ASTRA: Adaptive Student-Teacher Method for Robust Aggregation and Client Drift Reduction in Federated Learning

João Gonçalves[1], John Sousa[1], Rafael Veiga[1], Lucas Bastos[2], Lucas Pacheco[1], Iago Medeiros[1], Denis Rosário[1], and Eduardo Cerqueira[1]

[1]Federal University of Para (UFPA)

[2]Federal University of South and Southeast of Para (UNIFESPA)

{joao.goncalves, rafael.teixeira.silva}@icen.ufpa.br

{john.sousa, lucas.pacheco}@itec.ufpa.br, bastos.lucas@unifesspa.edu.br,

{iagomedeiros, denis, cerqueira}@ufpa.br

Abstract. *Federated Learning (FL) faces critical convergence challenges when client data is heterogeneous (Non-IID). Existing solutions often trade computational efficiency for stability, incurring a high overhead. This paper proposes ASTRA (Adaptive Student-Teacher Method for Robust Aggregation and Client Drift Reduction in Federated Learning), a method that integrates geometric regularization with a dynamic self-distillation mechanism. Unlike static approaches, ASTRA employs a curriculum-based schedule that transitions from intensive guidance to periodic semantic correction. This strategy ensures that local updates remain aligned with the global objective (mitigating client drift), while drastically reducing the computational cost of the teacher model. Experimental results in Non-IID CIFAR-10 demonstrate that ASTRA outperforms the structural baseline (FedProx) by 32.7% in accuracy under severe heterogeneity, effectively preventing catastrophic divergence while maintaining training speeds within 2.3% of the lightweight FedAvg algorithm.*

1. Introduction

The exponential growth of data generation, ranging from mobile devices to Internet of Things (IoT) sensors, has fundamentally shifted the landscape of Machine Learning (ML) applications. Traditional centralized ML applications rely on aggregating raw data on a single central server, raising critical issues regarding privacy and communication bandwidth [Rodríguez-Barroso et al. 2024]. In contrast, Federated Learning (FL) emerged as the standard for decentralized intelligence, allowing models to train collaboratively on user devices while sharing only weight updates [Amiri et al. 2024]. However, the FL paradigm introduces a complex challenge regarding the statistical behavior of the data, since edge data is inherently heterogeneous in terms of data distributions that vary significantly across clients, *i.e.*, non-independent and identically distributed (Non-IID) [Jimenez G. et al. 2024].

This inherent statistical divergence creates a critical failure mode known as client drift, where the model degrades its performance due to inconsistent optimization objectives [Yan et al. 2023]. Specifically, as each client optimizes its local model based solely on its skewed subset of data, the resulting weight updates bias the model toward local optima that are frequently incompatible with the global objective. Consequently, when

the central server aggregates these divergent updates, the global model suffers from catastrophic forgetting of minority classes or fails to converge altogether [Cooray et al. 2025]. Empirical experiments indicate that in highly heterogeneous settings, such as partitioned CIFAR-10 with a Dirichlet concentration of $\alpha = 0.1$, standard proposals, such as FedAvg, suffer performance degradations of up to 30% compared to IID data distribution [Li et al. 2020]. Hence, the central challenge is not simply to define how to aggregate models, but how to reconcile conflicting gradients without imposing prohibitive computational overhead on resource-constrained edge devices.

Current mitigation strategies typically address this conflict through one of two isolated approaches, neither of which is fully sufficient. Structural approaches, such as FedProx [Li et al. 2020, Mhou and Senekane 2026], apply proximal constraints (L2-norm) to penalize weights that deviate too far from the global model. Although this stabilizes training, it imposes a structural rigidity that restricts the ability of the local model to learn unique features, stifling adaptation to local distributions. In contrast, semantic approaches utilize Knowledge Distillation (KD) to align model outputs (logits) rather than weights [Mora et al. 2024, Qin et al. 2024]. While promising, standard KD methods impose a constant computational penalty by requiring a teacher forward pass in every training step. Thus, there is a lack of methods capable of ensuring both structural stability and semantic consistency without incurring the continuous computational cost of traditional distillation.

This paper presents ASTRA (Adaptive Student-Teacher method for Robust Aggregation and Client Drift Reduction in Federated Learning). ASTRA addresses the challenges of Non-IID data through two complementary approaches designed to mitigate drift while maintaining efficiency: Dual-Space Regularization and Curriculum-Based Periodic Correction. In its operation, ASTRA employs a dual-space regularization approach to ensure structural stability by combining geometric constraints inspired by proximal terms to limit physical divergence in client weights. With the Semantic Self-Distillation, the method transfers the logits of a frozen teacher model (the global state) to the active local student, enforcing semantic consistency. We also introduce a Curriculum-Based Periodic Correction approach that dynamically balances global stability with local plasticity. This allows the model to transition from intensive guidance to periodic correction, ensuring high accuracy even in highly Non-IID environments while drastically reducing the computational cost of the teacher model. Evaluation results demonstrate that ASTRA prevents catastrophic divergence compared to baseline methods, outperforming FedProx by up to 32.7% in accuracy. Furthermore, ASTRA maintains training speeds within 2.3% of the lightweight FedAvg algorithm, proving that robust semantic alignment is achievable with minimal resource overhead.

