leq F_i(\boldsymbol{x}^*) + \langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^* \rangle + \overbrace{\frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}_i^{(t,k)}\|^2}^{\square} \end{aligned}$$

Now in order to get a bound on the ($\square$) term we will try to split the $\boldsymbol{x}_i^{(t,k)}$ variable and express our bound as a function of the pre step progress and the distance between clients (which we will bound using the client drift lemma later). To do so we use the triangle inequality ($\|a+b\| \leq \|a\| + \|b\|$):

$$\begin{aligned} \overbrace{\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}_i^{(t,k)}}^{a+b} &= \overbrace{\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}}^{a} + \overbrace{\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}}^{b} \\ \square &= \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}_i^{(t,k)}\|^2 \\ &\leq \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 \\ &\leq L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 \end{aligned}$$

The last step is just done for readability, we could actually get a tighter bound by keeping the factor two in there. Plugging this back in the $F_i(\bar{\boldsymbol{x}}^{(t,k+1)})$ bound we get the following inequality:

$$\begin{aligned} F_i(\bar{\boldsymbol{x}}^{(t,k+1)}) \leq F_i(\boldsymbol{x}^*) &+ \langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^* \rangle \\ &+ L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 \end{aligned}$$

Let's plug what we just derived in equation (8):

$$\begin{aligned} (\heartsuit) = F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*) &= \frac{1}{M}\sum_{i=1}^{M}\Big(F_i(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)\Big) \\ &\leq \frac{1}{M}\sum_{i=1}^{M}\Big(F_i(\boldsymbol{x}^*) + \langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^* \rangle \\ &+ L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 - F(\boldsymbol{x}^*)\Big). \end{aligned}$$

Using the linearity of the sum we can split this expression into separate terms:

$$
\begin{aligned}
(\heartsuit) \le\ & \overbrace{\frac{1}{M}\sum_{i=1}^{M} F_i(\boldsymbol{x}^*) - F(\boldsymbol{x}^*)}^{=0} + \frac{1}{M}\sum_{i=1}^{M}\langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle \\
& + L\frac{1}{M}\sum_{i=1}^{M}\overbrace{\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2}^{\text{not a function of } i} + L\frac{1}{M}\sum_{i=1}^{M} \|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 \\
& = \overbrace{\frac{1}{M}\sum_{i=1}^{M}\langle \nabla F_i(\boldsymbol{x}_i^{(t,k)}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle}^{(\blacklozenge)} \\
& + L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + \overbrace{L\frac{1}{M}\sum_{i=1}^{M} \|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2}^{\blacksquare}
\end{aligned}
$$

This brings us closer to bounding the $(\heartsuit)$ term, remembering that we are working with values inside of an expectation, we clearly see that term $(\blacksquare)$ is a measure of client drift (which we will bound later with the client drift lemma). Taking a look at the sum the term that requires and explanation is the $(\blacklozenge)$ term, which we will now take a look at. Recall that the FedAVG algorithm queries stochastic gradients g_i, which are unbiased, variance bounded estimators of the true gradient ∇F_i. Observe that we can write out the local true client gradient as:

$$
\nabla F_i(\boldsymbol{x}) = \overbrace{g_i(\boldsymbol{x})}^{(\triangleleft)} + \overbrace{(F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}))}^{(\triangleright)}, \tag{9}
$$

where $(\triangleleft)$ represents the sampled gradient of client objective F_i (which is used for gradient descent steps) and $(\triangleright)$ the deviation from the true gradient value value. This is useful as we know from the assumptions that the deviation is 0 in expectation (g_i is an unbiased estimator of ∇F_i). Using the decomposition (9) we can develop the $(\heartsuit)$ bound as follows:

$$
\begin{aligned}
(\heartsuit) \le\ & \frac{1}{M}\sum_{i=1}^{M}\langle g_i(\boldsymbol{x}) + (F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}))\,, \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle \\
& + L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\frac{1}{M}\sum_{i=1}^{M} \|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2 \\
= & \overbrace{\frac{1}{M}\sum_{i=1}^{M}\langle g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle}^{(\blacktriangleleft)} + \overbrace{\frac{1}{M}\sum_{i=1}^{M}\langle F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle}^{(\blacktriangleright)} \\
& + L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\frac{1}{M}\sum_{i=1}^{M} \|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2
\end{aligned}
$$

