Convergence analysis

Ahmad DABAJA, Rachid EL-AZOUZI

April 2025

1 Introduction

1.1 Objective

In this section, we present a detailed convergence analysis of any **static partial model training scheme**, including our proposed method **FedPLT**. The objective is to show that, under standard assumptions, such methods converge.

We consider the case where each client's objective is smooth and strongly convex, and the global objective is defined as a weighted sum of local losses. Our analysis takes into account the key conditions:

- **Static Partial Layer Training:** Clients update only a fixed subset of model parameters using predefined binary masks over all training rounds.
- **Stochastic Gradient Descent (SGD):** Clients train locally using mini-batch SGD, leading to stochastic updates.

The convergence analysis aims to:

- Quantify how these factors (partial updates and stochastic gradients) influence the global model update and convergence.
- Derive a bound on the expected squared distance between the global model at any round and its optimal value over t rounds.
- Establish a convergence rate to stationary points in the non-convex setting.

The overall goal is to demonstrate that despite the constraint imposed by restricted update masks, FedPLT, as well as other static partial model update methods, remains an effective and theoretically grounded algorithm with provable convergence guarantees.

1.2 Problem Formulation

We consider the standard federated learning optimization problem, where the goal is to minimize a global objective function defined as the weighted average

of local losses across K clients:

$$\min_{W \in \mathbb{R}^d} F(W) := \sum_{k=1}^{K} \frac{n_k}{n} F_k(W), \quad \text{with } n = \sum_{k=1}^{K} n_k, \tag{1}$$

where $F_k(W)$ is the empirical risk on client k given by:

$$F_k(W) := \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(W; x_i^{(k)}, y_i^{(k)}), \tag{2}$$

and $\ell(\cdot; x, y)$ is the per-sample loss function evaluated on a data point $(x_i^{(k)}, y_i^{(k)})$ from client k.

1.3 Partial Update Rule

Let $W^t \in \mathbb{R}^d$ denote the global model at communication round t. Each client k is associated with a binary mask vector $m_k \in \{0,1\}^d$, which specifies the subset of model parameters that client k is responsible for updating. Specifically, for each coordinate $i \in \{1, \ldots, d\}$:

$$(m_k)_i = \begin{cases} 1 & \text{if client } k \text{ is assigned to update coordinate } i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The mask m_k reflects the client's computational budget and its assigned sub-layers in the model. Let $r_k = \|m_k\|_0 / d$ denote the fraction of coordinates that client k is allowed to update.

At the beginning of round t, each selected client k receives the current global model W^t and initializes its local model:

$$W_k^{t,0} := W^t.$$

Each client performs τ local update steps, only modifying the coordinates specified by m_k.
We distinguish between two possible settings:

1. **Local Stochastic Gradient Descent (SGD):** In the default and practical setting, each client uses mini-batch SGD. At each local step s, a random mini-batch $\xi_k^{(s)}$ is sampled from client k's local dataset, and the update rule is:

$$W_k^{t,s+1} = W_k^{t,s} - \eta_k \left(m_k \odot \nabla \ell(W_k^{t,s}; \xi_k^{(s)}) \right), \quad s = 0, \ldots, \tau - 1, \tag{4}$$

where $\eta_k > 0$ is the local learning rate, and $\nabla \ell(W; \xi)$ is the stochastic gradient over the mini-batch ξ.

2. **Local Full Gradient Descent (FGD):** In the idealized setting, each client uses full-batch gradient descent on its local data. The update rule becomes:

$$W_k^{t,s+1} = W_k^{t,s} - \eta_k \left(m_k \odot \nabla F_k(W_k^{t,s}) \right), \quad s = 0, \ldots, \tau - 1, \tag{5}$$

where $\nabla F_k(W)$ is the full local gradient:

$$\nabla F_k(W) := \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla \ell(W; x_i^{(k)}, y_i^{(k)}).$$

After completing local training, the client returns only its local updates: $m_k \odot U_k^t$ in case of SGD or $m_k \odot \bar{U}_k^t$ for FGD, where:

$$U_k^t = W_k^{t,\tau} - W_k^{t,0} = -\eta_k \sum_{s=0}^{\tau-1} \nabla \ell(W_k^{t,s}; \xi_k^{(s)}) \tag{6}$$

and

$$\bar{U}_k^t = W_k^{t,\tau} - W_k^{t,0} = -\eta_k \sum_{s=0}^{\tau-1} \nabla F_k(W_k^{t,s}) \tag{7}$$

The server aggregates the received partial local updates using coordinate-wise weighted averaging.

We define the normalized weight of client k as:

$$c_k := \frac{n_k}{\sum_{j=1}^K n_j}, \quad \text{so that } \sum_{k=1}^K c_k = 1.$$

Then the global model update at round t can be written as:

$$\Delta^t := \left(\sum_{k=1}^K c_k \cdot (m_k \odot U_k^t) \right) \odot \psi,$$

where the compensation vector $\psi \in \mathbb{R}^d$ is defined coordinate-wise as:

$$\psi_i := \begin{cases} \dfrac{\sum_{j=1}^K n_j}{\sum_{k=1}^K n_k (m_k)_i} & \text{if } \sum_{k=1}^K (m_k)_i > 0, \\ \text{any real number} & \text{if } \sum_{k=1}^K (m_k)_i = 0. \end{cases}$$

This vector represents an amplification factor that compensates for the unequal number of clients updating each coordinate. Specifically, ψ_i rescales the aggregated update at coordinate i to reflect the weighted average when only a subset of clients contributes due to masking. A value of $\psi_i = 1$ means all clients updated that coordinate (no compensation needed), while higher values indicate fewer clients participated, requiring stronger correction.

Remark. For coordinates i not updated by any client, i.e., $\sum_k (m_k)_i = 0$, the aggregated update value $\left(\sum_k c_k \cdot (m_k \odot U_k^t)\right)_i$ is exactly zero. Therefore, the corresponding ψ_i is multiplied by zero, and its value does not affect the global update. In such cases, ψ_i may be assigned arbitrarily within $\mathbb{R}$.

2 Convergence Analysis

In this section, we present the convergence analysis of static partial training schemes, including our proposed method, FedPLT.

