Federated Learning by Utility-Constrained Stochastic Aggregation for Improving Rational Participation

M Yashwanth[1] **Arunabh Singh**[1] **Ashok Nayak**[2] **Saikiran Bulusu**[3] **Anirban Chakraborty**[1]

[1]**Indian Institute of Science, Bangalore** [2]**Indian Institute of Technology Bombay** [3]**IIIT Hyderabad**

yashwanthm@iisc.ac.in arunabhsingh25@gmail.com 22b0455@iitb.ac.in
saikiran.bulusu@iiit.ac.in anirban@iisc.ac.in

Abstract

Federated Learning (FL) algorithms implicitly assume that clients passively comply with server-side orchestration, sharing local model updates upon the server's request. However, this overlooks an important aspect of real-world cross-silo environments: clients are often rational agents who may prioritize their own interests, such as local model performance, over that of the global model. In settings with significant statistical heterogeneity, rational clients may opt out of the federation if the perceived benefits of collaboration fail to meet their local utility thresholds. Such attrition degrades global model performance and can lead to the collapse of the federated training process. To avoid such a scenario, in this work, we introduce *FedUCA (Federated Learning by Utility-Constrained Stochastic Aggregation for Improving Rational Participation)*, a non-monetary framework that establishes a non-payment channel through which clients communicate their willingness to participate via their utilities. The server acts as an optimizer, identifying aggregation parameters that satisfy clients' individual rationality (IR) constraints to sustain the global model performance. FedUCA achieves this through a stochastic aggregation mechanism over Dirichlet-sampled mixture strategies, which induces an implicit utility surplus that helps satisfy participation constraints. Supported by theory and extensive experiments[1], FedUCA's focus on participation feasibility improves client retention and global model performance.

1 Introduction

Federated Learning (FL) enables multiple data holders to collaboratively train models without sharing raw data [28]. In the standard paradigm, a central server coordinates iterative rounds, aggregating local updates to learn a global model across decentralized data silos. This framework is essential in settings constrained by privacy, data sovereignty, and institutional restrictions. Most prior work focuses on improving statistical efficiency, communication, and robustness [1, 14, 18], while implicitly assuming passive client participation. This assumption is reasonable in cross-device FL, but breaks down in cross-silo settings, where participation incurs non-trivial operational costs.

In cross-silo FL, participants such as hospitals, research labs, and enterprises face costs related to computation, networking, compliance, and infrastructure. As a result, participation becomes a strategic decision based on whether collaboration improves local performance. Empirical deployments reflect this behavior: clinical FL systems prioritize intermediate utility and deployment readiness [7], while platforms such as Fed-BioMed emphasize governance and integration constraints [6]. Similar considerations arise in industrial settings, including energy and infrastructure costs [4].

This cost-driven behavior motivates viewing cross-silo FL through a mechanism-design lens. In settings where external payments are infeasible or undesirable, the primary basis for sustained

[1]The implementation is provided here.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

collaboration is the value of the model itself. We therefore model each client as a rational agent that participates when the perceived utility of the global model exceeds its local baseline. Sustaining the federation then requires satisfying individual rationality (IR) constraints across clients. Thus, each client evaluates a simple, performance-based proposition: they will participate if and only if their perceived utility from the global model strictly eclipses their local baseline. However, under the severe statistical heterogeneity typical of cross-silo data, a single deterministic aggregation vector often cannot simultaneously satisfy the diverse IR thresholds of all participants, making attrition likely.

Breaking this deadlock requires the right model of client preferences. As detailed in Section 3.2, we model client utility as a concave function of its assigned aggregation weight. Concavity is the standard primitive in mechanism design [32], capturing risk aversion and diminishing marginal returns: a client benefits from increased influence up to an optimal informational threshold, beyond which over-representation degrades the model's generalization. We adopt this primitive directly and verify empirically that it holds in the federated learning setting.

This inherent concavity allows us to introduce a stochastic aggregation framework that exploits *Jensen's gap*, the difference between the utility of the expected aggregation and the expected utility across a distribution of weights. By optimizing a set of candidate aggregation strategies evaluated under a Dirichlet-sampled mixture, the server uses this gap to generate a utility surplus that acts as a safety margin. The magnitude of this induced surplus is structurally regulated via the Dirichlet concentration parameter δ, sustaining participation across heterogeneous clients.

Our key contributions are

- We introduce **FedUCA**, a federated aggregation framework that establishes a non-payment channel through which clients communicate their desire to participate via their utilities. The server acts as an optimizer, tasked with identifying aggregation parameters that satisfy the individual rationality (IR) demands communicated through this channel.
- We introduce a stochastic aggregation mechanism with Dirichlet-sampled mixture weights over multiple aggregation strategies, and show that optimizing the resulting Jensen lower-bound induces an implicit utility surplus under concave-modeled client utilities. This surplus acts as a buffer that helps satisfy participation constraints under heterogeneity, and is regulated by the Dirichlet concentration parameter δ. We validate our theoretical findings through empirical evaluations on standard datasets.
- We provide a convergence characterization for this endogenous system where participation sets and aggregation weights co-evolve. The mechanism-induced aggregation discrepancy is the mathematical footprint of rational participation management and is directly controlled by the Dirichlet concentration parameter δ, connecting our Server Optimization design and Convergence.

2 Related Work

Federated Learning (FL) was introduced via FedAvg [28], which assumes fully compliant clients. Subsequent work improves robustness to data heterogeneity, including FedProx [24], SCAFFOLD [18], and FedDyn [1]. While these methods address statistical drift, they implicitly assume passive participation. Fairness and contribution-aware approaches have also been studied. q-FFL [25] reweights client losses to improve equity, while Data Shapley [12] and its extensions to FL [35] quantify data contribution via cooperative game theory. A separate line of work models rational participation. IncFL [5] incentivizes clients based on performance gains, but tends to favor clients whose local models align with the global model, potentially degrading performance under heterogeneity.

Other approaches incorporate explicit incentives. Model-quality rewards [21] and sequential decision-based aggregation [13] aim to improve participation and fairness. Economic formulations include Stackelberg frameworks [33], pricing-based methods [27], and contract-theoretic approaches [17]. Broader surveys are provided in [40]. Auction-based methods [41] and hierarchical game formulations [36] further extend this paradigm. Recent work also considers dynamic and learning-based incentives. DaringFed [38] uses Bayesian persuasion with monetary rewards, while accuracy-shaping mechanisms [19] penalize free-riders through degraded model quality. Across these approaches,

randomness is primarily used for client selection rather than as a mechanism for shaping client utilities.

Table 1: Comparison of incentive mechanisms in federated learning.

Method	Uses Monetary Incentives	Requires Equilibrium Modeling	Handles Rational Dropout	Direct Aggregation Optimization
Contract-based [17]	✓	✓	✓	Indirect
Pricing-based [27]	✓	✓	✓	Indirect
DaringFed [38]	✓	✓	✓	Pricing-driven
IncFL [5]	×	×	✓	Similarity-based
FedUCA (Ours)	×	×	✓	✓

Table 1 summarizes key differences between FedUCA and existing approaches. FedUCA adopts a non-monetary, performance-driven formulation and does not rely on equilibrium modeling. Instead, it encodes rationality as performance-based constraints and directly optimizes aggregation weights to satisfy participation requirements. Unlike pricing and contract-based approaches that depend on external rewards, FedUCA sustains participation purely through model utility. Accordingly, we focus on baselines within the same non-monetary setting, while noting that monetary and aggregation-based approaches are complementary.

3 Method

FedUCA uses stochastic aggregation to exploit concavity-induced Jensen surplus in client utilities, enabling participation beyond deterministic aggregation. We first show how aggregation weights define client utility and rational participation (Sections 3.1, 3.2), then formulate the resulting constrained optimization problem (Section 3.3), describe the client-server protocol (Section 3.4), and provide theoretical guarantees. We consider a standard FL setting with N rational clients and a global objective $f(\theta) = \sum_{k=1}^{N} \beta_k f_k(\theta)$, where f_k is the local objective at client k and $\{\beta_k\}_{k=1}^{N}$ is a simplex. Training proceeds in rounds with server-side aggregation of client updates.

3.1 Rational Participation

In standard FL, the server broadcasts the aggregated global model $\boldsymbol{\theta}^t = \sum_{k=1}^{N} \alpha_k^{t-1} \boldsymbol{\theta}_k^{t-1}$. A rational client evaluates this model against its current local model using a validation-based rationality check:

$$acc(\boldsymbol{\theta}^t, \mathcal{D}_k^{\text{val}}) > acc(\boldsymbol{\theta}_k^{t-1}, \mathcal{D}_k^{\text{val}}), \tag{1}$$

where $acc(\cdot)$ denotes accuracy and $\mathcal{D}_k^{\text{val}}$ is the held-out validation set of client k (one can use loss as well, instead of accuracy). If Eq. 1 holds, the client accepts $\boldsymbol{\theta}^t$, trains from it, and shares an update in the next round; otherwise, it rejects the global model and does not participate. However, this binary accept/reject mechanism is a "black box" to the server. It has no way to control this check as it is agnostic to the clients' data and may lead to federation collapse. Since the server aggregates client models using simplex weights, FedUCA makes this rationality check weight-aware by sharing the aggregation weight α_k^{t-1} with client k. The client uses this information to estimate a utility function $u_k(r; \zeta_k)$ over possible aggregation weights, and reports the fitted utility parameters ζ_k together with a minimum acceptable threshold γ_k. The threshold γ_k captures the client's individual rationality requirement: client k participates only if

$$u_k(r_k; \zeta_k) > \gamma_k. \tag{2}$$

By collecting these utilities and thresholds, the server observes heterogeneous participation requirements and can optimize aggregation weights to satisfy individual rationality constraints.

3.2 Estimation of Clients' Utility Functions

The assumption that client utility is concave in its aggregation weight is standard in the mechanism design and contract theory literature [32]. Concavity captures the economic property of diminishing marginal returns: a client benefits from increased weight up to its optimal contribution level, beyond which additional weight reduces the informational benefit drawn from the rest of the federation. This is the natural analogue of risk aversion in expected utility theory [32] and is the standard primitive in incentive-compatible mechanism design. We adopt this assumption throughout, and verify empirically in Appendix A.4 that it holds in the federated deep learning setting studied here.

If the client k knows the information of the aggregation weight α_k^{t-1} used in aggregating the client models, it can exploit this information to design the utility as a function of the aggregation weight r. We now estimate the utility function of the client based on the global performance on its held-out dataset. The server sends the value α_k^{t-1} for the client k based on the distribution $\mathbf{s}$ along with the global model $\boldsymbol{\theta}^t$. Client k estimates its impact on the aggregated global model by computing the model that would result if its own update were excluded, denoted as $\boldsymbol{\theta}_{-k}^t$. Since the client knows the exact weight, it uses it to remove its own contribution.

$$\boldsymbol{\theta}_{-k}^t = \frac{\boldsymbol{\theta}^t - \alpha_k^{t-1}\boldsymbol{\theta}_k^{t-1}}{1 - \alpha_k^{t-1}} \tag{3}$$

where $\boldsymbol{\theta}_k^{t-1}$ is the client model in round $t-1$. [2] The client now creates a model $\tilde{\boldsymbol{\theta}}_k(r)$ with an aggregation parameter r as $\tilde{\boldsymbol{\theta}}_k(r) = (1-r)\boldsymbol{\theta}_{-k}^t + r\boldsymbol{\theta}_k^t$, $\tilde{\boldsymbol{\theta}}_k(r)$ can be interpreted as the global model when the aggregation weight given to client k is r. The client uses $\tilde{\boldsymbol{\theta}}_k(r)$ to assess its utility from the server model as a function of its aggregation weight. For a specific value of r, this is computed by evaluating the validation loss/accuracy on the client held-out data, $acc(.)$ denotes the accuracy.

$$\hat{u}_k(r) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_k^{val}}[acc(h(\mathbf{x};\tilde{\boldsymbol{\theta}}_k(r)), y)] \tag{4}$$

$h(x;\tilde{\boldsymbol{\theta}}_k(r))$ denotes the prediction made by client k on data point $\mathbf{x}$ with the model $\tilde{\boldsymbol{\theta}}_k(r)$. We observe that the clients attain best accuracies when $r \in (0,1)$. When r is 0, the accuracy will be lower as the client model is not utilized; on the other hand, when r is 1, the server model is ignored. This motivates us to parameterize the utility curve by ζ_k for each individual client k.

