Notes on Federated Optimization

Titouan Renard

May 26, 2022

Contents

1 Setting and Algorithm **1**
1.1 Problem definition 1
1.2 The Federated Averaging Algorithm 1

2 Analysis of the Algorithm **2**
2.1 Setting and assumptions 2
2.2 Convergence of FedAVG 3
2.3 Properties used for the proof 13

3 A toy example to better understand FedAVG **15**
3.1 A simple learning problem: 15

1 Setting and Algorithm

1.1 Problem definition

TODOs

- Setting description
- Cost function

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 η

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 $k \in [0, k)$.

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.* In order to get to a bound we will have to prove two lemmas to show that. 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:

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.

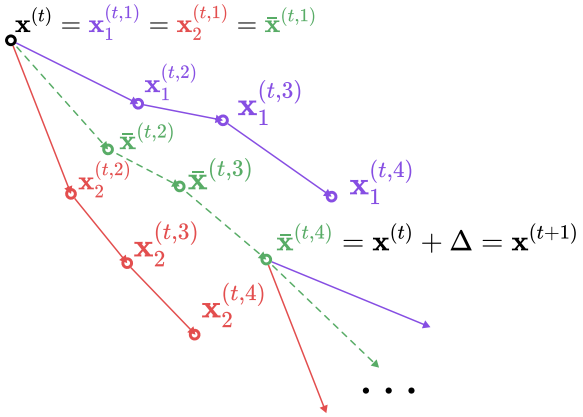

Figure 1: Illustration of the progression of one 4 step round of algorithm 1 with the shadow sequence represented on green.

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:*

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

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:*

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

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 progress which we bounded in lemma 2.2:

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

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.

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

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 ($\dagger$):

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

   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:

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

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

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

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 the lemmas.

Proving Lemma 2.2 (per round progression)

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 (†), most of the work we will have to do will lie in understanding the (♡) term on which we take the expectation. Writing it out we can first split it between the separate client objective functions:

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

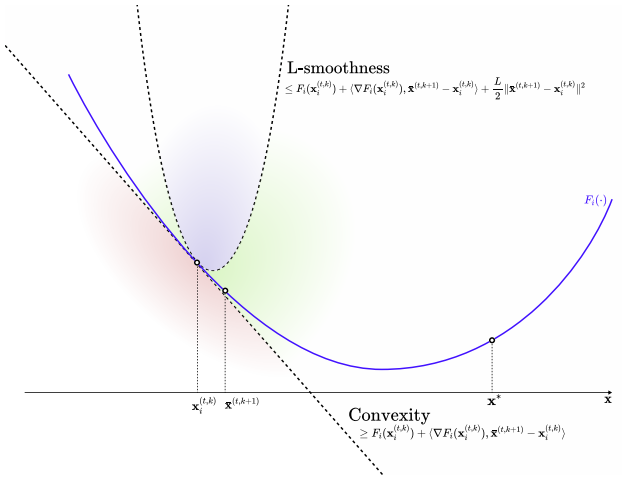

Figure 2: Illustration of the L-smooth/convex properties of F_i. Convexity implies that the function is above the half space tangent to the graph at any point (in red) and that is is below the parabola with parameter L (in blue). In other words, the function being L-smooth and convex is necessarily in 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 (♡) 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 2.1 (which we can do as F_i is convex) we further bound the (♠) 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:

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

This brings us closer to bounding the (♡) term, remembering that we are working with values inside of an expectation, we clearly see that term (■) is a measure of variance (which we will bound later with the client drift lemma). Taking a look at the sum the term that requires and explanation is the (⧫) 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 write out the local client gradient as:

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

where (◁) represents the true gradient of client objective F_i and (▷) the deviation from that 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 (♡) bound as follows:

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

We will analyze (▶) later, when we take expectations, let us now focus on (◀). 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 (◀) 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):

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

The (◀) 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 (♡):

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

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

$$
\begin{aligned}
(\dagger) &= \mathbb{E}\Big[\overbrace{F(\bar{\boldsymbol{x}}^{(t,k+1)}) - F(\boldsymbol{x}^*)}^{(\heartsuit)}\,|\mathcal{F}^{(t,k)}\Big] \\
&\leq \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2\right] - \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] \\
&\quad - \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\right] + \overbrace{\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle\right]}^{(\mathbb{E}[\blacktriangleright])} \\
&\quad + L\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] + \frac{L}{M}\mathbb{E}\left[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\right] \\
&= \overbrace{\left(\frac{1}{2\eta} + L\right)\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}^*\|^2\right]}^{\geq 0 \iff \eta \leq \frac{1}{4L}} - \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right] \\
&\quad - \frac{1}{2\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\right] + \overbrace{\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle\right]}^{(\mathbb{E}[\blacktriangleright])} \\
&\quad + \frac{L}{M}\mathbb{E}\left[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\right]
\end{aligned}
$$