We begin by introducing a set of standard assumptions commonly adopted in the federated learning literature to ensure theoretical rigor. Additionally, we propose a new assumption specifically tailored to the partial update setting, which captures the alignment properties between masked and full gradients.

Our proof follows the general structure of the analysis in [1], which we extend to accommodate the static partial update mechanisms used in schemes like FedPLT. We detail each step of the convergence proof, highlighting where our contributions diverge from or generalize existing results.

In the cited paper, virtual aggregation is applied at each local step s, whereas in a real setting, aggregation occurs only once every τ steps. We will be using their proving flow by using this notation and then, when necessary in the analysis, we revert to the per-step formulation by explicitly indexing local iterations s.

Thus, we use the simplified update notation:

$$U_k^t = -\eta^t \nabla \ell(W_k^t; \xi_k), \quad \text{and} \quad \bar{U}_k^t = -\eta^t \nabla F_k(W_k^t),$$

2.1 Assumptions

We make the following standard assumptions commonly used in the convergence analysis of federated learning. These ensure that both the local and global objective functions behave in a well-structured and stable manner during training.

- **Assumption 1 (Smoothness):** Each local objective function F_k is L-smooth. That is, for all $W, W' \in \mathbb{R}^d$,

$$\|\nabla F_k(W) - \nabla F_k(W')\| \leq L\|W - W'\|. \tag{8}$$

  This implies that the gradient of F_k does not change too rapidly. Equivalently, by the Descent Lemma, we have:

$$F_k(W') \leq F_k(W) + \nabla F_k(W)^\top (W' - W) + \frac{L}{2}\|W' - W\|^2. \tag{9}$$

- **Assumption 2 (Strong Convexity):** Each local objective function F_k is μ-strongly convex. That is, for all $W, W' \in \mathbb{R}^d$,

$$F_k(W') \geq F_k(W) + \nabla F_k(W)^\top (W' - W) + \frac{\mu}{2}\|W' - W\|^2. \tag{10}$$

- **Assumption 3 (Unbiased Gradient Estimates):** The stochastic gradients computed from local mini-batches are unbiased estimators of the true local gradient:

$$\mathbb{E}_{\xi_k}[\nabla \ell(W; \xi_k)] = \nabla F_k(W), \tag{11}$$

  where ξ_k denotes the randomness in the mini-batch sampling at client k.

- **Assumption 4 (Bounded Variance):** The variance of the stochastic gradients is uniformly bounded:

$$\mathbb{E}_{\xi_k}\left[\|\nabla \ell(W;\xi_k) - \nabla F_k(W)\|^2\right] \leq \sigma^2, \tag{12}$$

for all $W \in \mathbb{R}^d$ and all clients k.

- **Assumption 5 (Bounded Global Gradient Norm):** The norm of the gradient of the global objective function

$$F(W) := \sum_{k=1}^{K} \frac{n_k}{n} F_k(W)$$

is uniformly bounded:

$$\|\nabla F(W)\|^2 \leq G^2, \quad \forall W \in \mathbb{R}^d. \tag{13}$$

- **Assumption 6 (Positive Masked Alignment):** For all clients k and all instances of the model W, we assume that the masking attenuation ratio

$$\rho_k := \frac{\langle m_k \odot \nabla F_k(W),\ W - W^* \rangle}{\langle \nabla F_k(W),\ W - W^* \rangle}$$

satisfies $\rho_k > 0$, where W^* denotes the global minimizer.

This ratio quantifies how masking affects the alignment between the local gradient and the global target. Using the identities:

$$\langle \nabla F_k(W),\ W - W^* \rangle = \|\nabla F_k(W)\| \cdot \|W - W^*\| \cdot \cos(\theta_1),$$

$$\langle m_k \odot \nabla F_k(W),\ W - W^* \rangle = \|m_k \odot \nabla F_k(W)\| \cdot \|W - W^*\| \cdot \cos(\theta_2),$$

$$\|m_k \odot \nabla F_k(W)\| = \|\nabla F_k(W)\| \cdot \cos(\theta_3),$$

we obtain:

$$\rho_k = \frac{\cos(\theta_3) \cdot \cos(\theta_2)}{\cos(\theta_1)}.$$

Interpretation. Since $\cos(\theta_3) > 0$ (the masked gradient is the projection of the gradient), the assumption $\rho_k > 0$ implies that $\cos(\theta_1)$ and $\cos(\theta_2)$ must share the same sign. In other words, if the full gradient $\nabla F_k(W)$ points toward the global optimum W^*, then the masked gradient $m_k \odot \nabla F_k(W)$ must also point in a generally similar direction.

Moreover, when masking improves the alignment with the descent direction (i.e., $\cos(\theta_2) > \cos(\theta_1)$), we get $\rho_k > \cos(\theta_3)$; when it degrades alignment, we get $\rho_k < \cos(\theta_3)$. Thus, ρ_k quantifies how masking affects the alignment between the local gradient and the global target.

Theoretical Justification: Under reasonable initialization and balanced participation, gradients across clients are typically aligned with global progress. This assumption rules out masking that would negate the gradient's direction.

2.2 Finding the Expected Distance Toward the Optimal Global Model

We aim to analyze the expected squared distance to the optimal global model W^* after round $t+1$ of federated training with partial updates. Specifically, our goal is to bound the following quantity:

$$\mathbb{E}\left[\|W^{t+1} - W^*\|^2\right],$$

where the expectation is taken over the randomness in local mini-batch selection. This quantity captures the expected gap between the global model and the optimum across rounds.