The empirical utility $\hat{u}_k(r)$ is evaluated along the interpolation path between the leave-one-out global model $\boldsymbol{\theta}_{-k}^t$ and the locally updated model $\boldsymbol{\theta}_k^t$. For convex models such as linear logistic regression, linear SVMs with convex surrogate losses, or last-layer fine-tuning with a convex loss, the validation loss along this path is convex in r; hence a loss-based utility, e.g., negative validation loss, is concave. For deep models and accuracy-based utilities, exact concavity is not guaranteed. However, parameter-space interpolations in deep networks are often observed to traverse smooth low-loss regions [11, 23], and we empirically observe smooth, unimodal utility profiles across clients, rounds, and datasets (Appendix A.4).

We model client utility as a concave function of the aggregation weight r, parameterized by ζ_k. The parameters ζ_k are estimated per round from held-out validation data:

$$\zeta_k := \arg\min_{\tilde{\zeta}_k} \mathbb{E}_{r\sim\mathcal{U}(0,1)} \left[\|u_k(r;\tilde{\zeta}_k) - \hat{u}_k(r)\|_2^2\right], \tag{5}$$

where $\mathcal{U}(0,1)$ denotes the uniform distribution. The concavity assumption is made at the modeling level; it characterizes client preferences in the sense of mechanism design, and any concave parametric family is admissible. The fitted parameters ζ_k are transmitted to the server as a lightweight proxy for the client's utility. The fitted utility $u_k^t(r;\zeta_k^t)$ tracks the empirical utility $\hat{u}_k^t(r)$ closely across rounds (Appendix A.4); improvements in the surrogate therefore translate to improvements in realized client utility, ensuring server-side optimization aligns with client-side rationality.

Discussion: This work is situated within the Cross-Silo Federated Learning (FL) paradigm, which involves a limited number of high-capacity clients such as institutions or large corporate entities. While compute, bandwidth, and power can still be practical concerns in this setting, they are not the focus of this work: we do not include them into the proposed utility function, and client participation is driven solely by model-performance utility. We leave the integration of resource-aware costs into the utility model to future work. We aim to satisfy

3.3 Solving Optimization

To maximize the participation of rational clients, the server must move beyond traditional deterministic aggregation. A single weight vector r_k often fails to satisfy the diverse utility thresholds γ_k of all silos

[2] This leave-one-out construction follows the broader counterfactual contribution-analysis perspective used in influence functions and data valuation methods [12, 20, 35]. Unlike these works, which quantify the contribution or value of data sources, we use the leave-one-out model as an anchor for estimating how client k's utility varies with its aggregation weight.

simultaneously. Rather than optimizing over a single, static weight vector r_k, the server generates a set of n_s candidate strategies $\{r_k^j\}_{j=1}^{n_s}$ for $k = 1, \ldots, N$. The influence of these strategies is determined by a simplex vector $\mathbf{s} = \{s_1, s_2, \ldots, s_{n_s}\}$, where s_j represents the probability of selecting the j^{th} strategy. Crucially, in our framework, the distribution $\mathbf{s}$ is not a learned parameter but is kept fixed for a communication round, with values drawn from a Dirichlet distribution characterized by a concentration parameter δ. By fixing this distribution, the server provides a predictable yet diverse set of outcomes. This diversity is essential for exploiting the Jensen's gap; as we show in Proposition 3.2, the concentration parameter δ modulates the mixture geometry of $\mathbf{s}$, which in turn shapes the induced Jensen surplus through the optimized strategy matrix $\mathbf{R}$. The server communicates to client k the expected aggregation weight $\alpha_k = \sum_i s_i r_k^i$, which determines the client's perceived utility. To sustain participation, the server seeks strategies under which each client's expected utility clears its threshold,

$$\sum_{i=1}^{n_s} s_i \, u_k(r_k^i) \geq \gamma_k + \epsilon, \tag{6}$$

where $\epsilon > 0$ is a margin enforcing strict satisfaction. Because the utilities are concave, Jensen's inequality gives $u_k(\sum_{i=1}^{n_s} s_i r_k^i) \geq \sum_{i=1}^{n_s} s_i \, u_k(r_k^i)$, so satisfying Eq. 6 guarantees the realized utility at α_k also clears the threshold.

We formulate the final optimization below via the slack variables $\mathbf{t} \coloneqq \{t_1, ..., t_N\}$.

$$\begin{aligned} \underset{\mathbf{t},\mathbf{R}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{t}\|_2^2 \\ \text{subject to} \quad & \mathbf{R}\mathbf{1}_N = \mathbf{1}_{n_s}, \mathbf{1}_{n_s}^T \mathbf{R} = \frac{n_s}{N}\mathbf{1}_N^T, \mathbf{r}_k \succeq 0, \mathbf{t} \succeq 0 \\ & \sum_{i=1}^{n_s} s_i u_k(r_k^i) \geq \epsilon + \gamma_k - t_k, \forall k. \end{aligned} \tag{7}$$

The above optimization becomes deterministic by setting $n_s = 1$ and ignoring the constraint $\mathbf{1}_{n_s}^T \mathbf{R} = \frac{n_s}{N}\mathbf{1}_N^T$. The constraint $\mathbf{1}_{n_s}^T \mathbf{R} = \frac{n_s}{N}\mathbf{1}_N^T$ ensures that rows of $\mathbf{R}$ are non-zero, if $N = n_s$ it implies $\mathbf{R}$ is doubly stochastic. The server recomputes $\mathbf{t}$ and $\mathbf{R}$ at every round; however, we suppress the temporal index to maintain readability.

Fallback Mechanism. If the optimization problem yields an outcome in which no client satisfies the individual rationality constraint, in a given round t, the server invokes a fallback strategy. It uniformly samples a client k and constructs a strategy that prioritizes this client by setting $r_k^i = r_k$ for all strategies i, where r_k is chosen such that $u_k(r_k; \zeta_k) > \gamma_k^t$. The remaining aggregation mass is distributed among the other clients so as to satisfy the aggregation constraints of $\mathbf{R}$. This guarantees that at least one client satisfies its individual rationality constraint and participates in the next round, preventing degenerate updates with zero participation.

3.4 Client-Server Algorithms and Communication Protocol

In the Client-Update of Algorithm 1, at each round t, client k evaluates the validation accuracy of the global model $\boldsymbol{\theta}^t$ and its local model $\boldsymbol{\theta}_k^{t-1}$ on $\mathcal{D}_k^{\text{val}}$. Using the announced expected aggregation weight $\alpha_k^{t-1} = \mathbb{E}_{i\sim\mathbf{s}^{t-1}}[r_k^{i,t-1}]$, it computes its utility $u_k(\alpha_k^{t-1}, \zeta_k^{t-1})$, reflecting the expected benefit under randomized aggregation. The client accepts the global model if the rationality condition holds; otherwise, it continues from its local model. From the chosen initialization, it performs τ local SGD steps to obtain an updated model, and updates its participation threshold and utility parameters using the realized validation performance and $\hat{u}_k(r)$.

If the rationality condition is satisfied, the client transmits both its updated model and ζ_k; otherwise, it reports only ζ_k. This captures rational, utility-driven participation. In the Server update routine of Algorithm 1, the server aggregates updates $\{\boldsymbol{\theta}_k^t, \gamma_k^t, \zeta_k^t\}_{k\in S_t}$, where S_t denotes participating clients. For non-participating clients, the server reuses their most recent updates. Retaining these stale updates, rather than re-normalizing over active clients, is critical for consistent utility estimation and sustained participation. Since utility depends on a client's historical contribution (tracked via past updates), discarding this information would disrupt future utility estimation and weaken the participation mechanism. Incorporating stale updates preserves continuity and supports long-term

engagement, as also observed in prior work such as FedVarp [16]. The full FedUCA algorithm is provided in Appendix A.2.

Scope: We assume truthful reporting of thresholds, utility parameters, and model updates. Appendix A.13 discusses the impact of threshold inflation on the proposed feasibility framework. Robustness to Byzantine behavior is orthogonal to this work and can be incorporated using existing verification or robust aggregation techniques.

3.5 Analysis

We present our theoretical guarantees in three complementary parts. Propositions 3.1 and 3.2 characterize participation conditions and the expected Jensen's gap as a continuous, tunable function of the Dirichlet concentration δ, establishing that surplus is available at moderate δ and vanishes at the extremes. Proposition 3.3 confirms that the event of a beneficial Jensen surplus occurs with non-zero probability, ensuring the mechanism is non-degenerate. Proposition 3.4 is a *conditional convergence result*: it characterizes the optimization landscape FedUCA operates in, given that participation sustains a minimum cohort size $S \geq 1$ per round. Together, the first two results motivate why FedUCA can sustain participation; the third characterizes convergence quality once participation is sustained.

Algorithm 1: FedUCA: Client-Server Updates

Function
`ClientUpdate(`$\boldsymbol{\theta}^t$, α_k^{t-1}, $\boldsymbol{\theta}_k^{t-1}$, γ_k^{t-1}, $\mathcal{D}_k$, $\mathcal{D}_k^{val}$`):`
- ▷ Evaluate metrics & check rationality limit
- $\boldsymbol{\theta}_k^{t,0} \leftarrow \boldsymbol{\theta}^t$ if $u_k(\alpha_k^{t-1}, \zeta_k^{t-1}) > \gamma_k^{t-1}$ else $\boldsymbol{\theta}_k^{t-1}$;
- ▷ Local SGD execution
- **for** $l \leftarrow 1$ **to** τ **do**
  - $\boldsymbol{\theta}_k^{t,l} \leftarrow \boldsymbol{\theta}_k^{t,l-1} - \eta_c \nabla f_k(\boldsymbol{\theta}_k^{t,l-1}; \xi_k^{t,l-1})$;
- ▷ Update utility parameters & transmit
- Evaluate $\hat{u}_k(r)$ as in (4) to compute ζ_k^t as Eq. 5;
- $\gamma_k^t \leftarrow u_k(\alpha_k^{t-1}, \zeta_k^t)$;
- **return** $(\boldsymbol{\theta}_k^{t,\tau}, \gamma_k^t, \zeta_k^t)$ if $u_k(\alpha_k^{t-1}, \zeta_k^{t-1}) > \gamma_k^{t-1}$ else (γ_k^t, ζ_k^t) ;