ASTRA achieves robust aggregation through a client-side regularization paradigm: rather than modifying the aggregation rule itself, the method ensures that local updates arriving at the server are already drift-corrected and semantically aligned with the global objective. This design philosophy recognizes that aggregation robustness is not solely determined by the weighting scheme, but fundamentally by the quality and consistency of the inputs being aggregated. By enforcing dual-space constraints locally, ASTRA guarantees that the standard FedAvg aggregation (weighted averaging by dataset size) operates on stabilized updates, thereby producing robust global models even under severe heterogeneity. This approach maintains full compatibility with existing FL infrastructure while achieving superior convergence stability compared to methods that rely solely on server-side aggregation modifications.

The remainder of this paper is organized as follows. Section 2 reviews the state-of-the-art in FL, focusing on heterogeneity mitigation and KD techniques. Section 3 presents the ASTRA proposal. Section 4 explores the simulation setup and presents comparative results against baseline methods, including a sensitivity analysis of data heterogeneity. Finally, Section 5 concludes the paper and introduces future works.

2. Related Work

This section presents the related works focusing on heterogeneity mitigation and KD techniques. In this sense, Li *et al.* introduced FedProx [Li et al. 2020], which is a structural optimization method designed to tackle heterogeneity in FL. Their approach modified the local objective by adding a proximal term (L2-norm regularization) to restrict the distance between the local model and the global server state. While this effectively stabilizes convergence by preventing local weights from drifting too far, it imposes a significant limitation: the proximal term acts as a rigid constraint that indiscriminately penalizes weight updates, potentially restraining the model's ability to learn beneficial unique features from local data.

Karimireddy *et al.* proposed SCAFFOLD to address the client drift phenomenon utilizing control variates [Karimireddy et al. 2020]. In this sense, SCAFFOLD estimates the direction of the drift in local updates and explicitly corrects it during aggregation. Although SCAFFOLD significantly reduces the variance among clients and accelerates convergence, it doubles the communication overhead by requiring the transmission of control variates alongside model parameters. This makes it less viable for bandwidth-constrained edge environments where uplink speed is a critical bottleneck.

Wang *et al.* introduced FedNova, which is a normalized averaging method that accounts for objective inconsistency when clients perform different numbers of local epochs [Wang et al. 2021]. FedNova ensures an unbiased global update by scaling local updates based on the amount of local computation. However, similar to FedProx and SCAFFOLD, FedNova operates exclusively in the parameter space. It neglects the semantic consistency of the model's predictions, failing to correct cases where clients learn conflicting class boundaries despite exhibiting similar weight magnitudes.

Jeong *et al.* proposed a KD-based method to leverage the exchange of logit outputs to align client models, called FedDistill [Jeong et al. 2018]. In its operation, the global model acts as a teacher, guiding local students to maintain consistency in the output. By operating directly on the prediction space without requiring auxiliary projection heads or complex regularization terms, FedDistill maintains high computational efficiency, incurring negligible overhead compared to FedAvg. Although promising for improving accuracy, a major drawback of such semantic approaches is their frequent reliance on shared public proxy data to align distributions, a requirement that violates the privacy-first principle of FL in sensitive domains where no public data is available.

Li *et al.* proposed a Model-Contrastive FL (MOON) that considers a contrastive loss to maximize the similarity between the local model's representations and the global model's representations [Li et al. 2021]. MOON effectively corrects local training by aligning feature representations in a latent space. However, MOON significantly increases the local computational burden by requiring multiple forward passes (Local, Global, and Previous Local models) for every single training batch. This drastically increases training time and energy consumption, making it prohibitive for resource-constrained edge

devices.

Table 1 summarizes the comparison of analyzed methods, evaluating their capabilities in drift mitigation, semantic awareness, efficiency, and data-free. We define Semantic Awareness as the ability to regularize training based on the model's decision boundaries (output logits) or feature representations, rather than relying solely on rigid parameter proximity (*e.g.*, Euclidean distance). Additionally, Data-Free operation refers to the capacity to perform knowledge transfer without requiring an auxiliary public proxy dataset. This distinction is critical, as methods relying on public data (*e.g.*, standard Co-Distillation) often compromise the privacy-first principle of FL in sensitive domains.

By analyzing the state-of-the-art methods, we conclude that there are significant trade-offs to consider when defining drift mitigation in an FL scenario. This analysis highlights the difficulty of balancing structural stability (FedProx) with semantic flexibility (MOON) without incurring prohibitive computational costs. In this context, a hybrid method that combines lightweight regularization with efficient and data-free distillation is essential. To the best of our knowledge, none of the existing works provide a method that considers a Curriculum-Based schedule to dynamically balance these objectives, ensuring high accuracy in Non-IID settings while maintaining training speeds.