We will analyze $(\blacktriangleright)$ later, when we take expectations, let us now focus on $(\blacktriangleleft)$. Using that client algorithm steps are computed as $\boldsymbol{x}_i^{(t,k+1)} = \boldsymbol{x}_i^{(t,k)} - \eta g_i(\boldsymbol{x}_i^{(t,k)})$, and that the inner product is distributive over addition we can develop $(\blacktriangleleft)$ as follows:

$$
\begin{aligned}
(\blacktriangleleft) &= \frac{1}{M}\sum_{i=1}^{M}\langle g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle = \langle \frac{1}{M}\sum_{i=1}^{M} g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle \\
&= \left\langle \frac{\bar{\boldsymbol{x}}^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k+1)}}{\eta}, \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^* \right\rangle = \frac{1}{\eta}\langle \bar{\boldsymbol{x}}^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k+1)}, \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle.
\end{aligned}
$$

Which we can convert into a sum of norms using the polarization identity (property 2.4):

$$
\begin{aligned}
(\blacktriangleleft) &= \frac{1}{\eta}\langle \bar{\boldsymbol{x}}^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k+1)}, \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle \\
&= \frac{1}{2\eta}\left(\|\bar{\boldsymbol{x}}^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k+1)} + \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\right) \\
&= \frac{1}{2\eta}\left(\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\right)
\end{aligned}
$$

The $(\blacktriangleleft)$ term now looks a lot like a sum which we will be able to telescope (which is what we intended). Let us plug it back into $(\heartsuit)$:

$$\begin{aligned}
(\heartsuit) &= F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*) \\
&\le \overbrace{\frac{1}{2\eta}\Big(\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 - \|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\Big)}^{(\blacktriangleleft)} \\
&\quad + \overbrace{\frac{1}{M}\sum_{i=1}^{M}\langle F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle}^{(\blacktriangleright)} \\
&\quad + L\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 + L\frac{1}{M}\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2,
\end{aligned}$$

and take expectations (we use the linearity of expectation to separate the terms and make the expression more readable):

$$\begin{aligned}
(\dagger) &= \mathbb{E}\Big[\overbrace{F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)}^{(\heartsuit)}\,|\mathcal{F}^{(t,k)}\Big] \\
&\le \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2\right] - \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] \\
&\quad -\frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\right] + \overbrace{\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle\right]}^{(\mathbb{E}[\blacktriangleright])} \\
&\quad +L\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] + \frac{L}{M}\mathbb{E}\left[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\right]
\end{aligned}$$

At this point we need to find a way of showing that the $\mathbb{E}[\blacktriangleright]$ expectation is bounded. To do so we proceed as follows, observing that since g_i is an unbiased estimator of ∇F_i we have that $\mathbb{E}[\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle] = 0$, which allows us to swap the term of the right side of the inner product as we please:

$$\begin{aligned}
\mathbb{E}[\blacktriangleright] &= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle\right] \\
&= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\rangle\right].
\end{aligned}$$

At this point we make use of Young's inequality (property 2.5):

$$\begin{aligned}
&= \frac{1}{M^2}\mathbb{E}\left[\frac{(\sqrt{2\eta})^2}{2}\sum_{i=1}^{M}\|\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2\right] + \mathbb{E}\left[\frac{1}{2(\sqrt{2\eta})^2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] \\
&= \frac{\eta}{M^2}\overbrace{\mathbb{E}\left[\sum_{i=1}^{M}\|\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2\right]}^{\le M\sigma^2 \text{ by assumption 6}} + \frac{1}{4\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] \\
&\le \frac{\eta\sigma^2}{M} + \frac{1}{4\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right].
\end{aligned}$$

Bringing it all together we have:

$$
\begin{aligned}
(\dagger) &= \mathbb{E}\Big[\overbrace{F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)}^{(\heartsuit)}\,|\mathcal{F}^{(t,k)}\Big] \\
&\leq \frac{1}{2\eta}\,\overbrace{\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2\Bigg]}^{=\|\bar{\boldsymbol{x}}^{(t,k)}-\boldsymbol{x}^*\|^2} - \frac{1}{2\eta}\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Bigg] \\
&\quad - \frac{1}{2\eta}\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\Bigg] + \overbrace{\frac{\eta\sigma^2}{M} + \frac{1}{4\eta}\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Bigg]}^{(\mathbb{E}[\blacktriangleright])} \\
&\quad + L\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Bigg] + \frac{L}{M}\mathbb{E}\Bigg[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\Bigg].
\end{aligned}
$$

reorganizing the sum we get to :

$$
\begin{aligned}
(\dagger) &\leq \frac{1}{2\eta}\left(\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\Big[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\Big]\right) \\
&\quad + \overbrace{\left(L - \frac{1}{4\eta}\right)}^{=L-\frac{1}{2\eta}+\frac{1}{4\eta}}\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Bigg] \\
&\quad + \frac{\eta\sigma^2}{M} + \frac{L}{M}\mathbb{E}\Bigg[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\Bigg].
\end{aligned}
$$