Function
`ServerUpdate(`$(\boldsymbol{\theta}_k^t, \gamma_k^t, \zeta_k^t)_{k \in S_t}, (\gamma_k^t, \zeta_k^t)_{k \in S_t^c}$`):`
- ▷ compute pseudo-gradients
- compute $\{\alpha_1^t, \ldots, \alpha_N^t\}$ from $\mathbf{R}^t$ and $\mathbf{s}^t$;
- $\mathbf{d}_k^t \leftarrow \frac{\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t}{\eta_c} \quad (\forall k \in S_t)$;
- ▷ Aggregate active and stale updates
- $\mathbf{g}^t \leftarrow \sum_{k \in S_t} \alpha_k^t \mathbf{d}_k^t + \sum_{k \in S_t^c} \alpha_k^t \mathbf{y}_k^t$;
- $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^{t,0} - \eta\tau \mathbf{g}^t$;
- ▷ Update stale trackers & solve for next round
- $\mathbf{y}_k^{t+1} \leftarrow \mathbf{d}_k^t$ if $k \in S_t$ else $\mathbf{y}_k^t$;
- Sample $\mathbf{s}^{t+1}$ and solve $\mathbf{R}^{t+1}$ as in Eq. (7);
- **return** $\boldsymbol{\theta}^{t+1}$, $\mathbf{s}^{t+1}$, $\mathbf{R}^{t+1}$;

Proposition 3.1. *Suppose $\epsilon = 0$, and all utilities $u_k(.)$ are concave then the solution to optimization 7 satisfies either $\sum_{i=1}^{n_s} s_i u_k(r_k^i) = \gamma_k - t_k$, with $t_k > 0$ or $\sum_{i=1}^{n_s} s_i u_k(r_k^i) \geq \gamma_k$, with $t_k = 0$. In the former case, the constraint is satisfied when $u_k(r_k^i) = \frac{s_i}{\sum_{j=1}^{n_s} s_j^2}(\gamma_k - t_k)$. If we further assume that $u_k(.)$ is strictly concave, the optimal solution $\mathbf{R}^*, \mathbf{t}^*$ is unique, if the former holds $\forall k \in [N]$.*

Proof Sketch. The proof follows by analyzing the KKT conditions. The details are provided in the appendix in proposition A.1. The generality of the result holds independent of ϵ as it can always be absorbed into the γ_k. □

The proposition states that the solution exists in two classes either satisfying $\sum_{i=1}^{n_s} s_i u_k(r_k^i) \geq \gamma_k$ (case 1) or $\sum_{i=1}^{n_s} s_i u_k(r_k^i) = \gamma_k - t_k$ (case 2) with $t_k > 0$. When the server gives the client k the aggregation weight $\alpha_k = \sum_{i=1}^{n} s_i r_k^i$. Owing to the Jensen's inequality, in the case 1 the clients participate as we have $u_k(\sum_{i=1}^{n_s} s_i r_k^i) \geq \sum_{i=1}^{n_s} s_i u_k(r_k^i) \geq \gamma_k$. In the case 2, we have $u_k(\sum_{i=1}^{n_s} s_i r_k^i) \geq \sum_{i=1}^{n_s} s_i u_k(r_k^i) = \gamma_k - t_k$. In this scenario, the client may or may not participate depending on whether $u_k(\sum_{i=1}^{n_s} s_i r_k^i) \geq \gamma_k$ or $u_k(\sum_{i=1}^{n_s} s_i r_k^i) < \gamma_k$. Especially the participation is possible if $u_k(\sum_{i=1}^{n} s_i r_k^i)$ exceeds $\sum_{i=1}^{n} s_i u_k(r_k^i)$ by at least t_k. Towards this end, we analyze the induced Jensen's gap associated with client k,

$$\mathcal{J}_k(\mathbf{s}) = u_k(\sum_{i=1}^{n_s} s_i r_k^i) - \sum_{i=1}^{n_s} s_i u_k(r_k^i), \tag{8}$$

as a function of $\mathbf{s}$. Since $\mathbf{s}$ is randomly chosen from the Dirichlet distribution parameterized by δ. Unlike standard stochastic regularization, the Jensen's gap arises from the concavity of client utilities and directly induces a surplus in expected utility, which is explicitly leveraged to satisfy participation constraints. We note that $\mathcal{J}_k(\mathbf{s})$ is a random variable, and we characterize its expectation under the Dirichlet mixture distribution.

Proposition 3.2. *Let the expected Jensen's gap for client k be defined as $\bar{\mathcal{J}}_k(\delta) \coloneqq \mathbb{E}_{\mathbf{s}\sim Dir(\delta)}[\mathcal{J}_k(\mathbf{s})]$ for a concentration parameter $\delta > 0$. The instantaneous gap is given by $\mathcal{J}_k(\mathbf{s}) \coloneqq u_k\big(\sum_{i=1}^{n_s} s_i r_k^i\big) - \sum_{i=1}^{n_s} s_i u_k(r_k^i)$, where $u_k(\cdot)$ is a strictly concave utility function and the aggregation weights r_k^i are obtained by solving Optimization 7, where the constraint is binding. The expected gap $\bar{\mathcal{J}}_k(\delta)$ vanishes at the extreme concentration regimes, specifically when $\delta \leq \delta_{min}$ or $\delta \geq \delta_{max}$ (where $\delta_{min} \ll 1$ and $\delta_{max} \gg 1$). Furthermore, $\bar{\mathcal{J}}_k(\delta)$ is a continuous function of δ on the open interval $(\delta_{min}, \delta_{max})$.*

Proof Sketch. The proof follows by analyzing the cases $\delta \ll 1$ and $\delta \gg 1$. We then argue the continuity of Jensen's gap as a function of the concentration parameter δ by appealing to the Dominated Convergence Theorem. Since the gap is positive and due to continuity it attains the maximum when $\delta \in (\delta_{min}, \delta_{max})$. The complete proof is provided in the proposition A.2. □

The proposition implies that choosing δ which is neither too small nor too big is better. Empirical measurements of $\bar{\mathcal{J}}_k(\delta)$ on CIFAR-10 confirm the predicted non-monotonic profile, with the gap smaller at $\delta = 0.01$ and $\delta = 100$ and peaking at $\delta \in [0.5, 1.0]$ (Section A.11, Table 7).

We now provide the sufficient conditions under which the Jensen's gap $\mathcal{J}_k(\mathbf{s})$ exceeds a particular threshold. In particular, we provide the lower bound on the probability that $\mathcal{J}_k(\mathbf{s}) \geq t$.

Proposition 3.3. *Suppose $u_k(\cdot)$ is concave, $\mathcal{J}_k(s) \geq 0$, and $t \leq \mathop{\mathbb{E}}_{\mathbf{s}\sim Dir(\delta)}[\mathcal{J}_k(\mathbf{s})]$ and $n_s \leq N$ then $Pr[\mathcal{J}_k(\mathbf{s}) \geq t] \geq \frac{(\mathbb{E}[\mathcal{J}_k(\mathbf{s})]-t)^2}{{M_J}^2}$, for some M_J and $M_J \leq 1$, if the utility is accuracy.*

Proof Sketch. The key is to note that $\mathcal{J}_k(\mathbf{s})$ is a non-negative random variable, the proof follows by using the Paley-Zygmund inequality [31]. The proof details are in the proposition A.3 of the Appendix. The important thing is to note that this probability is non-zero. □

Discussion: The stochastic formulation does not enlarge the simplex of aggregation weights. Rather, under concave utilities, it enables Jensen's gap surplus that can satisfy IR constraints when a single deterministic aggregation vector cannot. Thus, randomization can induce better participation.

Assumptions for Convergence: We adpot the standard non-convex FL assumptions used in partial-participation analyses [16, 30]: each f_k is L-smooth, stochastic gradients are unbiased with variance at most σ^2, client heterogeneity satisfies $\|\nabla f_k(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|^2 \leq \sigma_g^2$, local gradients are bounded as $\|\nabla f_k(\boldsymbol{\theta})\|^2 \leq G$, stale delays satisfy $\nu_k^t \leq \nu_{\max}$. We assume that the participation set S_t is uniformly lower bounded, i.e., $1 \leq S \leq |S_t| \leq N$, where S is the minimum number of participating clients per round. In our fallback implementation, the weakest guarantee corresponds to $S = 1$.

Proposition 3.4 (Convergence decomposition)**.** *Under standard smoothness, bounded variance, bounded gradient, and bounded stale-delay assumptions, FedUCA satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \leq \mathcal{E}_{\text{init}} + \mathcal{E}_{\text{agg}} + \mathcal{E}_{\text{var}} + \mathcal{E}_{\text{drift}} + \mathcal{E}_{\text{sys}}.$$

Here, $f(\boldsymbol{\theta}) = \sum_{k=1}^N \beta_k f_k(\boldsymbol{\theta})$, $\sum_{k=1}^N \beta_k = 1$, T denotes rounds, $\mathcal{E}_{\text{init}} = \mathcal{O}(1/\sqrt{(T)})$ is the optimization term, $\mathcal{E}_{\text{agg}}$ is proportional to $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2$, and the remaining terms capture stochastic variance, local-update drift, stale-update drift, and data heterogeneity. The $\mathcal{E}_{\text{sys}}$ contains a non-vanishing bias term $\mathcal{O}((N-S)\sigma^2)$. The full statement and proof are in the Appendix A.21.

Proof Sketch. We write the exact FedUCA average-update direction as $\bar{\mathbf{v}}^t = \sum_{k=1}^N \Big(I_k^t \alpha_k^t \mathbf{h}_k^t + (1 - I_k^t)\alpha_k^t \mathbf{h}_k^{t-\nu_k^t}\Big)$ where I_k^t denotes participation, α_k^t is the FedUCA aggregation weight, and $\nu_k^t \leq \nu_{\max}$ is the stale delay. Applying the smoothness descent lemma to this trajectory separates the progress term from stochastic variance, local drift, stale-update drift, and aggregation discrepancy.

Conditioning on the history H_t makes the reported utilities, thresholds, participation decisions, and aggregation weights fixed for round t. We then add and subtract the full gradient $\nabla f(\boldsymbol{\theta}^{(t,0)}) = \sum_{k=1}^N \beta_k \nabla f_k(\boldsymbol{\theta}^{(t,0)})$, which yields the aggregation-discrepancy penalty $\mathcal{E}_{agg}$ upper bounded by $\mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2$. If $\boldsymbol{\beta} = (1/N)\mathbf{1}_N$, from Lemma A.10, we have $\mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2 \leq (n_s - 1)/[N(n_s\delta + 1)]$ and eventually, we have $\mathcal{E}_{agg} \leq C(n_s - 1)/[S(n_s\delta + 1)]$, where C is a constant. The active-client drift follows from standard local-SGD bounds, while stale-client drift is controlled by the bounded-delay condition $\nu_k^t \leq \nu_{\max}$ and the minimum active cohort size S. Telescoping the resulting descent

Table 2: Server model performance on CIFAR-10, F-MNIST, and CIFAR-100 under different data splits. The data partition Dirichlet $(0.5, 1.0)$ implies that training data is distributed according to the Dirichlet distribution based on the concentration parameter 0.5, and the validation data uses 1.0.