Table 1. Comparison of state-of-the-art FL methods against ASTRA

Method	Drift Mitigation	Semantic Awareness	Comm. Efficient	Comp. Efficient	Data-Free
FedProx	✓	✗	✓	✓	✓
SCAFFOLD	✓	✗	✗	✓	✓
FedNova	✓	✗	✓	✓	✓
FedDistill	✓	✓	✓	✗	✗
MOON	✓	✓	✓	✗	✓
ASTRA (Ours)	✓	✓	✓	✓	✓

3. ASTRA Overview

This section presents ASTRA (Adaptive Student-Teacher method for Robust Aggregation), which considers two complementary approaches (Dual-Space Regularization and Curriculum-Based Periodic Correction) to mitigate drift while maintaining efficiency. ASTRA employs geometric regularization inspired by proximal terms to limit physical divergence in client weights. In addition, it considers a Self-Distillation approach to transfer the logits of a frozen teacher model (the global state) to the active local student, enforcing semantic consistency. Rather than treating drift mitigation as a static constraint, ASTRA models local training as a dynamic mentorship process. Hence, the local model (Student) learns from its private data while periodically being guided by the global model (Teacher) to prevent semantic divergence.

However, we argue that parameter proximity does not guarantee semantic preservation, especially in deep non-convex networks. ASTRA introduces Dual-Space Regularization, which simultaneously constrains on two distinct levels. Structurally, it enforces Geometric Stability in the parameter space via a proximal term ($||w_k - w^t||$) that limits the physical divergence of weights from the global initialization. Simultaneously, it enforces Semantic Consistency in the function space via Knowledge Distillation. This secondary constraint ensures that even if the weights shift to accommodate local data, the decision boundaries (output logits) remain aligned with the global teacher's consolidated knowledge.

While Dual-Space Regularization ensures robustness, it typically incurs high computational costs due to the need for a secondary Teacher model. To solve this, ASTRA implements Curriculum-Based Periodic Correction. This temporal strategy dynamically adjusts the regularization intensity. This time is (tR_{boot}), which defines the bootcamp or period of a fully active teacher. The method uses a "Foundation Phase" ($t \leq R_{boot}$), during which the teacher remains continuously active to enforce a rigid trajectory during the volatile early rounds. Once geometric stability is established, the system transitions to a "Mentorship Phase," where the teacher is activated only periodically (every k-th round). That reduces the total floating-point operations (FLOPs) by ignoring the teacher in the majority of rounds during the "Mentorship Phase", thereby balancing high accuracy with edge-device efficiency.

We consider an FL scenario involving a set of devices $\mathcal{P} = \{p_1, \ldots, p_K\}$ that receive the global model from the central server over the communication rounds, participate in training using their local data, and exchange model parameters without exposing raw sensing data. In a typical round t, the central server selects a subset of clients $\mathcal{C}_t \subset \mathcal{P}$ to participate in the training round. Each client k holds a private local dataset $\mathcal{D}_k$, where samples (x, y) follow a local distribution P_k. The global objective is to minimize the weighted average loss, as shown in Eq. 1. We represent $F_k(w)$ as the empirical local loss function.

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} F_k(w) \tag{1}$$

However, in realistic scenarios with Non-IID distributions ($P_i \neq P_j$), optimizing $F_k(w)$ in isolation biases local weights towards client-specific optima. This phenomenon, known as client drift, causes local updates to diverge from the global objective w^*, degrading the aggregated model performance.

Figure 1 illustrates the ASTRA architecture, which consists of three zones: i) the top zone contains the server layer, which hosts the global model and the curriculum scheduler method; ii) the middle layer visualizes the transferring parameters between the global model and the client model; iii) the bottom layer shows the process of client training and how it deals with the dataset to train the local model.

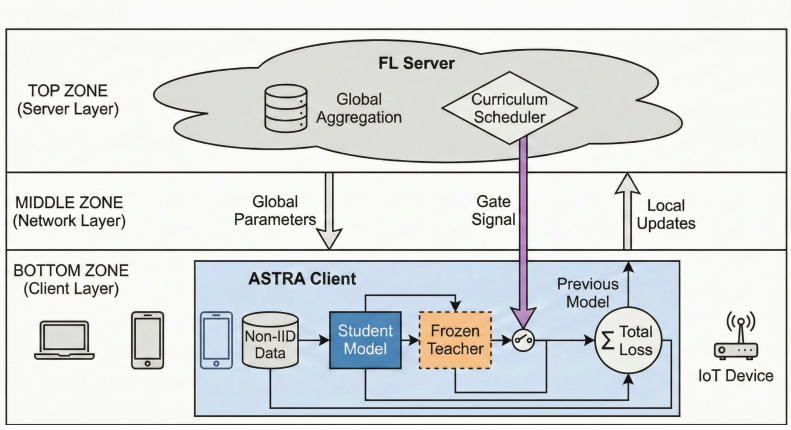

Figure 1. ASTRA Federated Learning Architecture.