Step 1: Expanding the Distance to the Optimum

We begin by expanding the global model's parameters using the global update as follow:

$$W^{t+1} = W^t + \Delta^t,$$

where the global update is defined as:

$$\Delta^t := \psi \odot \sum_{k=1}^{K} c_k \left(m_k \odot U_k^t\right),$$

and each client's stochastic local update U_k^t is given by:

$$U_k^t = -\eta^t \nabla \ell(W_k^t; \xi_k),$$

thus we have:

$$\|W^{t+1} - W^*\|^2 = \|W^t + \Delta^t - W^*\|^2.$$

We then introduce the full (non-stochastic) update:

$$\bar{\Delta}^t := \psi \odot \sum_{k=1}^{K} c_k \left(m_k \odot \bar{U}_k^t\right), \quad \text{where} \quad \bar{U}_k^t = -\eta^t \nabla F_k(W_k^t).$$

by add and subtract it inside the norm:

$$\|W^{t+1} - W^*\|^2 = \left\|(W^t - W^* + \bar{\Delta}^t) + (\Delta^t - \bar{\Delta}^t)\right\|^2.$$

Applying the Euclidean norm expansion (a.k.a. the cosine or parallelogram identity), we obtain:

$$\|W^{t+1} - W^*\|^2 = \underbrace{\|\Delta^t - \bar{\Delta}^t\|^2}_{A_1} + \underbrace{\|W^t - W^* + \bar{\Delta}^t\|^2}_{A_2} + \underbrace{2\left\langle \Delta^t - \bar{\Delta}^t,\ W^t - W^* + \bar{\Delta}^t \right\rangle}_{A_3}. \tag{14}$$

Step 2: Expectation of the Cross-Term A_3

We now analyze the expectation of A_3. The key idea is that the stochastic error term $\Delta^t - \bar{\Delta}^t$ arises from the randomness due to mini-batch sampling at round t, while the other term in the inner product is deterministic.

Specifically, we compute:

$$\mathbb{E}_{\zeta^t}[A_3] = \mathbb{E}_{\zeta^t}\left[2\langle W^t - W^* + \bar{\Delta}^t, \Delta^t - \bar{\Delta}^t\rangle\right].$$

Note that:

- W^t, W^*, and $\bar{\Delta}^t$ are deterministic with respect to the mini-batch sampling at round t,
- $\Delta^t - \bar{\Delta}^t$ is a zero-mean random variable:

$$\mathbb{E}_{\zeta^t}[\Delta^t - \bar{\Delta}^t] = 0.$$

Therefore, the expectation of the inner product vanishes:

$$\mathbb{E}_{\zeta^t}\left[\langle W^t - W^* + \bar{\Delta}^t,\ \Delta^t - \bar{\Delta}^t\rangle\right] = 0,$$

which implies that:

$$\mathbb{E}[A_3] = 0.$$

The term A_3 contributes nothing in expectation. Hence, the expected squared distance simplifies to:

$$\mathbb{E}\left[\|W^{t+1} - W^*\|^2\right] = \mathbb{E}\left[A_1\right] + \mathbb{E}\left[A_2\right].$$

Step 3: Bounding the Stochastic Gradient Variance Term (A_1)

We now provide a concise upper bound for the variance of the global update due to stochastic gradient noise. Recall that the actual update is:

$$\Delta^t := W^{t+1} - W^t = \psi \odot \sum_{k=1}^{K} c_k (m_k \odot U_k^t),$$

and the full gradient update is:

$$\bar{\Delta}^t := \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \bar{U}_k^t),$$

where U_k^t and $\bar{U}_k^t$ are the stochastic and full local updates, respectively.

Letting $\delta_k^t := U_k^t - \bar{U}_k^t$, the update difference becomes:

$$\Delta^t - \bar{\Delta}^t = \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \delta_k^t).$$

We now bound the squared norm:

$$\|\Delta^t - \bar{\Delta}^t\|^2 = \left\| \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \delta_k^t) \right\|^2 .$$

Applying the inequality $\|a \odot v\| \leq \|a\|_\infty \cdot \|v\|$ for any vectors $a, v \in \mathbb{R}^d$ and donating $\overline{\psi} := \|\psi\|_\infty$, we obtain:

$$\|\Delta^t - \bar{\Delta}^t\|^2 \leq \overline{\psi}^2 \cdot \left\| \sum_{k=1}^{K} c_k (m_k \odot \delta_k^t) \right\|^2 .$$

We take the expectation of the squared norm:

$$\mathbb{E}\left[\|\Delta^t - \bar{\Delta}^t\|^2\right] \leq \overline{\psi}^2 \cdot \mathbb{E}\left[\left\| \sum_{k=1}^{K} c_k (m_k \odot \delta_k^t) \right\|^2 \right] \leq \overline{\psi}^2 \cdot \sum_{k=1}^{K} c_k^2 \mathbb{E}\left[\|m_k \odot \delta_k^t\|^2\right].$$

Now applying the masking inequality $\|m_k \odot \delta_k^t\| \leq \alpha_{r_k} \|\delta_k^t\|$, we get:

$$\mathbb{E}\left[\|\Delta^t - \bar{\Delta}^t\|^2\right] \leq \overline{\psi}^2 \cdot \sum_{k=1}^{K} c_k^2 \alpha_{r_k}^2 \cdot \mathbb{E}[\|\delta_k^t\|^2].$$

Finally, applying Assumption 3 (bounded variance):

$$\mathbb{E}[\|\delta_k^t\|^2] = \mathbb{E}[\|U_k^t - \bar{U}_k^t\|^2] \leq (\eta^t)^2 \sigma^2,$$

We conclude:

$$\mathbb{E}[\|\Delta^t - \bar{\Delta}^t\|^2] \leq (\eta^t)^2 \sigma^2 \cdot \overline{\psi}^2 \sum_{k=1}^{K} c_k^2 \alpha_{r_k}^2 .$$

Step 4: Decomposition of A_2 and Finding its expectation

Recall the second term in our main inequality:

$$A_2 := \|W^t - W^* + \bar{\Delta}^t\|^2,$$

We expand this term as:

$$A_2 = \|W^t - W^*\|^2 + \underbrace{\|\bar{\Delta}^t\|^2}_{A_{2.1}} + \underbrace{2\langle W^t - W^*, \bar{\Delta}^t \rangle}_{A_{2.2}} .$$

We define the following components for further analysis:

$$A_{2.1} := \|\bar{\Delta}^t\|^2,$$
$$A_{2.2} := 2\langle W^t - W^*, \bar{\Delta}^t \rangle.$$

Step 4.1: Bounding $A_{2.1}$ (Ideal Update Norm)

Recall the definition of the ideal global update:

$$\bar{\Delta}^t := \psi \odot \sum_{k=1}^{K} c_k \left(m_k \odot \bar{U}_k^t \right),$$

where $\bar{U}_k^t = -\eta^t \nabla F_k(W_k^t)$ is the full local gradient descent.