	CIFAR-10		F-MNIST		CIFAR-100	
	Dirichlet 0.5,1.0	**Dirichlet 1.0,1.0**	**Dirichlet 0.5,1.0**	**Dirichlet 1.0,1.0**	**Dirichlet 0.5,1.0**	**Dirichlet 1.0,1.0**
FedAvg	$16.63_{\pm 1.16}$	$18.31_{\pm 1.66}$	$53.26_{\pm 2.78}$	$54.64_{\pm 2.05}$	$04.05_{\pm 0.06}$	$03.63_{\pm 0.50}$
IncFL	$32.35_{\pm 0.96}$	$34.60_{\pm 0.16}$	$37.07_{\pm 0.20}$	$36.69_{\pm 0.04}$	$04.18_{\pm 0.41}$	$12.89_{\pm 0.94}$
FedProx	$17.44_{\pm 0.86}$	$17.63_{\pm 2.14}$	$54.68_{\pm 2.01}$	$53.79_{\pm 0.55}$	$06.35_{\pm 2.91}$	$04.06_{\pm 0.22}$
AAggFF-S	$16.13_{\pm 0.50}$	$17.90_{\pm 0.33}$	$52.84_{\pm 3.18}$	$68.92_{\pm 6.22}$	$5.60_{\pm 0.33}$	$04.53_{\pm 0.86}$
FedAvg-WS	$47.70_{\pm 0.47}$	$51.63_{\pm 0.71}$	$77.60_{\pm 0.57}$	$75.75_{\pm 0.23}$	$33.50_{\pm 0.26}$	$48.93_{\pm 0.05}$
FedUCA (Ours)	$\mathbf{66.89_{\pm 01.56}}$	$\mathbf{71.86_{\pm 1.17}}$	$\mathbf{83.18_{\pm 0.40}}$	$\mathbf{83.48_{\pm 0.85}}$	$\mathbf{42.50_{\pm 0.29}}$	$\mathbf{50.15_{\pm 0.32}}$

inequality over T rounds gives $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}^{(t,0)})\|^2 \leq \mathcal{E}_{\text{init}} + \mathcal{E}_{\text{agg}} + \mathcal{E}_{\text{var}} + \mathcal{E}_{\text{drift}} + \mathcal{E}_{\text{sys}}$. All terms except the structural aggregation bias $\mathcal{E}_{\text{agg}}$ and the system error floor $\mathcal{E}_{\text{sys}}$ decay at a rate of $\mathcal{O}(1/T)$ or $\mathcal{O}(1/\sqrt{T})$. Consequently, FedUCA converges to an error neighborhood defined by these non-vanishing terms. □

FedUCA should not be interpreted as an unbiased variant of FedAvg. The server deliberately modifies aggregation weights to satisfy individual rationality constraints, rather than solely approximating the full-participation descent direction. Consequently, the aggregation-discrepancy term $\mathcal{E}_{agg}$ is not merely a proof artifact; it captures the mechanism-induced bias incurred when the server trades aggregation fidelity for sustained rational participation. In the limiting regime where participation is full and the FedUCA weights match the target objective weights, this discrepancy vanishes and the standard FL behavior is recovered.

The bound $\mathbb{E}\|(1/N)\mathbf{1}_N - \boldsymbol{\alpha}^t\|^2 \leq (n_s - 1)/[N(n_s\delta + 1)]$ (Lemma A.10) reveals a quantitative tradeoff governing FedUCA's convergence. The Dirichlet concentration δ simultaneously controls two opposing forces: increasing δ tightens the aggregation discrepancy (the bound decays as $1/(n_s\delta + 1)$) but reduces the Jensen surplus that sustains participation (Proposition 3.2). The optimal δ balances these forces, which is consistent with the empirical participation peak at $\delta \in [0.5, 1.0]$ (Figure 2) and correlates with Jensen's-gap peak. Similarly, the strategy count n_s enters the numerator linearly but also appears in the denominator through $n_s\delta$; for large n_s the bound saturates at $\approx 1/(N\delta)$, confirming that additional strategies yield diminishing returns.

4 Experiments

Datasets and Models: We benchmark on three different datasets: CIFAR-10, CIFAR-100 [22], F-MNIST [37]. For the correct evaluation of a method's persuasive power, the design of the data split amongst clients is of utmost importance. We provide an extensive discussion of the data splits in A.6 of the Appendix and request the readers to refer to the same. The data partition Dirichlet $(0.5, 1.0)$ implies that training data is distributed according to the Dirichlet distribution based on the concentration parameter 0.5, and the validation data uses 1.0 concentration parameter similar to [1]. We use the CVXPY [2, 9] to solve the optimization 7 for FedUCA. For CIFAR-10, we adopt a CNN architecture akin to that of McMahan et al. [28] and [1], comprising two convolutional layers with 64 5 × 5 filters, followed by two fully connected layers with 384 and 192 units, respectively, and a softmax output layer. We use a fully-connected neural network architecture for F-MNIST with 2 layer MLP. The number of neurons in the layers are 200 and 100. For evaluation on CIFAR-100, we use a ResNet-18 model [15] pre-trained on ImageNet-1K [8] dataset where we use group normalization instead of batch normalization as in [1]. While FedUCA theory is generic, for experiments we prefer our utility to be the concave quadratic, $u_k(r, \zeta_k) = -a_k r^2 + b_k r + c_k$, where a_k is positive, and $\zeta_k := \{a_k, b_k, c_k\}$. Since these are parameterized by 3 real numbers, this is communication efficient. The hyperparameter details are in the Appendix A.9.

Baselines and evaluation protocol: Since FedUCA operates without monetary incentives, market-based incentive baselines are not directly applicable. We equip four standard FL methods,

FedAvg, FedProx, IncFL, and AAggFF-S [13], with a rationality check (Eq. 1): a client accepts the global model only if it improves upon the local model on the client's held-out set. This preserves each method's original update rule and isolates the effect of aggregation design under rational behavior. To ensure a rigorous and fair comparison, we equip all baseline methods with the exact same stale update retention mechanism utilized by FedUCA. To further test whether baseline degradation is a transient initialization effect, we additionally evaluate FedAvg under a warm-start protocol (FedAvg-WS) with $K = 5$ rounds of forced full participation before the rationality check activates. All methods use identical architectures, data splits, training budgets, and participation dynamics; implementation details are in Appendix A.8.

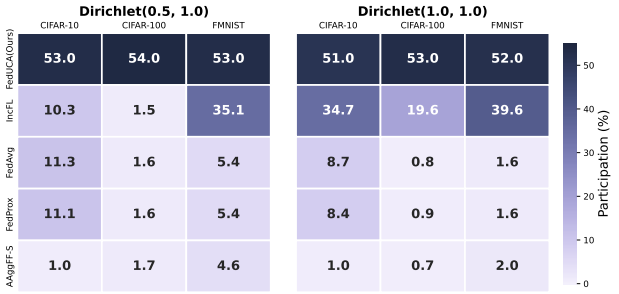


Figure 1: Comparison of participation rates across different FL methods on CIFAR-10, CIFAR-100, and FMNIST under varying degrees of data heterogeneity Dirichlet$(0.5, 1.0)$ and Dirichlet$(1.0, 1.0)$.

Results: As shown in Table 2 and Figure 1, FedUCA achieves substantially higher accuracy and participation across the datasets and splits. Under rational participation, baselines suffer from a self-reinforcing collapse: the initial global model (an average of heterogeneous local models) often degrades individual client performance, clients reject it, the server receives fewer updates, and the next global model deteriorates further. This cycle drives participation below 12% for FedAvg, FedProx, and AAggFF-S on CIFAR-10 and CIFAR-100 (Figure 1). IncFL partially mitigates this but still plateaus at $\sim 35\%$ participation on CIFAR-10, Dir(1.0,1.0). FedUCA breaks this cycle by optimizing aggregation weights to satisfy individual rationality constraints directly, sustaining 51-54$\%$ participation and achieving higher accuracy than the baselines. We provide further discussion of the rationality check and alternative thresholding strategies in Appendix A.15. We additionally evaluate on 20 Newsgroups dataset [29] in Appendix A.10. In A.3 of the Appendix, we provide a detailed discussion on how we introduce the rationality check by clients when evaluated for the benchmarking methods. In the Sec A.6, we provide details on the data preparation.

Warm-start evaluation: To test whether the baseline gap is a transient initialization effect, we evaluate FedAvg under a warm-start protocol (FedAvg-WS) with $K = 5$ rounds of forced full participation. Under the heterogeneous Dir$(0.5, 1.0)$ splits, FedUCA leads FedAvg-WS by $+19.2$ points on CIFAR-10, $+5.6$ on F-MNIST, and $+9.0$ on CIFAR-100; under Dir$(1.0, 1.0)$, FedUCA leads FedAvg-WS by $+20.2$ points on CIFAR-10, $+7.7$ points on F-MNIST, and $+1.2$ points on CIFAR-100. FedUCA's advantage is consistent across varying heterogeneity regimes. Warm-start is not a benign baseline modification: in a rational cross-silo deployment, forced participation presupposes monetary compensation or regulatory authority [17], system components external to the aggregation method. FedUCA replaces these external components endogenously through participation-management, achieving superior performance without requiring infrastructure outside the aggregation mechanism.

4.1 Impact on varying the Dirichlet Concentration δ

Figure 2 illustrates the effect of the Dirichlet concentration parameter δ. Very small δ concentrates strategies near simplex vertices, while very large δ concentrates them near the uniform vector. In both extremes, the induced Jensen surplus is limited. Intermediate values provide useful strategy diversity without collapsing to either extreme, yielding higher participation. This empirical trend provides a strong validation of our theoretical claim: strategic randomization is not a mere heuristic, but a mathematically grounded method to bridge the Jensen's gap and stabilize participation in rational FL. The participation

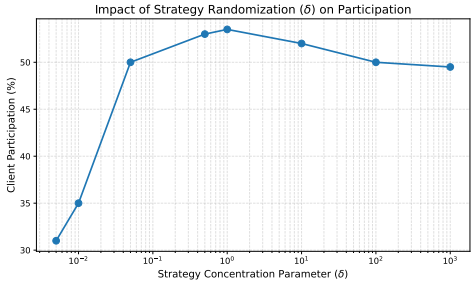


Figure 2: Impact of Dirichlet concentration δ on participation for CIFAR-10.

rate exhibits the non-monotonic dependence on δ predicted by Proposition 3.2, rising from $\sim 35\%$ at $\delta = 0.01$ to a peak of $\sim 54\%$ at $\delta = 1$, before declining to $\sim 49\%$ at $\delta = 1000$. Independent measurements of the Jensen gap on the same setup Appendix A.11) show a similar non-monotonic trend, with larger gaps occurring in the same intermediate concentration regime. This alignment supports the proposed mechanism: intermediate Dirichlet concentrations induce useful strategy diversity, which increases the Jensen surplus available to Optimization 7 and thereby improves participation feasibility.

4.2 Comparison with single strategy

In the table. 3 we compare the performance of the server model with $n_s = 1$ vs $n_s = 5$. By increasing the number of strategies ($n_s = 5$), the solver optimizes the pessimistic surrogate, generating utility surplus via the Jensen's gap. This empirically confirms our theory that multiple strategies exploit the Jensen's gap and increase the participation.

Table 3: Server Accuracy between single and multiple strategies

Strategies	**CIFAR-10** (ACC %/ Participation %)	**CIFAR-100** (ACC % / Participation %)
$n_s = 1$	63.80 (29)	36.80 (37)
$n_s = 5$	66.89 (53)	42.50 (54)

4.3 Computation

We evaluate FedUCA's computational footprint on NVIDIA-RTX A5000 GPUs and find the overhead to be minimal. On the client-side, estimating the utility curve only requires forward passes on the validation set and a lightweight least-squares fit. On the server-side, the server Optimization is a convex program independent of the model dimension d. The detailed breakdown is provided in the Appendix A.14.

5 Limitations and Future work

FedUCA addresses cross-silo federated learning under rational, performance-based client participation, with concave client utilities standard in mechanism design [32] and verified empirically in this setting (Appendix A.4). The framework assumes truthful reporting of utility parameters, which is supported in cross-silo deployments by repeated interaction and institutional accountability; extensions incorporating formal incentive-compatibility guarantees are a natural direction. On the theoretical side, Proposition 3.4 together with Lemma A.10 couples the aggregation-discrepancy term to the Dirichlet concentration δ, identifying a quantitative tradeoff with Jensen surplus. Bounding S in terms of system parameters is a refinement we leave for future work.