The local training workflow proceeds as follows: upon receiving the global model w^t, the client creates a reference copy $w_{teach} \leftarrow w^t$. The active student model w_k then minimizes the hybrid loss function $\mathcal{L}_{ASTRA}$, as shown in Eq. 2. The first term, $\mathcal{L}_{CE}$, refers to the standard Cross-Entropy loss, which drives the optimization of the model's performance on the client's private dataset. The second term (Proximal Regularization) imposes a geometric constraint (controlled by μ) penalizing the Euclidean distance between the evolving local weights w_k and the initial global state w^t. The third term imposes a semantic constraint via Self-Distillation, minimizing the Kullback-Leibler (KL) Divergence, aligning the student's logits (z_s) with the Teacher's logits (z_t). The temperature T softens the probability distributions, facilitating the transfer of inter-class relationship knowledge, also known as dark knowledge. At the same time, the dynamic coefficient α_t regulates the intensity of this guidance according to the curriculum schedule.

$$\mathcal{L}_{ASTRA}(w_k) = \mathcal{L}_{CE}(w_k) + \underbrace{\frac{\mu}{2}||w_k - w^t||^2}_{\text{Geometric Constraint}} + \underbrace{\alpha_t T^2 \mathcal{KL}\left(\sigma\left(\frac{z_s}{T}\right), \sigma\left(\frac{z_t}{T}\right)\right)}_{\text{Semantic Constraint}} \quad (2)$$

Where $z_s = f_{w_k}(x)$ and $z_t = f_{w_{teach}}(x)$ denote the raw output logits (pre-softmax activations) of the student and teacher models, respectively, and $\sigma(\cdot)$ represents the softmax function that converts logits into probability distributions. The temperature parameter T controls the smoothness of probability distributions in knowledge distillation. Following standard KD practices [Hinton et al. 2015]. We set $T = 3.0$ to soften the Teacher's output, amplifying the importance of low-probability classes (dark knowledge). Higher temperatures produce softer distributions, facilitating transfer of inter-class relationships rather than hard labels. We empirically validated this choice through preliminary experiments comparing $T \in \{1, 2, 3, 4, 5\}$, where $T = 3$ balanced convergence speed and final accuracy.

ASTRA operates simultaneously in the parameter space (weights) and the function space (outputs), which are complementary and not redundant. While the geometric constraint (FedProx) keeps weights close in magnitude, two models can be geometrically similar yet produce divergent decision boundaries due to heterogeneous data. The semantic constraint bridges this gap by explicitly requiring the Student's predictions to align with the global consensus, thereby correcting drift that geometric constraints alone might miss.

ASTRA employs a curriculum-based schedule to balance efficiency and model stability, using continuous distillation that increases the computational cost of local training, as shown in Figure 2. In the early rounds ($t \leq R_{boot}$), the Teacher is always active to prevent unstable updates and early divergence under heterogeneous data. After this stage, the Teacher activates only periodically, every k-th round. That allows most training rounds to proceed without distillation, reducing overhead while still providing regular semantic corrections.

We employ the Curriculum Scheduling function to dynamically regulate the Teacher model's influence, as defined in Eq. 3. This function linearly decays the distillation coefficient α_t from its initial value α_{init}, ensuring strong semantic guidance during the volatile early rounds and gradually reducing the constraints to allow for local specialization to enforce two distinct operational modes, as shown in Eq. 3. First, during the

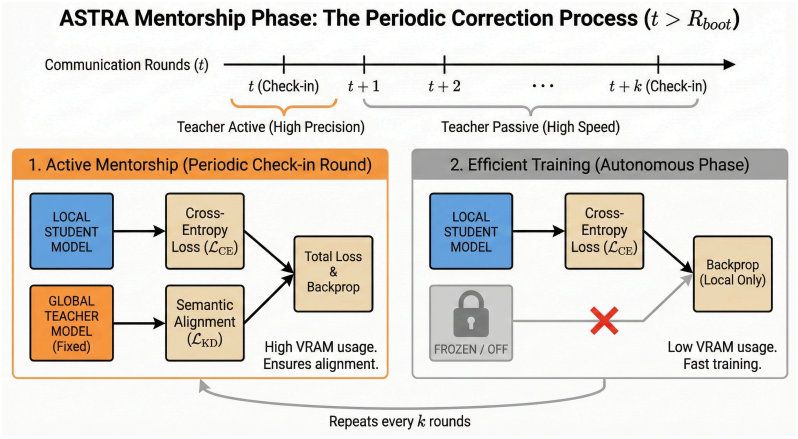

Figure 2. Operational workflow of the ASTRA Mentorship Phase.

initial Phase ($t \leq R_{boot}$), the system maintains a continuous teacher signal to align with the global trajectory strictly. Second, in the Mentorship Phase, the method activates the teacher only periodically every k-th round. In all intermediate rounds, the system otherwise sets $\alpha_t = 0$, granting the student model full autonomy to optimize for local data specificity without external regularization.

$$\alpha_t = \begin{cases} \alpha_{init} \times (1 - \frac{t}{R_{max}}) & \text{if } t \leq R_{boot} \text{ or } t \pmod k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Algorithm 1 details the execution flow of the ASTRA method, integrating the Curriculum-Based Periodic Correction strategy. The process begins with the server determining the global training state (Step One). Specifically, the algorithm checks if the system is in the Foundation Phase (early stabilization) or an active round of the Mentorship Phase. If either condition is met, the system sets the Teacher status to active. Otherwise, the system defaults to Autonomous Mode, bypassing the Teacher instantiation entirely (Step Two) to reduce computational overhead to the level of standard FedAvg.