Using the masking inequality $\|m_k \odot v\| \leq \alpha_{r_k} \|v\|$, and applying the convexity of the squared norm $\|\cdot\|^2$, we get:

$$\|\bar{\Delta}^t\|^2 = \left\| \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \bar{U}_k^t) \right\|^2 \leq \overline{\psi}^2 \left\| \sum_{k=1}^{K} c_k (m_k \odot \bar{U}_k^t) \right\|^2 \leq \overline{\psi}^2 \sum_{k=1}^{K} c_k \|m_k \odot \bar{U}_k^t\|^2.$$

Applying the masking bound:

$$\|m_k \odot \bar{U}_k^t\| \leq \alpha_{r_k} \|\bar{U}_k^t\|, \quad \text{so} \quad \|m_k \odot \bar{U}_k^t\|^2 \leq \alpha_{r_k}^2 \|\bar{U}_k^t\|^2,$$

we obtain:

$$\|\bar{\Delta}^t\|^2 \leq \overline{\psi}^2 \sum_{k=1}^{K} c_k \alpha_{r_k}^2 \|\bar{U}_k^t\|^2.$$

We now bound $\|\bar{U}_k^t\|^2$:

$$\|\bar{U}_k^t\|^2 = (\eta^t)^2 \|\nabla F_k(W_k^t)\|^2.$$

Using the L-smoothness of F_k and the local optimal value $F_k^* := \min_W F_k(W)$, we apply:

$$\|\nabla F_k(W_k^t)\|^2 \leq 2L \left(F_k(W_k^t) - F_k^* \right).$$

Hence,

$$\|\bar{U}_k^t\|^2 \leq 2(\eta^t)^2 L \left(F_k(W_k^t) - F_k^* \right).$$

Substituting into the bound on $\|\bar{\Delta}^t\|^2$, we obtain:

$$A_{2.1} = \|\bar{\Delta}^t\|^2 \leq 2L(\eta^t)^2 \overline{\psi}^2 \sum_{k=1}^{K} \alpha_{r_k}^2 \; c_k \left(F_k(W_k^t) - F_k^* \right).$$

Step 4.2: Bounding the Inner Product Term $A_{2.2}$

We now focus on bounding the term:

$$A_{2.2} := 2 \left\langle W^t - W^*, \bar{\Delta}^t \right\rangle.$$

We start by expanding $\bar{\Delta}^t$ using the aggregation rule:

$$\bar{\Delta}^t := \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \bar{U}_k^t), \quad \text{where} \quad \bar{U}_k^t := -\eta \nabla F_k(W_k^t),$$

We decompose $W^t - W^*$ as:

$$W^t - W^* = (W^t - W_k^t) + (W_k^t - W^*).$$

Substituting this into the inner product and splitting the sum:

$$\begin{aligned} A_{2.2} &= 2\left\langle W^t - W^*,\ \psi \odot \sum_{k=1}^{K} c_k (m_k \odot \bar{U}_k^t) \right\rangle \\ &= 2\sum_{k=1}^{K} \left[\left\langle W^t - W_k^t,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle + \left\langle W_k^t - W^*,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle \right]. \end{aligned}$$

Part 1: Applying the Cauchy–Schwarz inequality, we get:

$$\begin{aligned} \left\langle W^t - W_k^t,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle &\le \|\psi \odot \left(c_k m_k \odot \bar{U}_k^t\right)\| \cdot \|W^t - W_k^t\| \\ &\le \overline{\psi} \cdot c_k \cdot \|m_k \odot \bar{U}_k^t\| \cdot \|W^t - W_k^t\|, \\ &= \overline{\psi} \cdot \eta^t \cdot c_k \cdot \|m_k \odot \nabla F_k(W_k^t)\| \cdot \|W^t - W_k^t\|, \end{aligned}$$

Then, applying the masking bound $\|m_k \odot v\| \le \alpha_{r_k} \|v\|$, where $\alpha_{r_k} := \min(1, \sqrt{d r_k})$, we obtain:

$$\left\langle W^t - W_k^t,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle \le \overline{\psi} \cdot \eta^t \cdot c_k \cdot \alpha_{r_k} \cdot \|\nabla F_k(W_k^t)\| \cdot \|W^t - W_k^t\|.$$

Now applying the AM–GM inequality with parameter $\eta^t > 0$, we get:

$$\|W^t - W_k^t\| \cdot \left(\overline{\psi} \cdot \alpha_{r_k} \cdot \|\nabla F_k(W_k^t)\|\right) \le \frac{1}{\eta^t} \|W^t - W_k^t\|^2 + \eta^t \left(\overline{\psi}\, \alpha_{r_k}\right)^2 \|\nabla F_k(W_k^t)\|^2.$$

Multiplying both sides by the positive scalar $c_k \cdot \eta^t$, we obtain:

$$\left\langle W^t - W_k^t,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle \le c_k \left(\|W^t - W_k^t\|^2 + \left(\eta^t\ \overline{\psi}\ \alpha_{r_k}\right)^2 \|\nabla F_k(W_k^t)\|^2 \right).$$

Recall the L-smoothness of F_k, which implies

$$\|\nabla F_k(W_k^t)\|^2 \le 2L \left(F_k(W_k^t) - F_k^*\right),$$

Substituting it into the previous inequality:

$$\left\langle W^t - W_k^t,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle \le c_k \left(\|W^t - W_k^t\|^2 + \left(\eta^t\ \overline{\psi}\ \alpha_{r_k}\right)^2 L \left(F_k(W_k^t) - F_k^*\right) \right).$$

Part 2: We aim to bound the inner product:

$$\left\langle W_k^t - W^*,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle, \quad \text{where } \bar{U}_k^t := -\eta^t \nabla F_k(W_k^t).$$

Expanding, we obtain:

$$\left\langle W_k^t - W^*,\ \psi \odot \left(c_k m_k \odot \bar{U}_k^t\right) \right\rangle = -\eta^t \left\langle W_k^t - W^*,\ c_k\ \psi \odot m_k \odot \nabla F_k(W_k^t) \right\rangle.$$