6 Conclusion

We introduced FedUCA, which reformulates federated aggregation as a constrained feasibility problem enforcing individual rationality through utility-threshold constraints. A stochastic variant samples weights from a Dirichlet distribution, exploiting utility concavity and Jensen's gap to offset utility deficits, with the concentration parameter δ governing the expected gap. Experiments on CIFAR-10, CIFAR-100, F-MNIST, and 20 Newsgroups demonstrate gains in federation stability and global accuracy across modalities. By reframing aggregation as a feasibility problem under rational participation, FedUCA provides a practical, non-monetary alternative to equilibrium-based and incentive-driven approaches, enabling stable cross-silo collaboration without external rewards or strategic modeling.

Acknowledgment

This work is partially supported by the P3DX project, seed-funded by the Ministry of Electronics and Information Technology (MeitY). The authors would also like to acknowledge compute support received from Kotak-IISc AI-ML Center (KIAC), IISc.

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

[2] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[3] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Studies in Economic Theory. Springer, 1999. ISBN 9783540658542. URL https://books.google.co.in/books?id=6jjY2Vi3aDEC.

[4] Hongliu Cao. Holistic analysis on the sustainability of federated learning across ai product lifecycle. *arXiv preprint arXiv:2312.14628*, 2023.

[5] Yae Jee Cho, Divyansh Jhunjhunwala, Tian Li, Virginia Smith, and Gauri Joshi. To federate or not to federate: Incentivizing client participation in federated learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022. URL https://openreview.net/forum?id=pG08eM0CQba.

[6] Francesco Cremonesi, Marc Vesin, Sergen Cansiz, Yannick Bouillard, Irene Balelli, Lucia Innocenti, Riccardo Taiello, Santiago Silva, Samy-Safwan Ayed, Melek Önen, et al. Fedbiomed: Open, transparent and trusted federated learning for real-world healthcare applications. In *Federated Learning Systems: Towards Privacy-Preserving Distributed AI*, pages 19–41. Springer, 2025.

[7] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10): 1735–1743, 2021.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[9] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[10] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/24389bfe4fe2eba8bf9aa9203a44cdad-Paper.pdf.

[11] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[12] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019.

[13] Seok-Ju Hahn, Gi-Soo Kim, and Junghye Lee. Pursuing overall welfare in federated learning through sequential decision making, 2024. URL https://arxiv.org/abs/2405.20821.

[14] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction, 2019. URL https://arxiv.org/abs/1811.03604.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916. PMLR, 2022.

[17] Jiawen Kang, Zehui Xiong, Dusit Niyato, Shengli Xie, and Junshan Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. volume 6, pages 10700–10714, 2019. doi: 10.1109/JIOT.2019.2940820.

[18] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[19] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael Jordan. Mechanisms that incentivize data sharing in federated learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022. URL `https://openreview.net/forum?id=Bx4Sz-N5K3J`.

[20] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[21] Shuyu Kong, You Li, and Hai Zhou. Incentivizing federated learning. *arXiv preprint arXiv:2205.10951*, 2022.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `https://api.semanticscholar.org/CorpusID:18268744`.

[23] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

[24] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[25] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ByexElSYDr`.

[26] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

[27] Bing Luo, Yutong Feng, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Incentive mechanism design for unbiased federated learning with randomized client participation. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, pages 545–555. IEEE, 2023.

[28] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[29] T. Mitchell. Twenty Newsgroups, 1997. URL `https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups`.

[30] John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. Federated learning with buffered asynchronous aggregation. In *International conference on artificial intelligence and statistics*, pages 3581–3607. PMLR, 2022.

[31] Valentin V. Petrov. On lower bounds for tail probabilities. *Journal of Statistical Planning and Inference*, 137(8):2703–2705, 2007. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2006.02.015. URL `https://www.sciencedirect.com/science/article/pii/S0378375807000213`. 5th St. Petersburg Workshop on Simulation.

[32] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913738`.

[33] Yunus Sarikaya and Ozgur Ercetin. Motivating workers in federated learning: A stackelberg game perspective, 2019. URL https://arxiv.org/abs/1908.03092.

[34] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

[35] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In *Federated Learning: Privacy and Incentive*, pages 153–167. Springer, 2020.

[36] Hongyi Wu, Xiaoying Tang, Ying-Jun Angela Zhang, and Lin Gao. Incentive mechanism for federated learning with random client selection. *IEEE Transactions on Network Science and Engineering*, 11(2):1922–1933, 2024. doi: 10.1109/TNSE.2023.3334476.

[37] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

[38] Yun Xin, Jianfeng Lu, Shuqin Cao, Gang Li, Haozhao Wang, and Guanghui Wen. Daringfed: a dynamic bayesian persuasion pricing for online federated learning under two-sided incomplete information. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6687–6695, 2025.

[39] Yunchao Yang, Yipeng Zhou, Miao Hu, Di Wu, and Quan Z. Sheng. Bara: Efficient incentive mechanism with online reward budget allocation in cross-silo federated learning. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4478–4485. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/498. URL https://doi.org/10.24963/ijcai.2023/498. Main Track.

[40] Rongfei Zeng, Chao Zeng, Xingwei Wang, Bo Li, and Xiaowen Chu. A comprehensive survey of incentive mechanism for federated learning, 2021. URL https://arxiv.org/abs/2106.15406.

[41] Ning Zhang, Xiaoqing Xu, Liuyihui Qian, Xiaojun Liu, Juan Wu, and Hong Tang. Auction-based incentive mechanism in federated learning considering communication path finding. *IEEE Access*, 12:139336–139345, 2024. doi: 10.1109/ACCESS.2024.3425948.

A Technical appendices and supplementary material

A.1 Notation 14
A.2 FedUCA Illustration and Algorithm 15
A.3 Rationality Check for benchmark methods 16
A.4 Empirical Verification of Concavity of the Utility 16
A.5 Discussion on Utility Design 17
A.6 Data Preparation 18
A.7 F-MNIST and CIFAR-100 19
A.8 Experimental and Implementation Details 19
A.9 HyperParameter Settings 20
A.10 Experiments on Natural Language Data 20
A.11 Empirical Verification of Jensen's Gap 20
A.12 Empirical Verification of the Aggregation Discrepancy Bound 21
A.13 Impact due to Inflated Thresholds 22
A.14 Computational Complexity 22

A.15 On the rationality Check and alternate thresholding strategies 23
A.16 Discussion on Comparison with Monetary Incentives 24
A.17 Additional Experiments 24
A.17.1 Comparison with Personalized baselines 24
A.17.2 Experiments with cubic utility fit and loss as utility 25
A.18 Proofs of Propositions 25
A.19 Assumptions for Convergence analysis 29
A.20 Update rules for the FedUCA 29
A.21 Convergence Analysis 30

A.1 Notation

- N: Number of clients
- $k \in [N]$: Client index
- t: Communication round index
- $\boldsymbol{\theta}^t$: Global server model at round t, we also use $\boldsymbol{\theta}^{t,0}$ for the same.
- $\boldsymbol{\theta}_k^t$: Local model/update of client k at round t
- $f_k(\boldsymbol{\theta})$: Local objective of client k
- $f(\boldsymbol{\theta}) = \sum_{k=1}^N \beta_k f_k(\boldsymbol{\theta})$: Global objective
- $\boldsymbol{\beta}$: Target objective weights over clients
- α_k^t: Effective aggregation weight assigned to client k at round t
- $\mathbf{s} = (s_1, \ldots, s_{n_s})$: Probability vector over aggregation strategies
- n_s: Number of aggregation strategies
- $\mathbf{R} \in \mathbb{R}^{n_s \times N}$: Strategy matrix; row i is one aggregation vector over clients
- r_k^i: Aggregation weight assigned to client k under strategy i
- $\hat{u}_k(r)$: Empirical utility of client k at aggregation weight r
- $u_k(r; \zeta_k)$: Fitted concave surrogate utility of client k
- ζ_k: Parameters of client k's fitted utility function
- γ_k: Minimum participation threshold of client k
- t_k: Slack variable for client k's rationality constraint
- S_t: Set of clients participating at round t
- S: Minimum number of participating clients per round
- I_k^t: Indicator that client k participates at round t
- ν_k^t: Staleness delay of client k at round t
- $\nu_{\max}$: Maximum allowed staleness delay
- $\mathbf{y}_k^t$: Stale update tracker for client k
- $\mathcal{J}_k(\mathbf{s})$: Jensen surplus for client k under mixture $\mathbf{s}$
- δ: Dirichlet concentration parameter for sampling $\mathbf{s}$
- $\mathcal{E}_{\text{init}}$: Transient optimization error term
- $\mathcal{E}_{\text{agg}}$: Aggregation-discrepancy / mechanism-induced bias term
- $\mathcal{E}_{\text{var}}$: Stochastic variance term
- $\mathcal{E}_{\text{drift}}$: Local and stale-update drift term
- $\mathcal{E}_{\text{sys}}$: System floor

Algorithm 2: FedUCA

Input: $\boldsymbol{\theta}$, $\mathbf{s}$; ▷ initial server model, initial aggregation strategy

Output: $\boldsymbol{\theta}^T$; ▷ final aggregated server model in T communication rounds

Hyperparameters : T, τ, n_s,η_c,η_s,η ; ▷ communication rounds, epochs, aggregation strategies, client learning rate, server learning rate and overall learning rate

Initialize:

$\boldsymbol{\theta}^0 \leftarrow \boldsymbol{\theta}$; $\mathbf{s}^{-1} \leftarrow \mathbf{s}$; $\mathbf{R}^{-1} \leftarrow \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathbf{T}}$; $\gamma_k^{-1} \leftarrow 0$, $\boldsymbol{\theta}_k^{-1} \leftarrow \boldsymbol{\theta}$;

for $t \leftarrow 0$ **to** $T-1$ **do**

  Server broadcasts $\boldsymbol{\theta}^t$ to all N clients with indices $[N]$;

  for $k \leftarrow 1$ **to** N **do**

    $\mathbf{r}_k^{t-1} \leftarrow \mathbf{R}_{:,k}^{t-1}$;

    compute $\alpha_k^{t-1} = \sum_{j=1}^{n_s} s_j^{t-1} r_k^{j,t-1}$;

    Server sends $\boldsymbol{\theta}^t$, $\mathbf{s}^{t-1}$, α_k^{t-1} to client k;

    $\boldsymbol{\theta}_k^t, \gamma_k^t, \zeta_k^t = \textbf{ClientUpdate}(\boldsymbol{\theta}^t, \alpha_k^{t-1}, \boldsymbol{\theta}_k^{t-1}, \gamma_k^{t-1}, \mathcal{D}_k, \mathcal{D}_k^{\text{val}})$;

    ; ▷ if participates

    or

    $\gamma_k^t, \zeta_k^t = \textbf{ClientUpdate}(\boldsymbol{\theta}^t, \alpha_k^{t-1}, \boldsymbol{\theta}_k^{t-1}, \gamma_k^{t-1}, \mathcal{D}_k, \mathcal{D}_k^{\text{val}})$;

    ; ▷ if doesn't participate

  $\boldsymbol{\theta}^{t+1}, \mathbf{s}^{t+1}, \mathbf{R}^{t+1} \leftarrow \textbf{SU}((\boldsymbol{\theta}_k^t, \gamma_k^t, \zeta_k^t)_{k\in S_t}, (\gamma_k^t, \zeta_k^t)_{k\in S_t^c})$;

A.2 FedUCA Illustration and Algorithm

Explanation of FedUCA. Algorithm 2 describes the full **FedUCA** mechanism, an optimization-based aggregation strategy that explicitly models rational client behavior and aims to satisfy client utility constraints through server-side aggregation design.