Inside the local optimization loop, the objective function enforces Dual-Space Regularization. The Student model always computes the Geometric Constraint (Proximal term) and Cross-Entropy loss to ensure physical stability. However, the Semantic Constraint ($\mathcal{L}_{KD}$) is conditionally activated. If the Teacher is present, the method computes the Teacher logits z_t and scales the divergence by the decaying factor α_t (Step 3); conversely, in Autonomous Mode, $\mathcal{L}_{KD}$ is set to zero. This selective activation ensures that ASTRA enforces rigid semantic alignment only when necessary, preventing client drift without incurring the continuous cost of dual-model forward passes. Finally, the server aggregates the updated weights using standard FedAvg (Step 4).

4. Evaluation

This section presents the simulation setup used to evaluate the performance and efficiency of ASTRA compared to baseline methods. We first describe the simulated FL scenario, including the underlying method, dataset characteristics, and simulation parameters. Subsequently, we present the results obtained, focusing on the metrics of accuracy, loss convergence, and computational overhead.

Algorithm 1 ASTRA

Require: K clients, learning rate η, proximal μ, initial KD α_{init}, Rounds R_{max}, Bootcamp R_{boot}, Interval k
Ensure: Global Model w^{final}

Initialize w^0
for round $t = 0$ **to** $R_{max} - 1$ **do**
  Server broadcasts w^t to selected clients C_t
  // Step 1: Determine Curriculum State
  if $t \leq R_{boot}$ **then**
    Foundation Phase: Teacher Always Active
    $Active \leftarrow$ **true**
  else
    Mentorship Phase: Periodic Check
    $Active \leftarrow (t \pmod k == 0)$
  end if
  for client $k \in C_t$ **in parallel do**
    $w_k \leftarrow w^t$
    // Step 2: Teacher Instantiation
    if $Active$ **then**
      $w_{teach} \leftarrow w^t$ **(Frozen Copy)**
      Calculate α_t (Linear Decay, Eq. 3)
    else
      Autonomous Mode: No Teacher Overhead
      $w_{teach} \leftarrow$ NULL; $\alpha_t \leftarrow 0$
    end if
    // Step 3: Local Optimization
    for batch $(x, y) \in \mathcal{D}_k$ **do**
      Compute student logits: $z_s = f_{w_k}(x)$ (raw outputs)
      if $Active$ **then**
        Compute teacher logits: $z_t = f_{w_{teach}}(x)$ (frozen)
        $\mathcal{L}_{KD} = \alpha_t T^2 \mathcal{KL}(\sigma(z_s/T), \sigma(z_t/T))$
      else
        $\mathcal{L}_{KD} = 0$
      end if
      $\mathcal{L}_{total} = \mathcal{L}_{CE} + \frac{\mu}{2}||w_k - w^t||^2 + \mathcal{L}_{KD}$
      Update: $w_k \leftarrow w_k - \eta \nabla \mathcal{L}_{total}$
    end for
    Send w_k to server
  end for
  // Step 4: Aggregation
  $w^{t+1} \leftarrow \sum_{k \in C_t} \frac{|D_k|}{\sum_{j \in C_t} |D_j|} w_k$ (FedAvg)
end for

4.1. Simulation Setup

The simulation environment is built upon FL-REST[1], which is an open-source tool designed to emulate the volatility of edge networks that rely on PyTorch for local model

[1]FL-REST: `https://github.com/twinte/fl-rest`

training. Our simulation code is available on GitHub[2]. We repeated all experiments across three independent trials using fixed random seeds $\{10, 42, 999\}$ for both data partitioning and weight initialization to enforce reproducibility. The simulation considers a heterogeneous cluster with $K = 20$ active clients, evenly distributed to mimic a diverse network. Specifically, 50% of clients are designated as "High-Performance" (simulated with NVIDIA GPU acceleration), while the remaining 50% are "Low-Resource" stragglers restricted to CPU-only execution.

Table 2 shows the main parameters used in our simulation, such as the dataset and the model used. We considered CIFAR-10 dataset[3], which comprises 60,000 labeled RGB images of size 32×32 pixels uniformly distributed across 10 distinct classes. We divided the training data among clients using a Dirichlet distribution $Dir(\alpha)$ over label ratios to build a realistic Non-IID setting. We set the concentration parameter $\alpha = 0.1$, creating an extreme skew in which most clients hold samples from only 2 or 3 distinct classes. This setup rigorously tests the model's ability to overcome client drift. We employ a compact Convolutional Neural Network (CNN) optimized for edge deployment, consisting of two convolutional blocks and three fully connected layers (≈62k parameters). This lightweight design ensures compatibility with the memory constraints of IoT-class devices.