We lower-bound the effect of the element-wise masking and compensation vector using:

$$\underline{\psi} := \min_i(\psi_i).$$

Then we apply the client-wise masked alignment ratio:

$$\rho_k^t := \frac{\langle m_k \odot \nabla F_k(W_k^t),\ W_k^t - W^* \rangle}{\langle \nabla F_k(W_k^t),\ W_k^t - W^* \rangle},$$

which measures how well the masked gradient aligns with the true gradient. Under Assumption 5, we assume the alignment is uniformly lower bounded across all rounds and clients:

$$\underline{\rho} := \min_{k \in [K],\ t \in \mathbb{N}} \rho_k^t > 0.$$

Using the strong convexity of F_k, we have:

$$\left\langle \nabla F_k(W_k^t),\ W_k^t - W^* \right\rangle \geq F_k(W_k^t) - F_k(W^*) + \frac{\mu}{2}\|W_k^t - W^*\|^2.$$

Combining the above, we obtain:

$$\begin{aligned}
\left\langle W_k^t - W^*,\ \psi \odot (c_k m_k \odot \bar{U}_k^t) \right\rangle &= -\eta^t c_k \left\langle W_k^t - W^*,\ \psi \odot m_k \odot \nabla F_k(W_k^t) \right\rangle \\
&\leq -\eta^t \underline{\psi} c_k \left\langle W_k^t - W^*,\ m_k \odot \nabla F_k(W_k^t) \right\rangle \\
&= -\eta^t \underline{\psi} \rho_k^t c_k \left\langle \nabla F_k(W_k^t),\ W_k^t - W^* \right\rangle \\
&\leq -\eta^t \underline{\psi} \rho_k^t c_k \left[F_k(W_k^t) - F_k(W^*) + \frac{\mu}{2}\|W_k^t - W^*\|^2 \right].
\end{aligned}$$

Finally, using $\rho_k^t \geq \underline{\rho}$, we get the uniform lower bound:

$$\left\langle W_k^t - W^*,\ \psi \odot (c_k m_k \odot \bar{U}_k^t) \right\rangle \leq -\eta^t\ \underline{\psi}\ \underline{\rho}\ c_k \left[F_k(W_k^t) - F_k(W^*) + \frac{\mu}{2}\|W_k^t - W^*\|^2 \right].$$

Part 3: Final Bound on $A_{2.2}$ Combining the two parts, we obtain:

$$\begin{aligned}
A_{2.2} = 2\left\langle W^t - W^*,\ \bar{\Delta}^t \right\rangle \leq \sum_{k=1}^{K} \Bigg[& c_k \|W^t - W_k^t\|^2 \\
& + 2(\eta^t\ \overline{\psi}\ \alpha_{r_k})^2\ L\left(F_k(W_k^t) - F_k^*\right) \\
& - 2\eta^t\ \underline{\psi}\ \underline{\rho}\ c_k \left(F_k(W_k^t) - F_k(W^*) + \frac{\mu}{2}\|W_k^t - W^*\|^2 \right) \Bigg],
\end{aligned}$$

Step 4.3: Grouping and Rearranging the Terms of A_2

Recall the decomposition:

$$A_2 := \|W^t - W^* + \bar{\Delta}^t\|^2 = \underbrace{\|W^t - W^*\|^2}_{\text{Initial distance}} + \underbrace{\|\bar{\Delta}^t\|^2}_{A_{2.1}} + \underbrace{2\langle W^t - W^*, \bar{\Delta}^t \rangle}_{A_{2.2}}.$$

Substituting the bounds from previous steps, we obtain:

$$
\begin{aligned}
A_2 \leq & \underbrace{\|W^t - W^*\|^2}_{\text{Initial distance}} \\
& + \underbrace{\sum_{k=1}^{K} \left[2(\eta^t)^2 \overline{\psi}^2 \ \alpha_{r_k}^2 L \ c_k (F_k(W_k^t) - F_k^*) \right]}_{\text{Ideal update norm } (A_{2.1})} \\
& + \underbrace{\sum_{k=1}^{K} \left[c_k \|W^t - W_k^t\|^2 - \mu \eta^t \ \underline{\psi} \ \underline{\rho} \ c_k \|W_k^t - W^*\|^2 \right]}_{\text{Model difference terms from } A_{2.2}} \\
& + \underbrace{\sum_{k=1}^{K} \left[2(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 \ L \ c_k (F_k(W_k^t) - F_k^*) - 2\eta^t \ \underline{\psi} \ \underline{\rho} \ c_k \ (F_k(W_k^t) - F_k(W^*)) \right]}_{\text{Loss terms from } A_{2.2}}.
\end{aligned}
$$

Combining the result and regrouping all components, we obtain:

$$
\begin{aligned}
A_2 \leq \|W^t - W^*\|^2 + \sum_{k=1}^{K} \Bigg[& c_k \|W^t - W_k^t\|^2 - \mu \eta^t \ \underline{\psi} \ \underline{\rho} \ c_k \|W_k^t - W^*\|^2 \\
& + 4(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 \ L \ c_k (F_k(W_k^t) - F_k^*) - 2\eta^t \ \underline{\psi} \ \underline{\rho} \ c_k (F_k(W_k^t) - F_k(W^*)) \Bigg].
\end{aligned}
$$

We recall that the global model at round t can be written as a weighted average of local models:

$$
W^t = \sum_{k=1}^{K} c_k W_k^t, \quad \text{with} \quad c_k := \frac{n_k}{\sum_{j=1}^{K} n_j}.
$$

Using the convexity of the squared norm, we apply:

$$
\|W^t - W^*\|^2 \leq \sum_{k=1}^{K} c_k \|W_k^t - W^*\|^2.
$$

Substituting into the earlier bound, we obtain:

$$
\begin{aligned}
A_2 \leq & (1 - \mu \eta^t \ \underline{\psi} \ \underline{\rho}) \cdot \|W^t - W^*\|^2 \\
& + \sum_{k=1}^{K} c_k \|W^t - W_k^t\|^2 \\
& + \underbrace{\sum_{k=1}^{K} \left[4(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 L \ c_k (F_k(W_k^t) - F_k^*) - 2\eta^t \ \underline{\psi} \ \underline{\rho} \ c_k (F_k(W_k^t) - F_k(W^*)) \right]}_{A_{2.4}}.
\end{aligned}
$$