The server maintains three evolving objects: the global model $\boldsymbol{\theta}^t$, a probability vector $\mathbf{s}^t$ governing the server's signaling strategy, and a matrix $\mathbf{R}^t$ whose columns represent client-specific aggregation weight strategies. These quantities are initialized prior to training and updated iteratively over T communication rounds.

At each round t, the server broadcasts its intent to all N clients. For each client k, the server extracts the corresponding aggregation vector $\mathbf{r}_k^{t-1}$ from the current matrix $\mathbf{R}^{t-1}$, computes the aggregation weight α_k^{t-1} and transmits to the client along with the current global model $\boldsymbol{\theta}^{t-1}$. This transmission constitutes the server's signaling action.

Upon receiving the signal, each client executes the **ClientUpdate** procedure. In this step, client k evaluates its utility under the received aggregation weight using its validation data. Based on this evaluation, the client updates its local model and determines a utility threshold γ_k^t reflecting its willingness to participate. The client then returns its updated local model and strategic information ζ_k^t to the server.

After collecting responses from all clients, the server invokes the Server Update **SU** routine. This step aggregates the local models while simultaneously sampling the distribution $\mathbf{s}^t$ and optimizing the aggregation weight matrix $\mathbf{R}^t$. The update is computed by solving a constrained optimization problem that balances model improvement against the probability of satisfying client participation constraints. As a result, the server refines both the global model and its future signaling strategy.

This iterative process continues for T rounds. By jointly updating the global model and the aggregation signaling mechanism, FedUCA enables the server to strategically influence client participation while accounting for rational, utility-driven behavior. The final output $\boldsymbol{\theta}^T$ represents the learned global model under this federated learning framework.

A note on round indexing. In our implementation, FedUCA's first communication round ($t = 0$) is a seeding round: all clients participate without an individual-rationality check, and the server simply aggregates the shared initialization, so no utility curves are fit and no learning occurs in this round. In the pseudocode this is abstracted by initializing the client models $\boldsymbol{\theta}_k^{-1}$ and the server model $\boldsymbol{\theta}^0$ to the

same state, so the $t = 0$ iteration leaves the model unchanged and the rational machinery engages from $t = 1$ onward. The rational FedAvg baseline, by contrast, trains all clients in round 0 itself, so its first round is a genuine training round. FedUCA therefore operates with one fewer effective training round than this baseline.

A.3 Rationality Check for benchmark methods

Algorithms 3 and 4 detail the integration of rationality check into FedAvg. Extending the rationality check to other algorithms follows a similar approach and is relatively straightforward, as it closely resembles the implementation used for FedAvg. The employed rationality check operates as follows: if the accuracy of the received global model is higher than that of the client's local model, the client accepts the global model and proceeds with training initialized from it. If the rationality check is not satisfied the client proceeds with its own training. It is to be noted that one can use the negative of the loss as well instead of accuracy.

Algorithm 3: FedAvg: *Client Update*

Function `ClientUpdate(`$\boldsymbol{\theta}^t$, $\boldsymbol{\theta}_m^{t-1}$, $\mathcal{D}_m$, $\mathcal{D}_m^{val}$, *client* m`):`

- ▷ `Rationality Check`
- **if** $(\mathbf{acc}(\boldsymbol{\theta}^t, \mathcal{D}_m^{val}) > \mathbf{acc}(\boldsymbol{\theta}_m^{t-1}, \mathcal{D}_m^{val}))$ **then**
  - $\boldsymbol{\theta}_m^{t,0} \leftarrow \boldsymbol{\theta}^t$
  - **for** $l \leftarrow 1$ **to** τ **do**
    - $\boldsymbol{\theta}_m^{t,l} \leftarrow \boldsymbol{\theta}_m^{t,l-1} - \eta_c \nabla f_k\big(\boldsymbol{\theta}_m^{t,l-1}; \xi_m^{t,l-1}\big)$ ▷ `mini-batch SGD step`
  - **return** $\boldsymbol{\theta}_m^{t,\tau}$;
- **else**
  - $\boldsymbol{\theta}_m^{t,0} \leftarrow \boldsymbol{\theta}_m^{t-1}$
  - **for** $l \leftarrow 1$ **to** τ **do**
    - $\boldsymbol{\theta}_m^{t,l} \leftarrow \boldsymbol{\theta}_m^{t,l-1} - \eta_c \nabla f_k\big(\boldsymbol{\theta}_m^{t,l-1}; \xi_m^{t,l-1}\big)$ ▷ `mini-batch SGD step`
  - **return**;

Algorithm 4: FedAvg with Rationality

Input: $\boldsymbol{\theta}^0, T, \nabla f_k(\boldsymbol{\theta}_k^0) = 0$
Output: $\boldsymbol{\theta}^T$
For $t = 0$ (round zero), initialize all client models to θ and all the clients train to form θ_k^0, and participate in Federation and server aggregates $\theta^1 = \frac{1}{N}\sum_{k=1}^{N} \theta_k^0$.
for $t \leftarrow 1$ **to** $T - 1$ **do**
- Server broadcasts to all N clients with indices $[N]$;
- **for** $m \leftarrow 1$ **to** N **do**
  - Server sends $\boldsymbol{\theta}^t$ to client m;
  - $\boldsymbol{\theta}_m^t \leftarrow$ **ClientUpdate**($\boldsymbol{\theta}^t$, $\boldsymbol{\theta}_m^{t-1}$, $\mathcal{D}_m$, $\mathcal{D}_m^{\text{val}}$, client m);
- Update the participating set S_t.
- $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}_k^{t-1}$, $\nabla f_k(\boldsymbol{\theta}_k^t) \leftarrow \nabla f_k(\boldsymbol{\theta}_k^{t-1})$, if $k \notin S_t, \forall k \in [N]$;
- $\boldsymbol{\theta}^{t+1} \leftarrow \frac{1}{N}\sum_k \boldsymbol{\theta}_k^t$

A.4 Empirical Verification of Concavity of the Utility

Table 4 reports the goodness-of-fit of the concave quadratic surrogate $u_k(r; \zeta_k) = -a_k r^2 + b_k r + c_k$ against the empirical utility profile $\hat{u}_k(r)$ across clients and communication rounds. Average R^2 values of 0.83 on CIFAR-10 and 0.82 on CIFAR-100 confirm that the concave quadratic family provides a strong fit to empirical utility profiles in the federated deep learning setting, supporting the modeling assumption introduced in Section 3.2.

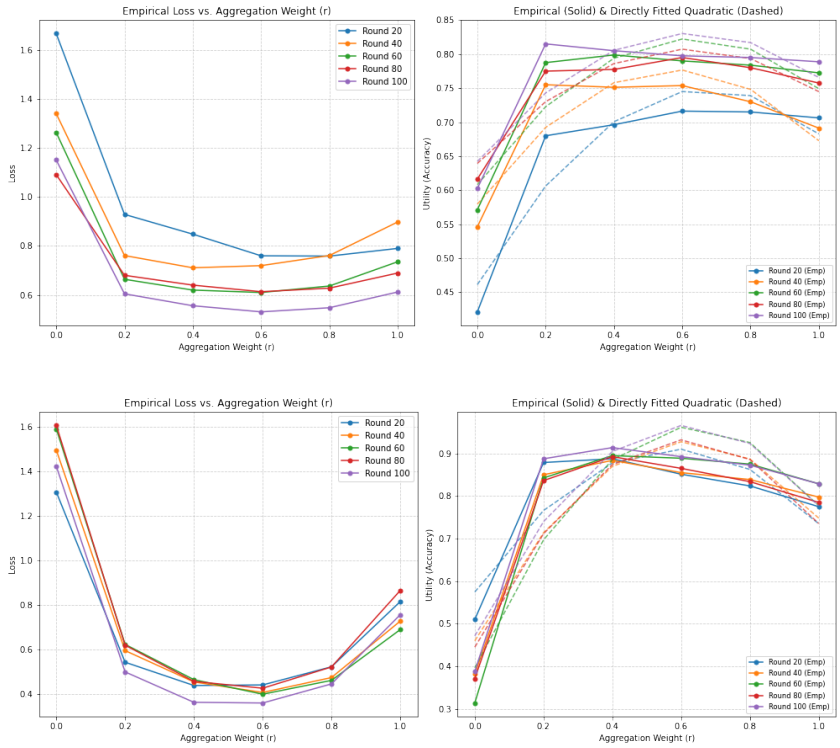


Figure 3: Representative empirical loss and utility profiles across two different clients, rounds, on the CIFAR-10 dataset. Across settings, loss curves exhibit convex-like behavior along the interpolation path, while utility curves remain smooth and unimodal. The dashed curves denote concave quadratic fits, supporting the surrogate utility model used by FedUCA.

Table 4: Goodness-of-fit of concave quadratic surrogate for client utility across datasets and rounds.

Dataset	Avg. R^2 Score
CIFAR-10	0.83
CIFAR-100	0.82

Figures 3 and 4 display representative empirical loss and utility profiles for two clients across multiple communication rounds. Across settings, loss curves exhibit convex-like behavior along the interpolation path while utility curves remain smooth, unimodal, and well-approximated by the concave quadratic fits (dashed). This confirms that the concavity assumption, standard in mechanism design, holds reasonably, as verified empirically in this setting.

A.5 Discussion on Utility Design

In our utility design, client k removes its own contribution from the global model using the weight sent by the server, giving the held-out model $\boldsymbol{\theta}^t_{-k} = \frac{\boldsymbol{\theta}^t - \alpha_k^{t-1}\boldsymbol{\theta}_k^{t-1}}{1-\alpha_k^{t-1}}$, and then, after local training, designs its utility $u_k(r)$ over the interpolated model $r\,\boldsymbol{\theta}^t_k + (1-r)\,\boldsymbol{\theta}^t_{-k}$, where $\boldsymbol{\theta}^t_k$ is its locally trained model. The endpoint $r = 1$ is the freshly trained local model, which also sets the rationality threshold, and $r = 0$ is the held-out rest. The curve does not pass through the realized global $\boldsymbol{\theta}^t$: by construction $\boldsymbol{\theta}^t = \alpha_k^{t-1}\boldsymbol{\theta}_k^{t-1} + (1-\alpha_k^{t-1})\boldsymbol{\theta}^t_{-k}$ lies on the line through the previous-round model $\boldsymbol{\theta}_k^{t-1}$, whereas the curve is built on the locally trained model $\boldsymbol{\theta}^t_k$. The check therefore evaluates a forward-looking model at the server's weighting rather than the model actually aggregated.