Table 2. Simulation Parameters

Parameter	**Configuration / Value**
Dataset	CIFAR-10 (Non-IID, $\alpha = 0.1$)
Model	SimpleCNN (2 Conv, 3 FC, ≈62k params)
Total Clients	20 (10 GPU / 10 CPU)
Clients per Round	5 (Random Selection)
Global Rounds	50
Local Epochs	5
Optimizer	SGD (Learning Rate: $\eta = 0.01$)
Curriculum Schedule	($R_{boot} = 10$), Interval=2
ASTRA Hyperparams	$\mu = 0.01$, $\alpha_{init} = 0.2$, $T = 3.0$

The hyperparameters were selected based on preliminary grid search experiments balancing convergence stability and computational efficiency. The proximal regularization strength $\mu = 0.01$ follows FedProx recommendations for lightweight constraint without over-restricting local adaptation. The initial distillation weight $\alpha_{init} = 0.2$ provides moderate semantic guidance while allowing local task learning. The Foundation Phase duration $R_{boot} = 10$ (20% of total rounds) stabilizes the volatile early training phase when models diverge rapidly from random initialization. The correction interval $k = 2$ balances frequent semantic alignment with computational savings, activating the teacher in approximately 60% of post-foundation rounds.

We consider three performance metrics to evaluate the methods, namely accuracy, loss, and computational efficiency. The global model accuracy is computed as the proportion of correct predictions among total predictions, quantifying generalization capabilities in Non-IID scenarios for each dataset client after aggregation. The global loss quantifies

[2]ASTRA: `https://github.com/IcsPingu/ASTRA-FL`

[3]CIFAR-10 Dataset: `https://www.cs.toronto.edu/~kriz/cifar.html`

the uncertainty or disorder between the predicted class probabilities and the actual labels, helping analyze convergence stability and the effectiveness of mitigating client drift. A lower entropy loss indicates better model performance, since it implies that the predicted probabilities are closer to the actual labels. Finally, Computational efficiency ($\mathcal{T}_{round}$) can be defined as the average wall-clock time per round. This last metric explicitly accounts for the "straggler effect" by measuring the time determined by the slowest client in the selected subset, capturing both local computation and network transmission delays.

Table 3 summarizes the implemented methods to cover the spectrum of drift mitigation in our evaluation, detailing how they mitigate client drift and the strategies employed. FedAvg [McMahan et al. 2017] is the standard FL algorithm and serves as a baseline in our evaluation, as it lacks explicit regularization mechanisms for addressing statistical divergence. FedAvg serves as the unregularized lower bound, illustrating the severity of performance degradation when heterogeneity is not addressed.

For mitigation strategies, we first evaluate FedProx [Li et al. 2020], which represents the structural regularization approach operating in the Parameter Space. By adding a proximal term $\frac{\mu}{2}||w - w^t||^2$ to the local objective, it theoretically constrains how far a client can diverge from the global model. We include it to assess whether rigid geometric constraints alone are sufficient to handle the extreme heterogeneity ($\alpha = 0.1$) of our scenario.

We also considered MOON [Li et al. 2021], which is a primary competitor representing semantic regularization in the Representation Space. MOON considers contrastive losses to align local feature representations with the global model. MOON allows us to quantify the specific benefits of ASTRA's student-teacher distillation over contrastive learning objectives.

Table 3. Comparison of Evaluated Methods and Drift Mitigation Strategies

Method	Regularization Type	Target Space	Mechanism	Teacher Signal
FedAvg [McMahan et al. 2017]	None	-	Weight Averaging	None
FedProx [Li et al. 2020]	Structural	Parameter Space	Proximal Term (L_2)	Static (Global Weights)
MOON [Li et al. 2021]	Semantic	Representation Space	Contrastive Loss	Representations (z)
ASTRA (Ours)	**Hybrid (Dual)**	**Dual Space (Param + Func)**	**Distillation + Proximal**	**Dynamic (Curriculum)**

4.2. Results

Figure 3 presents the comparative analysis of accuracy for the methods evaluated in the highly heterogeneous CIFAR-10 dataset ($\alpha = 0.1$). By analyzing the results of Figure 3, we observe that while FedAvg achieves a final accuracy of 45.0%, its trajectory is characterized by significant instability. In particular, FedProx suffers a catastrophic collapse in round 45, dropping from 40.8% to 30.9%. This is a classic symptom of the client drift phenomenon, where the proximal term fails to reconcile divergent local updates. ASTRA successfully avoids this failure mode, maintaining a stable accuracy plateau (≈41-44%) and proving that Curriculum-Based Periodic Correction effectively regularizes training without the rigidity of FedProx.

Figure 4 presents the comparative analysis of losses for the methods evaluated in the highly heterogeneous CIFAR-10 dataset ($\alpha = 0.1$). As observed, ASTRA consistently maintains the lowest training loss throughout the optimization process (min: 1.57 vs FedAvg: 1.86), demonstrating superior semantic alignment. Specifically, in the early