Step 4.4: Bounding $A_{2.4}$

We define $A_{2.4}$ to capture all the loss-related contributions arising in the expansion of A_2:

$$A_{2.4} := \sum_{k=1}^{K} \Bigg[4(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 L \cdot c_k (F_k(W_k^t) - F_k^*) - 2\eta^t \ \underline{\psi} \ \underline{\rho} \cdot c_k (F_k(W_k^t) - F_k(W^*)) \Bigg].$$

We regroup the above term by adding and subtracting F_k^* in the expression $(F_k(W_k^t) - F_k(W^*))$:

$$A_{2.4} := \sum_{k=1}^{K} \Bigg[\underbrace{\left(4(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 L - 2\eta^t \ \underline{\psi} \ \underline{\rho}\right)}_{:= \ -\gamma_k^t} \cdot c_k (F_k(W_k^t) - F_k^*) + \underbrace{2\eta^t \ \underline{\psi} \ \underline{\rho}}_{:= \ \beta^t} \cdot c_k (F_k(W^*) - F_k^*) \Bigg].$$

We insert and subtract F^* inside both gap terms accordingly:

$$A_{2.4} = \sum_{k=1}^{K} \Bigg[- \gamma_k^t \cdot c_k \left(F_k(W_k^t) - F^* + F^* - F_k^*\right) + \beta^t \cdot c_k \left(F_k(W^*) - F^* + F^* - F_k^*\right) \Bigg].$$

We regroup and simplify to highlight the three key components:

$$\begin{aligned} A_{2.4} = \sum_{k=1}^{K} \Bigg[& - \gamma_k^t \cdot c_k \left(F_k(W_k^t) - F^*\right) \\ & + (\beta^t - \gamma_k^t) \cdot c_k \left(F^* - F_k^*\right) \\ & + \beta^t \cdot c_k \left(F_k(W^*) - F^*\right) \Bigg]. \end{aligned}$$

Noting that $F^* = \sum_{k=1}^{K} c_k F_k(W^*)$, this implies:

$$\sum_{k=1}^{K} c_k (F_k(W^*) - F^*) = 0.$$

We thus eliminate the final term and rewrite the expression as:

$$A_{2.4} = \sum_{k=1}^{K} \Bigg[- \gamma_k^t \cdot c_k \left(F_k(W_k^t) - F^*\right) + (\beta^t - \gamma_k^t) \cdot c_k \left(F^* - F_k^*\right) \Bigg].$$

Bounding $(F_k(W_k^t) - F^*)$ gap term By adding and subtracting $F_k(W^t)$, we write:

$$F_k(W_k^t) - F_k^* = \left(F_k(W_k^t) - F_k(W^t)\right) + \left(F_k(W^t) - F^*\right).$$

Using the convexity of F_k, we have:

$$F_k(W_k^t) - F_k(W^t) \geq \left\langle \nabla F_k(W^t), \ W_k^t - W^t \right\rangle.$$

Applying the AM–GM inequality with parameter η^t, we get:

$$\left\langle \nabla F_k(W^t),\ W_k^t - W^t \right\rangle \geq -\frac{\eta^t}{2}\|\nabla F_k(W^t)\|^2 - \frac{1}{2\eta^t}\|W_k^t - W^t\|^2.$$

Using the L-smoothness of F_k:

$$\|\nabla F_k(W^t)\|^2 \leq 2L\left(F_k(W^t) - F_k^*\right),$$

we obtain the following bound:

$$F_k(W_k^t) - F^* \geq F_k(W^t) - F^* - \eta^t L\left(F_k(W^t) - F_k^*\right) - \frac{1}{2\eta^t}\|W_k^t - W^t\|^2.$$

Now we decompose the bound by adding and subtracting F^* in $(F_k(W^t) - F_k^*)$:

$$\begin{aligned} F_k(W_k^t) - F_k^* \geq\ & \left(1 - \eta^t L\right)\left(F_k(W^t) - F^*\right) \\ & - \eta^t L(F^* - F_k^*) - \frac{1}{2\eta^t}\|W_k^t - W^t\|^2. \end{aligned}$$

We substitute the lower bound on $F_k(W_k^t) - F^*$ into the original upper bound for $A_{2.4}$. This gives:

$$\begin{aligned} A_{2.4} \leq \sum_{k=1}^{K} \Bigg[& \gamma_k^t \cdot c_k \left((\eta^t L - 1)(F_k(W^t) - F^*) + \eta^t L(F^* - F_k^*) + \frac{1}{2\eta^t}\|W_k^t - W^t\|^2 \right) \\ & + (\beta^t - \gamma_k^t) \cdot c_k (F^* - F_k^*) \Bigg]. \end{aligned}$$

We now regroup the terms in the corrected upper bound of $A_{2.4}$:

$$\begin{aligned} A_{2.4} \leq \sum_{k=1}^{K} \Bigg[& \gamma_k^t (\eta^t L - 1) \cdot c_k\ \left(F_k(W^t) - F^*\right) \\ & + \left[\gamma_k^t (\eta^t L - 1) + \beta^t\right] \cdot c_k\ (F^* - F_k^*) \\ & + \gamma_k^t \cdot \frac{1}{2\eta^t} \cdot c_k\ \|W_k^t - W^t\|^2 \Bigg]. \end{aligned}$$

We now simplify the bound on $A_{2.4}$ under the following conditions:

- $\eta^t < \frac{1}{L} \Rightarrow (\eta^t L - 1) < 0$,
- $\gamma_k^t > 0$, so $\gamma_k^t(\eta^t L - 1) < 0$,
- $\frac{\gamma_k^t}{2\eta^t} < 1$, so $\gamma_k^t \cdot \frac{1}{2\eta^t} \cdot c_k\ \|W_k^t - W^t\|^2 < c_k\ \|W_k^t - W^t\|^2$,
- The combined coefficient:

$$\gamma_k^t(\eta^t L - 1) + \beta^t = 2(\eta^t)^2 L \cdot \left[2\overline{\psi}^2 \alpha_{r_k}^2 (1 - \eta^t L) + \underline{\psi\rho}\right] \leq 2(\eta^t)^2 L\nu,$$

  where $\nu := \max_t \left[2\overline{\psi}^2 \alpha_{r_k}^2 (1 - \eta^t L) + \underline{\psi\rho}\right] > 0$.