Finally, observing that if $\eta \leq \frac{1}{4L}$ the $\mathbb{E}\Bigg[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Bigg]$ term can be dropped from the sum we get a lighter, less tight expression of our bound:

$$
\begin{aligned}
(\dagger) &= \mathbb{E}\Big[F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)|\mathcal{F}^{(t,k)}\Big] \\
&\leq \frac{1}{2\eta}\left(\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2 - \mathbb{E}\Big[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\Big]\right) \\
&\quad + \frac{\eta\sigma^2}{M} + \frac{L}{M}\mathbb{E}\Bigg[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\Bigg].
\end{aligned}
$$

Similarly to what we did in the case of the proof of theorem 2.1 we get to the bound of the lemma by telescoping this sum across all client steps $k = 1, ..., \tau$. We omit the details as they are identical to the ones preformed in the previous proof, just note that the terms

$$
\frac{\eta\sigma^2}{M} \qquad \text{and} \qquad \mathbb{E}\Bigg[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\Bigg]
$$

are simply summed across all steps and are not telescoped. □

Proving Lemma 2.3 (client drift)

Proof: (Of Lemma 2.3.)
The second lemma is a bit less technical. Here the main idea is to look at the expected drifts between any two client (we pick clients 1 and 2 *wlog*), find an upper bound for that drift and then to use that to bound the difference between the average progression and any specific client i. Again we will try to find an expression that we can telescope across the algorithm's steps. Starting with both our clients at the same point

$\boldsymbol{x}^{(t)} = \boldsymbol{x}_1^{(t,0)} = \boldsymbol{x}_2^{(t,0)}$, we will look at the expected drift for a given client step k to $k+1$.

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}_1^{(t,k+1)} - \boldsymbol{x}_2^{(t,k+1)}\|^2 | \mathcal{F}^{(t,k)}\right] \\
&= \mathbb{E}\left[\overbrace{\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)} - \eta\left(g_1(\boldsymbol{x}_1^{(t,k)}) - g_2(\boldsymbol{x}_2^{(t,k)})\right)\|^2 | \mathcal{F}^{(t,k)}}^{\text{By definition of the gradient step}}\right] \\
&= \|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 - 2\eta\langle g_1(\boldsymbol{x}_1^{(t,k)}) - g_2(\boldsymbol{x}_2^{(t,k)})\rangle \\
&\quad + \eta^2\|g_1(\boldsymbol{x}_1^{(t,k)}) - g_2(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&\leq \|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 - 2\eta\overbrace{\langle \nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)}), \boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\rangle}^{(\Diamond)} \\
&\quad + \eta^2\overbrace{\|\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)})\|^2}^{(\triangle)} + 2\eta^2\sigma^2
\end{aligned}$$

Where the last step is computed by using assumption 6 of the algorithm (*each client queries an unbiased stochastic gradient with σ^2-uniformly bounded variance in l_2 norm*). Using assumption 7 (*the difference of local gradient $\nabla F_i(\boldsymbol{x})$ and the global gradient $\nabla F(\boldsymbol{x})$ is ζ-uniformly bounded in l_2 norm.*) we can easily bound the $\|\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)})\|^2$ term by $\|\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)})\|^2 \leq \zeta^2$. We care about bound the $(\triangle)$ and $(\Diamond)$ terms. Let us start with the $(\Diamond)$ term:

$$\begin{aligned}
-(\Diamond) &= -\langle \nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)}), \boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\rangle \\
&\leq -\langle \nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)}), \boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\rangle \\
&\quad + 2\zeta\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\| && \text{by assumption 7} \\
&\leq -\frac{1}{L}\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&\quad + 2\zeta\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\| && \text{by L-smoothness} \\
&\leq -\frac{1}{L}\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&\quad + \frac{1}{2\eta\tau}\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\| + 2\eta\tau\zeta^2 && \text{by AM-GM inequality.}
\end{aligned}$$