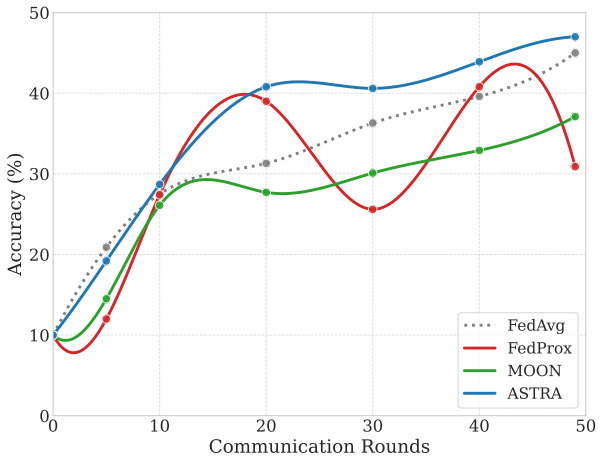

Figure 3. Accuracy over the communication rounds for the evaluated methods

training stages, ASTRA exhibits a steeper convergence trajectory compared to the baselines, driven by the rigid guidance of the Foundation Phase which prevents the initial divergence common in random initializations. While FedAvg and FedProx display high-variance oscillations due to the conflicting gradients of heterogeneous clients, ASTRA's curve remains smooth and monotonic. This stability indicates that the periodic teacher corrections effectively dampen client drift, preventing the local models from "forgetting" the global consensus during their private updates. The lower loss achieved by ASTRA (1.57 vs FedAvg 1.86) directly translates to better calibrated predictions and reduced uncertainty in class assignments. Moreover, the monotonic convergence without oscillations suggests that the curriculum mechanism successfully prevents the accumulation of conflicting gradients that would otherwise manifest as loss spikes. The Foundation Phase establishes a stable optimization trajectory, while the Mentorship Phase maintains this trajectory through periodic corrections, effectively balancing local adaptation with global consistency. This contrasts sharply with FedProx's loss trajectory, which exhibits increasing variance in later rounds, culminating in the catastrophic divergence observed in Figure 3 at round 45.

For validation of the consistency of the method, we extended the evaluation to include moderate heterogeneity settings ($\alpha = 0.3$ and $\alpha = 0.5$). Figure 5 shows all methods accuracy when being simulated in different heterogenity patterns. Under extreme heterogeneity ($\alpha = 0.1$), all methods struggle due to the severe lack of class overlap, yet ASTRA still secures the highest accuracy of 47.0% compared to FedAvg (45.0%) and a big advantage on FedProx (30.9%). However, the advantages of the ASTRA method are not only applicable on Severe Non-IID distributions as the lead in accuracy become more pronounced as the data distribution relaxes. At $\alpha = 0.5$, ASTRA achieves 65.80%, significantly outperforming FedAvg (58.00%) by +7.8% and MOON (63.50%) by +2.3%. This result highlights that while structural methods like FedProx stabilize training, ASTRA's semantic distillation scales better by effectively leveraging the improved local data quality.

Table 4 presents the average training time per run. Despite maintaining a teacher model, ASTRA achieves only 2.3% overhead compared to FedAvg (1132s vs 1107s), outperforming both MOON (7.5%, 1190s) and FedProx (19.9%, 1327s). This efficiency

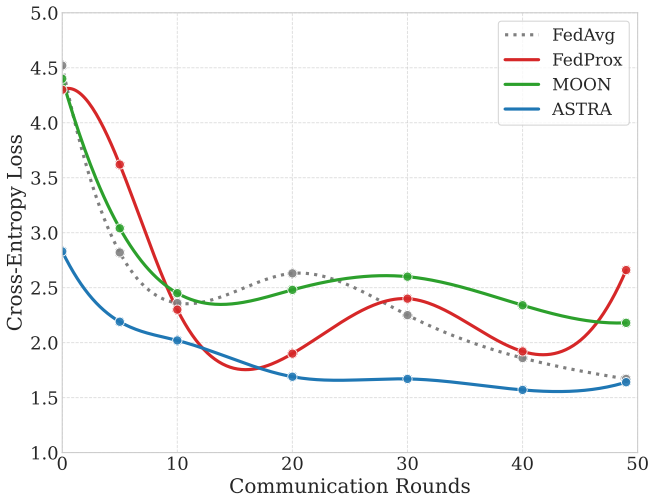

Figure 4. Loss over the communication rounds for the evaluated methods

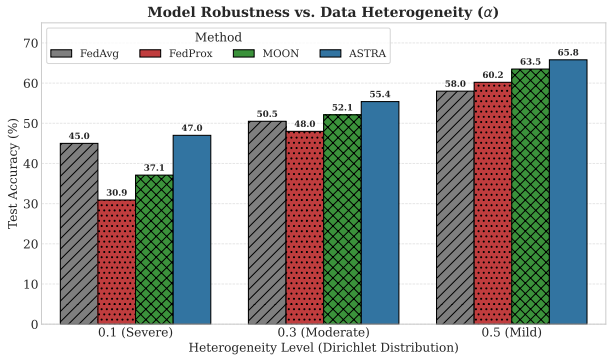

Figure 5. Impact of Data Heterogeneity on the accuracy for the evaluated methods

stems from ASTRA's curriculum mechanism, which activates the teacher only in the Foundation Phase ($R_{boot} = 10$ rounds) and periodically thereafter (every $k = 2$ rounds in the Mentorship Phase). In non-active rounds, ASTRA operates identically to FedProx with only the proximal term, avoiding the computational cost of teacher forward passes and KL divergence calculations. The amortized teacher cost across all rounds results in minimal overhead, as the expensive distillation operation is strategically limited to critical correction points rather than applied continuously. In contrast, FedProx's higher overhead can be attributed to the proximal term being computed at every gradient step across all local epochs, which accumulates over 5 local epochs and multiple batches per round. While each individual proximal computation is inexpensive, the cumulative cost over thousands of local iterations becomes significant. ASTRA mitigates this by balancing the proximal term (continuous, lightweight) with periodic distillation (intermittent, heavier), achieving superior drift mitigation with lower aggregate computational cost.