Applying these observations to the previous bound:

$$A_{2.4} \leq \sum_{k=1}^{K} \Big[2(\eta^t)^2 L\nu \cdot c_k (F^* - F_k^*) + c_k \|W_k^t - W^t\|^2 \Big].$$

We define the aggregate heterogeneity gap:

$$\Lambda := \sum_{k=1}^{K} c_k (F^* - F_k^*),$$

thus, Applying these observations to the previous bound:

$$A_{2.4} \leq \sum_{k=1}^{K} \Big[2(\eta^t)^2 L\nu \cdot c_k \ \Lambda + c_k \|W_k^t - W^t\|^2 \Big].$$

Step 4.5: Regrouping the Final Bound on A_2

We now regroup and simplify the final upper bound on A_2, using the decompositions derived in previous steps.

Recall that:

$$\begin{aligned} A_2 \leq & (1 - \mu\eta^t \ \underline{\psi} \ \underline{\rho}) \cdot \|W^t - W^*\|^2 \\ & + \sum_{k=1}^{K} c_k \|W^t - W_k^t\|^2 \\ & + \underbrace{\sum_{k=1}^{K} \Big[4(\eta^t \ \overline{\psi} \ \alpha_{r_k})^2 L \ c_k (F_k(W_k^t) - F_k^*) - 2\eta^t \ \underline{\psi} \ \underline{\rho} \ c_k (F_k(W_k^t) - F_k(W^*)) \Big]}_{A_{2.4}}. \end{aligned}$$

We substitute the simplified bound for the loss-related component $A_{2.4}$ derived in step 6.4, yielding:

$$\begin{aligned} A_2 \leq & (1 - \mu\eta^t \ \underline{\psi} \ \underline{\rho}) \|W^t - W^*\|^2 \\ & + \sum_{k=1}^{K} 2c_k \|W_k^t - W^t\|^2 \\ & + 2(\eta^t)^2 L\nu \cdot \Lambda. \end{aligned}$$

Step 4.6: Taking the Expectation of A_2

To compute $\mathbb{E}[A_2]$, we note that:

$$\mathbb{E}[A_2] \leq (1 - \mu\eta^t \underline{\psi}\, \underline{\rho})\|W^t - W^*\|^2 + 2\,\mathbb{E}\left[\sum_{k=1}^{K} c_k \|W_k^t - W^t\|^2\right] + 2(\eta^t)^2 L\nu \cdot \Lambda.$$

We now bound the expected local divergence term:

$$\mathbb{E}\left[\sum_{k=1}^{K} c_k \|W_k^t - W^t\|^2\right].$$

Recall that:

- $W_k^t = W_k^{t,\tau}$ is the local model of client k after τ local steps,
- $W^t = W_k^{t,0}$ is the global model at round t, identical to the local initialization,
- Local updates are computed with coordinate-wise masks:

$$W_k^{t,s+1} = W_k^{t,s} - \eta^t \left(m_k \odot \nabla \ell(W_k^{t,s}; \xi_k^{(s)})\right).$$

By unrolling the update over τ local steps, we obtain:

$$W_k^t - W^t = W_k^{t,\tau} - W_k^{t,0} = -\eta^t \sum_{s=0}^{\tau-1} \left(m_k \odot \nabla \ell(W_k^{t,s}; \xi_k^{(s)})\right).$$

Taking the squared norm and applying Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left[\|W_k^t - W^t\|^2\right] &= (\eta^t)^2 \cdot \mathbb{E}\left[\left\|\sum_{s=0}^{\tau-1} m_k \odot \nabla \ell(W_k^{t,s}; \xi_k^{(s)})\right\|^2\right] \\ &\leq (\eta^t)^2 \cdot \tau \sum_{s=0}^{\tau-1} \mathbb{E}\left[\|m_k \odot \nabla \ell(W_k^{t,s}; \xi_k^{(s)})\|^2\right].\end{aligned}$$

Applying the masking bound $\|m_k \odot v\| \leq \alpha_{r_k} \cdot \|v\|$ and the variance assumption:

$$\mathbb{E}\left[\|\nabla \ell(W_k^{t,s}; \xi_k^{(s)})\|^2\right] \leq G^2 + \sigma^2,$$

We obtain:

$$\mathbb{E}\left[\|W_k^t - W^t\|^2\right] \leq (\eta^t)^2 \cdot \tau^2 \cdot \alpha_{r_k}^2 \cdot (G^2 + \sigma^2).$$

Finally, summing over clients with weights c_k, we get:

$$\mathbb{E}\left[\sum_{k=1}^{K} c_k \|W_k^t - W^t\|^2\right] \leq (\eta^t)^2 \cdot \tau^2 \cdot (G^2 + \sigma^2) \sum_{k=1}^{K} c_k \alpha_{r_k}^2.$$

We denote this term by:

$$\Gamma := \sum_{k=1}^{K} c_k \alpha_{r_k}^2,$$

and thus conclude:

$$\mathbb{E}\left[\sum_{k=1}^{K} c_k \|W_k^t - W^t\|^2\right] \leq (\eta^t)^2 \cdot \tau^2 \cdot (G^2 + \sigma^2) \cdot \Gamma.$$

Combining this with the earlier result for A_2, we obtain the following upper bound in expectation:

$$\mathbb{E}[A_2] \leq (1 - \mu\eta^t \ \underline{\psi} \ \underline{\rho}) \cdot \|W^t - W^*\|^2 + 2 \ (\eta^t)^2 \cdot \tau^2 \cdot (G^2 + \sigma^2) \cdot \Gamma + 2 \ (\eta^t)^2 L\nu \cdot \Lambda,$$