The present design carries a one-round lag. The server aggregates using clients' current-round models, but each client builds $u_k(r)$ from the held-out global $\boldsymbol{\theta}^t_{-k}$, which is anchored to $\boldsymbol{\theta}^t$ and the

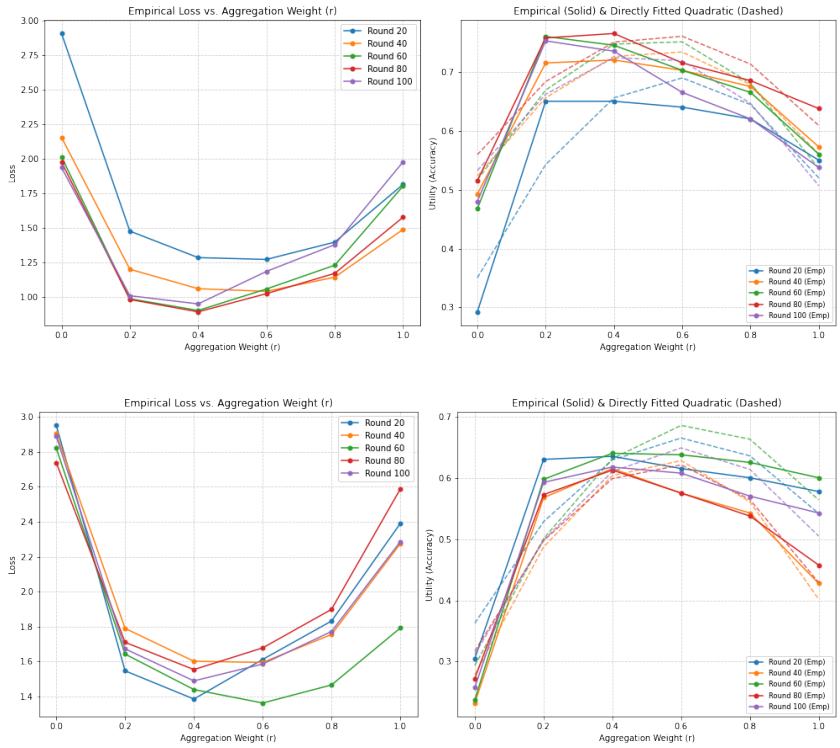


Figure 4: Representative empirical loss and utility profiles across two different clients, rounds, on the CIFAR-100 dataset. Across settings, loss curves exhibit convex-like behavior along the interpolation path, while utility curves remain smooth and unimodal. The dashed curves denote concave quadratic fits, supporting the surrogate utility model used by FedUCA.

round $t-1$ client model $\boldsymbol{\theta}_k^{t-1}$, so the fitted curve is mildly optimistic relative to the realized global. On CIFAR-10 this appears as a small reconcilable gap: the fitted-curve mechanism reaches 53% participation and 66% accuracy, a strict check on the realized global accuracy gives 40% participation and 58% accuracy, and a 3% slack on the realized check recovers the fitted-curve numbers, with the gap shrinking as training proceeds. The accurate alternative is a two-pass protocol per round, in which clients refit their utilities against the current global without further local training and the server re-aggregates, which removes the lag at the cost of roughly doubling communication. Closing this one-round optimism while staying single-pass is the open problem for a future utility design.

A.6 Data Preparation

For the correct evaluation of the persuasion capabilities of any FL algorithm, we propose that, for a cross-silo setup, the data distribution amongst clients should allow for some individual training capabilities of clients while limiting their scope. This would allow for clients to be "self-interested", which could result in a poor server model if clients do not choose to participate. Such a distribution also enables the server to meaningfully influence participation through aggregation design. In general, the data distribution should satisfy:

- Clients have the option to train a local model to a reasonable accuracy if the server model does not benefit the clients.
- If the clients do choose to participate in the federation there are some benefits to gain from the federated training.
- It models real-world settings where the distribution of train and validation data split may not be identical.

- Clients should not participate by default regardless of the aggregation method, as can occur in nearly i.i.d. settings; otherwise, it becomes difficult to compare how effectively different FL algorithms sustain participation.

The notation $(0.5, 1.0)$ indicates that a Dirichlet concentration parameter of 0.5 is used to generate the training data split, while a concentration of 1.0 is used for the validation data split. We first divide the dataset into sets of classes assigned to different clients. Within each client, the training data is sampled using the 0.5 concentration, and the validation data using the 1.0 concentration. The resulting training and validation split for the $(0.5, 1.0)$ setting is illustrated in Figure 5. A similar training and validation split for the $(1.0, 1.0)$ setting is illustrated in Figure 6.

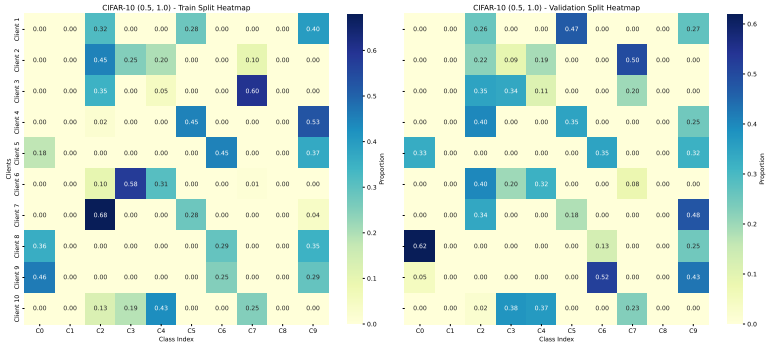


Figure 5: Data distribution heatmap for CIFAR-10 (0.5, 1.0) showing Training and Validation splits across 10 clients.

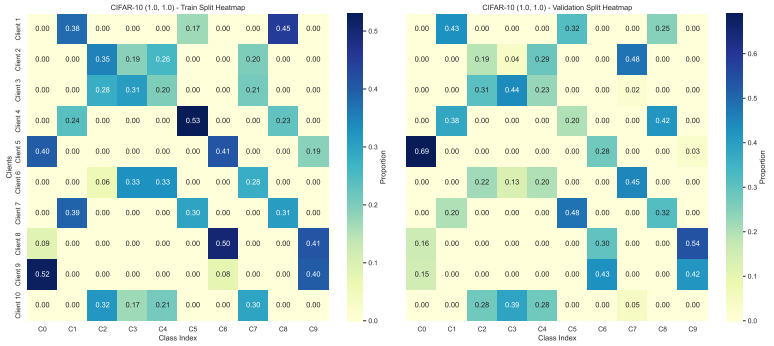


Figure 6: Data distribution heatmap for CIFAR-10 (1.0, 1.0) showing Training and Validation splits across 10 clients.

A.7 F-MNIST and CIFAR-100

For F-MNIST, we employ the same data distribution strategy as for CIFAR-10. For CIFAR-100, since we have 100 classes, we distribute 20 classes to each client in a similar fashion as CIFAR-10.

A.8 Experimental and Implementation Details

We adopt hyperparameter settings similar to those used in FedDyn. We develop our experiments on the code released by FedDyn [1]. Specifically, we set the number of local training epochs to $\tau = 5$. The local learning rate is $\eta_c = 0.1$, and the global learning rate is $\eta_s = 1.0$. For the CIFAR-10/100 and Fashion-MNIST (F-MNIST) datasets, we conduct training for a total of 100 communication rounds. We report the performance as the average accuracy over the final 10 rounds to ensure stability and reflect steady-state behavior.
We use mini-batch SGD as the optimizer for all our experiments. To account for variability in initialization, we repeat each experiment using three different random seeds. We report both the

mean and standard deviation of the performance metrics across these three runs.

A.9 HyperParameter Settings

Table 5: Experimental setup and hyperparameter configurations for the benchmark datasets.

Dataset	**Clients** (N)	**Rounds** (T)	**Epochs** (τ)	**Model Architecture**	**Optimizer** (η)
CIFAR-10	10	100	5	5-layer CNN	SGD (0.1)
F-MNIST	10	100	5	2-layer MLP	SGD (0.1)
CIFAR-100	20	100	5	ResNet-18	SGD (0.1)

Table 5 details the foundational experimental setup and hyperparameter configurations across our benchmark datasets. To ensure a rigorous evaluation across varying task complexities, we scale the model capacity and federation size accordingly: utilizing a 2-layer MLP for FMNIST and a 5-layer CNN for CIFAR-10 in a 10-client cross-silo environment, while deploying a deeper ResNet-18 backbone across 20 clients for the more challenging CIFAR-100 dataset. To explicitly isolate the performance impact of our proposed aggregation mechanism, local optimization dynamics are held strictly constant across all environments, employing standard SGD (learning rate $\eta = 0.1$) for $\tau = 5$ local epochs over $T = 100$ global communication rounds. For FedUCA, we set $n_s = 5$ and $\delta = 0.5$ by default. The value of ϵ for 10 clients is set to 0.1 and for 20 clients it is set to 0.05.

A.10 Experiments on Natural Language Data

Table 6: Performance comparison on the NLP dataset (20 Newsgroups). We use Bi-directional LSTM as the model Architecture. We run for 50 communication rounds with 10 clients. FedUCA outperforms all baselines.

Method	**Participation Rate (%)**	**Acc (%)**
IncFL	07.00	$14.25_{\pm 0.20}$
FedProx	07.00	$22.34_{\pm 0.40}$
AAggFF-S	07.60	$22.82_{\pm 0.30}$
FedAvg	06.50	$16.25_{\pm 0.18}$
FedUCA	**66.0**	$\mathbf{48.33_{\pm 0.57}}$

Table 6 evaluates FedUCA beyond vision tasks on the 20 Newsgroups NLP dataset using a bidirectional LSTM. FedUCA achieves 48.33% server accuracy, outperforming the strongest standard baseline, AAggFF-S, by 25.51 percentage points.

This shows that the utility-constrained aggregation continues to improve rational participation and global performance in an NLP setting. These results suggest that FedUCA's participation-management mechanism is not limited to vision datasets.

A.11 Empirical Verification of Jensen's Gap

Table 7 reports the empirical Jensen's gap $\sum_k \mathcal{J}_k^t$ across communication rounds on CIFAR-10, for five Dirichlet concentrations spanning four orders of magnitude. We summarize the per-round gap three ways: the unconditional mean over all 100 rounds, and the mean over the top 50 and top 25 rounds. The latter two characterize the rounds in which the surplus mechanism is most active, which are typically the rounds in which the optimizer is working hardest to satisfy IR constraints for a heterogeneous set of clients.

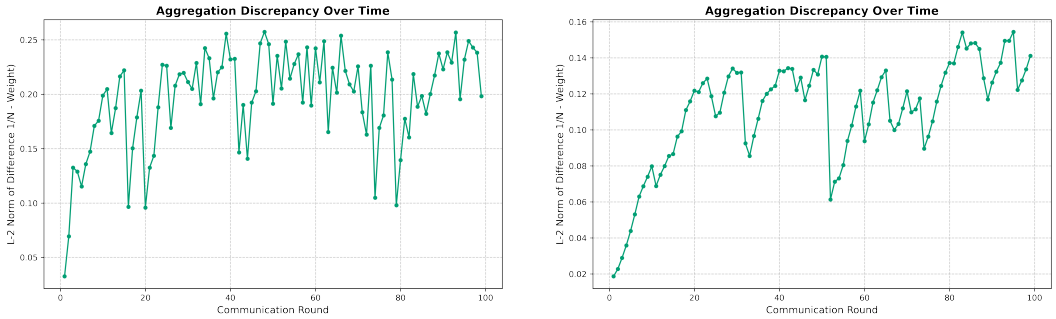
Two observations align the data with Proposition 3.2. First, the dependence on δ is non-monotonic: the gap vanishes at $\delta = 0.01$ (where the Dirichlet collapses to a vertex) and at $\delta = 100$ (where it concentrates near the uniform vector), and peaks at $\delta \in [0.5, 1.0]$. Second, this peak location correlates with the participation maximum in Figure 2, providing two independent measurements that confirm the theoretical prediction. The amplification across the Top-50 and Top-25 columns indicates

Table 7: Empirical verification of Proposition 3.2. We report the per-round total Jensen's gap $\sum_k \mathcal{J}_k^t$ measured over 100 communication rounds on CIFAR-10, summarized three ways: the mean over all rounds, the mean over the 50 rounds with the largest gap (Top-50), and the mean over the 25 rounds with the largest gap (Top-25). The Top-50 and Top-25 columns characterize the rounds in which the surplus mechanism is most active, typically rounds in which a substantial fraction of clients have binding IR constraints. The non-monotonic dependence on δ predicted by Proposition 3.2 is observed across all three columns: the gap is smallest at extreme concentrations ($\delta = 0.01$ and $\delta = 100$) and peaks at moderate values $\delta \in [0.5, 1.0]$.