Now for the $(\triangle)$ term:

$$\begin{aligned}
&\|\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&= \|\nabla F(\boldsymbol{x}_1^{(t,k)}) + (\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_1^{(t,k)})) \\
&\quad - \nabla F_2(\boldsymbol{x}_2^{(t,k)}) - (\nabla F_2(\boldsymbol{x}_2^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)}))\|^2 \\
&= \|(\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F_2(\boldsymbol{x}_2^{(t,k)})) + (\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})) \\
&\quad - (\nabla F_2(\boldsymbol{x}_2^{(t,k)}) - \nabla F(\boldsymbol{x}_1^{(t,k)}))\|^2 \\
&\leq \|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&\quad + \|\nabla F_1(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2 \\
&\quad + \|\nabla F_2(\boldsymbol{x}_2^{(t,k)}) - \nabla F(\boldsymbol{x}_1^{(t,k)})\|^2 \\
&\leq 3\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2 + 6\zeta^2.
\end{aligned}$$

Plugging those bounds in to the expected per-round drift we get:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}_1^{(t,k+1)} - \boldsymbol{x}_2^{(t,k+1)}\|^2 | \mathcal{F}^{(t,k)}\right] \\
&\leq \|\boldsymbol{x}_1^{(t,k)} + \boldsymbol{x}_2^{(t,k)}\|^2 \\
&- \overbrace{\frac{2\eta}{L}}^{L\leq\frac{1}{4\eta}\Rightarrow\geq 8\eta^2} \overbrace{\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2}^{\geq 0} + \frac{1}{\tau}\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + 4\eta^2\tau\zeta^2 \\
&+3\eta^2 \overbrace{\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2}^{\geq 0} + 6\eta^2\zeta + 2\eta^2\sigma^2,
\end{aligned}$$

which we rearrange to show that the gradient terms can be removed by loosening the bound:

$$\begin{aligned}
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + \overbrace{(-8\eta^2 + 3\eta^2)\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2}^{\leq 0} \\
&+4\eta^2\tau\zeta^2 + 6\eta^2\zeta + 2\eta^2\sigma^2 \\
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + 4\eta^2\tau\zeta^2 + 6\eta^2\zeta + 2\eta^2\sigma^2 \\
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + 10\eta^2\tau\zeta^2 + 2\eta^2\sigma^2.
\end{aligned}$$

Telescoping the per-step drift over one round gives us:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}_1^{(t,k+1)} - \boldsymbol{x}_2^{(t,k+1)}\|^2 | \mathcal{F}^{(t,k)}\right] \\
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + \overbrace{(-8\eta^2 + 3\eta^2)\|\nabla F(\boldsymbol{x}_1^{(t,k)}) - \nabla F(\boldsymbol{x}_2^{(t,k)})\|^2}^{\leq 0} \\
&+4\eta^2\tau\zeta^2 + 6\eta^2\zeta + 2\eta^2\sigma^2 \\
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + 4\eta^2\tau\zeta^2 + 6\eta^2\zeta + 2\eta^2\sigma^2 \\
&\leq (1 + \frac{1}{\tau})\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 + 10\eta^2\tau\zeta^2 + 2\eta^2\sigma^2.
\end{aligned}$$

Since $\|\boldsymbol{x}_1^{(t,0)} - \boldsymbol{x}_2^{(t,0)}\|^2 = 0$ we have that the drift is bounded by a geometric series as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{x}_1^{(t,k)} - \boldsymbol{x}_2^{(t,k)}\|^2 | \mathcal{F}^{(t,0)}\right] \\
&\leq (10\eta^2\tau\zeta^2 + 2\eta^2\sigma^2)\left(k + \sum_{t=2}^{k}(1 + \frac{1}{\tau})^k\right) \\
&\leq 18\eta^2\tau^2\zeta^2 + 4\eta^2\tau\sigma^2.
\end{aligned}$$

And by convexity this result generalizes to each client w.r.t the shadow sequence:

$$\mathbb{E}\left[\|\boldsymbol{x}_i^{(t,k)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2 | \mathcal{F}^{(t,0)}\right] \leq 18\eta^2\tau^2\zeta^2 + 4\eta^2\tau\sigma^2.$$

□

2.3 Properties used for the proof

In the following subsection, we list the different properties and lemmas that are useful to the proofs of the convergence of algorithm 1. First, we consider convexity and smoothness lemmas:

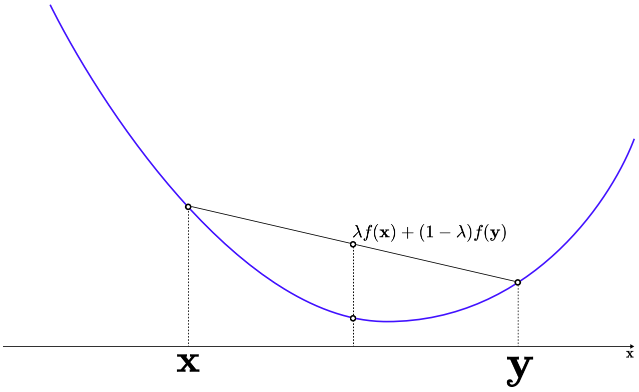

Fig. 3: Visual illustration of convexity.

Property 2.1. *(Convexity) consider a real value function $f : \boldsymbol{dom}(f) \to \mathbb{R}$, we say that f if*

1. $\boldsymbol{dom}(f)$ *is convex*
2. $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f)$ *we have:*

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

Property 2.2. *(First order characterization of convexity) suppose $\boldsymbol{dom}(f)$ is open and that $f : \boldsymbol{dom}(f) \to \mathbb{R}$ is differentiable (it's gradient $\nabla f(x)$ exists $\forall \boldsymbol{x} \in \boldsymbol{dom}(f)$), then f is convex $\iff$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f).$$

Here a geometrical interpretation of convexity which is useful, convexity implies that the segment between any two points in the domain passes above the function f. This implies the following observation: For $f : \boldsymbol{dom}(f) \to \mathbb{R}$ convex, we have the following property. Any hyperplane parallel to the hyperplane tangent to f in $\boldsymbol{x}$ which intersects f on some point $\boldsymbol{z}$ in $\boldsymbol{dom}(f)$ is "above" the tangent hyperplane, more rigorously:

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{z} \rangle, \ \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{dom}(f).$$

We will also use the definition of L-smoothness:

Property 2.3. *(L-Smoothness) let $f : \boldsymbol{dom}(f) \to \mathbb{R}$ be a differentiable function and $X \subseteq \boldsymbol{dom}(f)$ convex and $L \in \mathbb{R}_+$. Function f is called smooth over X if:*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f).$$

Next we look at a geometric identity: the parallelogram law:

Property 2.4. *(Polarization identity) let $a, b \in \mathbb{R}^d$ be two vectors, let $\| \cdot \|$ denote their 2-norm and $\langle a, b \rangle$ denote the inner (dot) product. Then the following equality is true:*

$$\langle a, b \rangle = \frac{1}{2} \|a + b\|^2 - \|a\|^2 - \|b\|^2 = \frac{1}{2} \|a\|^2 + \|b\|^2 - \|a - b\|^2.$$

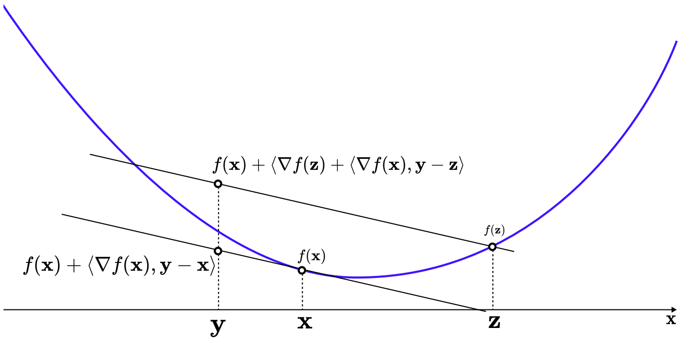
$f(\mathbf{x}) + \langle \nabla f(\mathbf{z}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{z} \rangle$
$f(\mathbf{z})$
$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$
$f(\mathbf{x})$
$\mathbf{y}$
$\mathbf{x}$
$\mathbf{z}$
x

Fig. 4: Visual illustration of observation 2.3.

Property 2.5. *(Young's inequality) let $a, b \in \mathbb{R}^d$ be two vectors, let $\| \cdot \|$ denote their 2-norm and $\langle a, b \rangle$ denote the inner (dot) product and $\lambda \in \mathbb{R}^+$, the following inequality is true:*

$$\langle a, b \rangle \leq \frac{\lambda^2}{2} \|a\|^2 + \frac{1}{2\lambda^2} \|b\|^2.$$