using the assumption $\eta \leq \frac{1}{4L}$ we can remove one term and get to the following form:

$$(\dagger) \leq \overbrace{\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\langle \nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x}), \bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\rangle\right]}^{(\mathbb{E}[\blacktriangleright])} + \frac{L}{M}\mathbb{E}\left[\sum_{i=1}^{M}\|\bar{\boldsymbol{x}}^{(t,k)} - \boldsymbol{x}_i^{(t,k)}\|^2\right]$$
$$-\frac{1}{2\eta}\left(\mathbb{E}\Big[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\Big] + \mathbb{E}\Big[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \boldsymbol{x}^*\|^2\Big]\right).$$

In that notation we see that the last term that poses problem is $(\mathbb{E}[\blacktriangleright])$, which we will be able to bound as follows. Using that $\mathbb{E}\big[\frac{1}{M}\sum_{i=1}^{M}\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\big] = 0$:

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

Using Young's inequality (property 2.5) we further get that:

$$(\mathbb{E}[\blacktriangleright]) \leq \mathbb{E}\left[\sum_{i=1}^{M}\|\frac{1}{M}\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2 + \|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right]$$
$$= \frac{1}{M^2}\mathbb{E}\left[\frac{(\sqrt{2\eta})^2}{2}\sum_{i=1}^{M}\|\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2\right] + \mathbb{E}\left[\frac{1}{2(\sqrt{2\eta})^2}\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right]$$
$$= \frac{\eta}{M^2}\overbrace{\mathbb{E}\left[\sum_{i=1}^{M}\|\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2\right]}^{\leq M\sigma^2} + \frac{1}{4\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right].$$

We can bound $\mathbb{E}\big[\sum_{i=1}^{M}\|\nabla F_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\|^2\big]$ with assumption 6,

$$(\mathbb{E}[\blacktriangleright]) \leq \frac{\eta\sigma^2}{M} + \frac{1}{4\eta}\mathbb{E}\left[\|\bar{\boldsymbol{x}}^{(t,k+1)} - \bar{\boldsymbol{x}}^{(t,k)}\|^2\right].$$

plugging this expression back into $(\dagger)$ we get:

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

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:

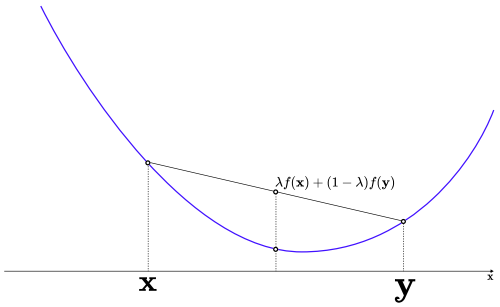

Figure 3: Visual illustration of convexity.

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:

Observation 2.1. *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:

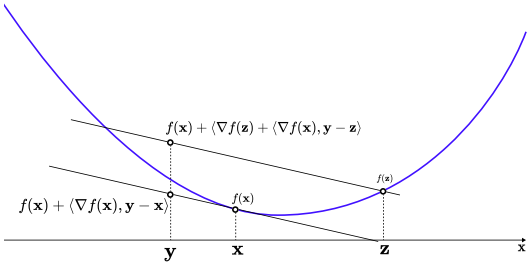

Figure 4: Visual illustration of observation 2.1.

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:*

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

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:*

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

3 A toy example to better understand FedAVG

In order to better understand the behavior of algorithm 1, we first consider a simple toy problem which we will use as a reference to analyze the behavior of the algorithm on a more complex dataset.

3.1 A simple learning problem:

Consider the following "*single layer linear neural network*" model:

$$\hat{y}(\boldsymbol{x}, (w)) = \boldsymbol{w}^T \boldsymbol{x}$$

which we use as a classifier for a simple two class classification problem. We train our model using FedAVG and the following quadratic loss function:

$$L_D(\boldsymbol{w}) = \frac{1}{|D|} \sum_{\boldsymbol{x}_i, \boldsymbol{y}_i}^{D} (\hat{y}(\boldsymbol{x}_i, (w)) - y_i)^2,$$

where D is our dataset containing $|D|$ $\boldsymbol{x}_i, \boldsymbol{y}_i$ input-class pairs. This loss function is convex and L-smooth.