Step 5: Recursive bound over the expected optimal gap

We recall the Euclidean expansion:

$$\|W^{t+1} - W^*\|^2 = \underbrace{\|\Delta^t - \bar{\Delta}^t\|^2}_{A_1} + \underbrace{\|W^t - W^* + \bar{\Delta}^t\|^2}_{A_2} + \underbrace{2\left\langle \Delta^t - \bar{\Delta}^t, \ W^t - W^* + \bar{\Delta}^t \right\rangle}_{A_3}.$$

Taking expectations over the stochasticity at round t, and applying the derived bounds, we get:

$$\begin{aligned}
\mathbb{E}\left[\|W^{t+1} - W^*\|^2\right] &= \mathbb{E}[A_1] + \mathbb{E}[A_2] + \mathbb{E}[A_3] \\
&\leq (\eta^t)^2 \sigma^2 \cdot \overline{\psi}^2 \Gamma \\
&\quad + (1 - \mu\eta^t \ \underline{\psi} \ \underline{\rho}) \cdot \mathbb{E}\left[\|W^t - W^*\|^2\right] \\
&\quad + 2 \ (\eta^t)^2 \tau^2 (G^2 + \sigma^2) \cdot \Gamma \\
&\quad + 2(\eta^t)^2 L\nu \cdot \Lambda.
\end{aligned}$$

We define the sequence:

$$D^t := \mathbb{E}\left[\|W^t - W^*\|^2\right],$$

and obtain the recursive inequality:

$$D^{t+1} \leq \left(1 - z \ \eta^t\right) \cdot D^t + (\eta^t)^2 \ B,$$

where:

$$z := \mu \ \underline{\psi} \ \underline{\rho}, \quad B := 2 \ \tau^2 (G^2 + \sigma^2) \ \Gamma + 2L\nu \ \Lambda + \sigma^2 \ \overline{\psi}^2 \Gamma.$$

Convergence Analysis of the Recursive Inequality

We recall the recursive inequality derived in Step 5:

$$D^{t+1} \leq (1 - z\eta^t) \cdot D^t + (\eta^t)^2 B,$$

We now analyze the convergence behavior under two scenarios: (1) constant step size, and (2) decaying step size.

Case 1: Constant Learning Rate

Assume a constant step size $\eta^t = \eta \in (0, \frac{1}{z})$. Then the recursion becomes:

$$D^{t+1} \leq (1 - z\eta)\, D^t + \eta^2 B.$$

Let $\rho := 1 - z\eta < 1$. Then by recursion:

$$D^{t+1} \leq \rho^{t+1} D^0 + \eta^2 B \sum_{j=0}^{t} \rho^j = \rho^{t+1} D^0 + \eta^2 B \cdot \frac{1 - \rho^{t+1}}{1 - \rho}.$$

Since $\frac{1-\rho^{t+1}}{1-\rho} \leq \frac{1}{1-\rho} = \frac{1}{z\eta}$, we obtain the final bound:

$$D^{t+1} \leq \rho^{t+1} D^0 + \frac{\eta B}{z}.$$

Interpretation: The first term decays geometrically, while the second term is a residual error that depends on η. Therefore, the convergence is fast, but with a gap of size: $\frac{\eta B}{z}$ around the optimal point W^*.

Implication: Choosing a smaller η reduces the residual error, but slows down convergence. This is a speed-error trade-off: a smaller learning rate implies lower asymptotic error but slower descent.

Case 2: Decaying Learning Rate

We now analyze the convergence behavior under a decaying learning rate of the form:

$$\eta^t = \frac{1}{z(t+1)},$$

Recall the recursive inequality:

$$D^{t+1} \leq (1 - z\eta^t) D^t + (\eta^t)^2 B.$$

Substituting the decaying step size into the recursion, we get:

$$D^{t+1} \leq \left(1 - \frac{1}{t+1}\right) D^t + \frac{B}{z^2(t+1)^2} = \frac{t}{t+1} D^t + \frac{B}{z^2(t+1)^2}.$$

We now prove by mathematical induction that for all $t \geq 0$, there exists a constant $C \geq \frac{B}{z^2}$ such that:

$$D^t \leq \frac{C}{t+1}.$$

Base Case: At $t = 0$,

$$D^1 \leq D^0 + \frac{B}{z^2} \leq \frac{C}{0+1},$$

provided that:

$$C \geq D^0 + \frac{B}{z^2} \geq \frac{B}{z^2}.$$

Inductive Step: Assume the hypothesis holds at step t, i.e.,

$$D^t \leq \frac{C}{t+1}.$$

Then:

$$\begin{aligned} D^{t+1} &\leq \frac{t}{t+1} \cdot \frac{C}{t+1} + \frac{B}{z^2(t+1)^2} \\ &= \frac{Ct + \frac{B}{z^2}}{(t+1)^2}. \end{aligned}$$

We now compare this to $\frac{C}{t+2}$ by defining the function:

$$f(t) := \frac{Ct + \frac{B}{z^2}}{(t+1)^2} - \frac{C}{t+2}.$$

It can be proved through typical methods of calculus that $f(t) \leq 0$ if $C \geq \frac{B}{z^2}$.

Thus, for any $C \geq \frac{B}{z^2}$, the inequality $D^{t+1} \leq \frac{C}{t+2}$ holds. This completes the inductive step.

Conclusion: By induction, for all $t \geq 0$, we have:

$$D^t \leq \frac{C}{t+1}, \quad \text{for some constant } C \geq \frac{B}{z^2}.$$

Therefore, the expected squared distance to the optimum decays at the rate:

$$\boxed{D^t = \mathcal{O}\left(\frac{1}{t}\right)}.$$

Interpretation: This gives asymptotic convergence to the optimum W^*, but at a sublinear rate. The error shrinks slowly over time, especially during early rounds.

Summary and Comparison

- With a constant step size, we converge fast to a small neighborhood around W^*, with asymptotic error $\frac{\eta B}{z}$.
- With a decaying step size, we converge exactly to W^*, but at a slower $\mathcal{O}(1/t)$ rate.

Practical Guideline: In practice, we can use a hybrid strategy, starting with a large constant step size to ensure fast initial progress, then decaying it slowly to improve final accuracy.

References

[1] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.