Finally, we analyze whether the proposed curriculum mechanism explicitly reduces client drift, the phenomenon in which heterogeneous local updates cause global divergence. Our results provide compelling evidence for this mitigation, particularly when comparing the stability of ASTRA with structural baselines. As observed in Figure 3,

Table 4. Average Training Time per Run (Seconds) and Computational Overhead.

Method	Time (s)	Overhead vs FedAvg	Efficiency Rank
FedAvg	1107 s	0.0% (Baseline)	1st
ASTRA (Ours)	**1132 s**	**+2.3%**	**2nd**
MOON	1190 s	+7.5%	3rd
FedProx	1327 s	+19.9%	4th

FedProx suffered a catastrophic performance collapse in Round 45, dropping from 40.8% to 30.9% accuracy. This failure highlights the limitations of operating strictly in Parameter Space: the proximal term used by FedProx blindly penalizes weight deviations, creating a rigid constraint that, under severe heterogeneity, conflicts with the local learning objective. In contrast, ASTRA maintained a monotonic convergence trajectory because it operates in function space. By regularizing the output predictions (via Knowledge Distillation) rather than the weights themselves, ASTRA allows the local models sufficient flexibility to learn specific local features, while the Teacher signal ensures the decision boundaries remain semantically aligned with the global consensus. This alignment is further corroborated by the loss landscape in Figure 4, where ASTRA achieves a significantly lower and smoother loss ($\mathcal{L} \approx 1.57$) compared to the high-variance convergence of FedAvg ($\mathcal{L} \approx 1.67$). The Curriculum Gate effectively acts as a periodic correction mechanism, preventing small local drifts from accumulating into the large-scale divergence observed in the baselines.

5. Conclusion

In this paper, we introduced the Adaptive Student-Teacher for Robust Aggregation (ASTRA) method designed to enhance model convergence in Federated Learning environments, especially in non-IID and resource-constrained scenarios. ASTRA improves the stability of FL training by synergizing geometric regularization with a curriculum-based self-distillation mechanism to ensure the semantic consistency of the model. Our approach achieves a final accuracy of 65.80% in moderate heterogeneity settings ($\alpha = 0.5$), surpassing FedAvg by +7.8% and the semantic baseline MOON by +2.3%.

For example, in drift-inducing conditions, the robustness of the model was evaluated by maintaining monotonic convergence while structural baselines suffered a performance collapse to 30.9% accuracy. Crucially, these gains are achieved with minimal computational cost. By dynamically toggling the teacher branch via the Curriculum Gate, ASTRA reduces the training overhead to just 2.3% compared to FedAvg. This effectively demonstrates that high-performance Knowledge Distillation is viable for edge devices when the distillation schedule is strategically managed.

Acknowledgment

This research was partially sponsored by CNPq grant 404186/2021-1, CAPES, the Brazilian Ministry of Science, Technology, and Innovations, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex, and published under Arquitetura Cognitiva (Phase 3), DOU 01245.003479/2024-10. This research is part of the INCT of Intelligent Communications Networks and the Internet of Things (ICoNIoT) funded by CNPq (proc. 405940/2022-0) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) Finance Code 88887.954253/2024-00.

References

Amiri, S. et al. (2024). Balancing privacy and performance in federated learning: A systematic literature review on methods and metrics. *Journal of Parallel and Distributed Computing*.

Cooray, L., Sendanayake, J., Vithanaarachchi, P., and Priyadarshana, Y. H. P. P. (2025). Deep federated learning: a systematic review of methods, applications, and challenges. *Frontiers in Computer Science*, Volume 7 - 2025.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., and Kim, S.-L. (2018). Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*.

Jimenez G., D. M., Solans, D., Heikkilä, M. A., Vitaletti, A., Kourtellis, N., Anagnostopoulos, A., and Chatzigiannakis, I. (2024). Non-IID data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. *arXiv preprint arXiv:2411.12377*.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR.

Li, Q., He, B., and Song, D. (2021). Model-contrastive federated learning. In *conference on computer vision and pattern recognition*, pages 10713–10722.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Machine learning and systems*, 2:429–450.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282.

Mhou, K. and Senekane, M. (2026). HAPI—FedProx: Heterogeneity—aware adaptive proximal optimization for federated learning. *Springer Nature Link*.

Mora, A. et al. (2024). Knowledge distillation in federated learning: a practical guide. In *33rd International Joint Conference on Artificial Intelligence*.

Qin, L. et al. (2024). Knowledge distillation in federated learning: a survey on long lasting challenges and new solutions. *arXiv preprint arXiv:2406.10861*.

Rodríguez-Barroso, N. et al. (2024). An overview of implementing security and privacy in federated learning. *Artificial Intelligence Review*, 57.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. (2021). A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249.

Yan, Y., Feng, C.-M., Ye, M., Zuo, W., Li, P., Goh, R. S. M., Zhu, L., and Chen, C. (2023). Rethinking client drift in federated learning: A logit perspective. *arXiv preprint arXiv:2308.10162*.