Dirichlet Concentration (δ)	**Mean**	**Top-50 Mean**	**Top-25 Mean**
0.01	0.20	0.40	0.81
0.5	2.48	4.10	5.32
1.0	2.77	4.26	5.18
10	1.19	1.60	1.82
100	0.31	0.48	0.67

that the surplus is concentrated in the rounds where the IR mechanism most needs to deliver, rather than spread uniformly across training.

A.12 Empirical Verification of the Aggregation Discrepancy Bound



(a) Results on CIFAR-10

(b) Results on CIFAR-100

Figure 7: We plot the discrepancy $\|\mathbf{1}/N - \boldsymbol{\alpha}^t\|_2$ across communication rounds for FedUCA with $n_s = 5$ on (a) CIFAR-10 ($N = 10$) and (b) CIFAR-100 ($N = 20$). The gradual increase reflects growing heterogeneity in fitted client utilities ζ_k as local models diverge, which the IR mechanism translates into larger deviations from uniform aggregation; this is the intended behavior of utility-constrained aggregation, not a stability issue.

Lemma A.11 establishes the point-wise bound $\|\beta - \alpha^t\|^2 \leq (n_s - 1)/N$, which holds almost surely at every round and follows from the constraints of Optimization 7 alone, where $\beta = \mathbf{1}/N$. Figure 7 plots the realized per-round discrepancy $\|\beta - \alpha^t\|_2$; taking square roots gives the almost-sure ceiling $\|\beta - \alpha^t\|_2 \leq \sqrt{(n_s - 1)/N}$. We track $\|\beta - \alpha^t\|_2$ empirically across all communication rounds for FedUCA with $n_s = 5$ on CIFAR-10 ($N = 10$) and CIFAR-100 ($N = 20$); the empirical values stay well below the ceiling throughout, peaking near 0.25 against $\sqrt{0.4} \approx 0.63$ on CIFAR-10 and near 0.15 against $\sqrt{0.2} \approx 0.45$ on CIFAR-100.

Two additional observations are worth noting. First, the discrepancy increases in the early rounds and saturates after roughly round 30. This trajectory reflects the dynamics of the utility-constrained aggregation: in the initial rounds, fitted client utilities $\{\zeta_k\}$ are similar across clients, and the optimizer can satisfy IR constraints with weights close to uniform. As local models diverge under heterogeneous data, client utility curves become more heterogeneous and the optimizer assigns weights further from β to maintain participation. The behavior is consistent with the IR mechanism functioning as intended-deviations from uniform aggregation are the means by which Jensen's-gap surplus is allocated to satisfy individual rationality. Second, CIFAR-100 exhibits smaller discrepancy than CIFAR-10 in absolute terms, consistent with the $1/\sqrt{N}$ dependence of the ceiling $\sqrt{(n_s - 1)/N}$:

doubling the federation size scales the ceiling by $1/\sqrt{2}$, and the empirical reduction roughly tracks this prediction.

A.13 Impact due to Inflated Thresholds

Table 8: Impact of strategic threshold inflation on FedUCA (CIFAR-10, Dir(0.5, 1.0)). Universal inflation of the individual rationality threshold γ_k^t shrinks the server's feasibility space, leading to participation collapse and severe accuracy degradation.

Threshold Inflation	**Participation (%)**	**Accuracy (%)**
0% (truthful)	53.0	$66.89_{\pm 1.56}$
$+5\%$	40.0	$63.90_{\pm 1.23}$
$+10\%$	1.3	$10.80_{\pm 0.01}$

In the table 8, we evaluate FedUCA's robustness against strategic misreporting by simulating clients who attempt to free-ride by inflating their reported thresholds γ_k by 5% and 10%. We model universal inflation: under common knowledge that an incentive to inflate available to one client is available to all, so the rational analysis must consider the deviation in which every client inflates.

Inflated thresholds shrink the server's feasibility space (Optimization 7): the server cannot satisfy clients whose demands exceed what the aggregation can deliver, and drops them from the participating cohort. At $+5\%$ inflation, participation falls from 53% to 40% and global accuracy degrades by roughly 3 points. At $+10\%$, the optimization becomes broadly infeasible, participation collapses to 1.3%, and accuracy degrades to random guessing (10.80%).

If any client believes inflation is profitable, every rational client believes the same and inflates simultaneously, but universal inflation triggers federation-wide collapse, in which no client receives utility. Importantly, we do not claim truthful reporting as a dominant strategy. Our framework explicitly assumes truthful reporting, and this outcome is purely an observation from this specific experiment we ran, in which clients report and apply inflated thresholds in their own participation check. Within that setting, inflation degrades feasibility and participation. This finding holds strictly for this tested scenario, and we claim nothing beyond it. Establishing formal incentive compatibility remains a direction for future work.

A.14 Computational Complexity

All of our experiments are run on NVIDIA-RTX A5000 GPUs. For CIFAR-10, with 10 clients trained sequentially, the average time of execution for FedUCA is 5575s (the time spent in optimizer is 6s) for 100 communication rounds, while for FedAvg we have an execution time of 4583s. The peak memory utilization for FedUCA is 5654 MB, while for FedAvg we get a maximum utilization of 5633 MB. For FedProx, we have an execution time of 4591s with a peak utilization of 5634 MB. IncFL has a peak memory utilization of 5640 MB with an average execution time of 4718s. For AAggFF-S we have an average execution time of 4761s with a peak memory utilization of 5640 MB.

Note: These wall-clock numbers reflect our current implementation. In particular, FedUCA, which is our method, runs substantially more logging relative to the baselines, potentially inflating its measured time; the numbers should therefore be treated as an upper bound on FedUCA's cost. The method's only intrinsic overhead over a rationality-check-equipped baseline is (i) validation forward passes for utility estimation, and (ii) the server-side aggregation solve.

On the client-side, the local model training incurs $n_{epoch} \cdot |D_k^{train}| \cdot (c_f + c_b)$, $|D_k^{train}|$ denotes the number of training points, c_f, c_b denotes the cost of forward and backward passes.

FedUCA incurs extra computation only for utility estimation. If G_r denotes model interpolations, we need G_r evaluations of $\hat{u}_k(r)$ on the validation set at $G_r \cdot |D_k^{val}| \cdot c_f$ and a least-squares fit of three parameters. ($|D_k^{val}|$ denotes validation data points).

Empirically, on the CIFAR-10 data set, with $n_{epoch} = 5$, 5000 training points, and 800 validation points on A5000-GPU, this takes on average $\sim$0.5s time while the local training takes $\sim$5s of time, per communication round. Compared to FedAvg, it takes $\sim$0.5s extra time to fit the utility curve.

On the server side, FedUCA solves the Optimization (7). This addition is a convex program in $n_s \cdot N$ variables, independent of the model dimension d. For $N = 10$, $n_s = 5$, it totals 0.06 s per round, which is negligible.

On the impact of the number of clients (N) and strategies n_s, it should be noted that client-side utility computation is independent of them; only server optimization depends on them. We quantify it in Table 9.

Table 9: Run-time of the server Optimization as a function of N and n_s

Strategies (n_s)	**N=10**	**N=20**	**N=40**	**N=80**	**N=160**
5	6.0	10.0	18.3	36.0	72.3
10	6.2	10.4	19.4	40.0	79.7
20	6.5	11.5	22.0	45.0	92.0
30	7.1	12.4	25.0	51.0	101.0
40	7.3	13.6	28.0	58.0	123.0

Column header = Clients (N). Row header = Strategies (n_s). Entries = time per-round in seconds $\times$ 100.

Linear in clients: doubling N roughly doubles time across all strategy counts (6.0 $\rightarrow$ 10 $\rightarrow$ 18.3 $\rightarrow$ 36 $\rightarrow$ 72.3). Cost scales as $\mathcal{O}(N)$.

Sublinear in strategies (n_s): 8$\times$ more strategies ($n_s = 5 \rightarrow 40$) adds only minor overhead depending on N (at $N = 10$, the time goes from 6.0 s to 7.3s when n_s goes from 5 to 40, for 100 rounds).

A.15 On the rationality Check and alternate thresholding strategies

Relaxed Rationality check Clients may adopt a relaxed form of the threshold, replacing the rationality check in Eq. 1 by $acc(\boldsymbol{\theta}^t, \mathcal{D}_k^{\text{val}}) + m_k > acc(\boldsymbol{\theta}_k^{t-1}, \mathcal{D}_k^{\text{val}})$ and the IR constraint in Eq. 2 by $u_k(r_k; \zeta_k) + m_k > \gamma_k$, where $m_k \geq 0$ is a client-chosen safety margin. We report $m_k = 0$ throughout, which is the strictest setting: any $m_k > 0$ only relaxes the bar, and therefore it cannot decrease participation. The optimization meets the constraints at a lower value of the slack variable t_k, so clients clear the threshold with a smaller required m_k, and in particular at $m_k = 0$.

Alternate thresholding strategies To evaluate the relative benefits of FedUCA under an alternate threshold choice, we re-run all methods under a single shared threshold. FedUCA's native threshold is the accuracy read from the updated utility curve at the current aggregation weight, which the baselines cannot form since they lack the aggregation-weight channel (α_k^{t-1} and $u_k(r)$). We therefore use a channel-free threshold common to every method, a convex combination of the local and global validation accuracies: a client participates in the next round if the assigned model clears $(1 - l_w)\, acc(\boldsymbol{\theta}^t, \mathcal{D}_k^{\text{val}}) + l_w\, acc(\boldsymbol{\theta}_k^t, \mathcal{D}_k^{\text{val}})$ with $l_w = 0.4$, and each baseline uses the analogous check at round $t + 1$, $acc(\boldsymbol{\theta}^{t+1}, \mathcal{D}_k^{\text{val}}) > (1 - l_w)\, acc(\boldsymbol{\theta}^t, \mathcal{D}_k^{\text{val}}) + l_w\, acc(\boldsymbol{\theta}_k^t, \mathcal{D}_k^{\text{val}})$. Under this shared threshold the baselines still collapse to a few percent participation while FedUCA sustains 33.3% in Dirichlet(1, 1) setting (Table 10).

Table 10: Average accuracy (standard deviation as subscript) and average participation across 3 seeds under the shared convex-combination threshold ($l_w = 0.4$).

Method	CIFAR-10 Dirichlet(0.5, 1)		CIFAR-10 Dirichlet(1, 1)	
	Accuracy (%)	Participation (%)	Accuracy (%)	Participation (%)
FedAvg	$25.73_{1.19}$	26	$16.33_{1.27}$	2
FedProx	$22.40_{1.97}$	23	$12.62_{4.49}$	1
IncFL	$24.95_{0.43}$	28	$19.42_{1.55}$	2
AAggFF-S	$27.38_{3.31}$	28	$16.50_{6.96}$	1
FedAvg-WS	$47.13_{1.03}$	43	$47.65_{1.08}$	5
FedUCA (Ours)	$\mathbf{48.08}_{3.60}$	44	$\mathbf{66.95}_{2.68}$	33

A.16 Discussion on Comparison with Monetary Incentives

Monetary mechanisms are not applicable baselines in our setting, which assumes no payment channel; participation costs are performance-based, not financial. We nonetheless quantify the comparison analytically.

Both schemes deal with an individual-rationality constraint, but in different currencies: FedUCA generates the needed slack internally as Jensen surplus (from utility concavity at no cost), whereas a monetary mechanism supplies it externally as payment. Neither framework fixes an exchange rate between accuracy-utility and money. What is well posed is the price at which the monetary route buys the same slack/participation.

Setup. We use the standard cross-silo reward-budget mechanism of [39]: a reverse auction with even allocation (per round budget) $B_t = B_{total}/T$, clients ranked by quality-to-bid ratio, admitted down the ranking until the round budget is exhausted, and winners paid the critical price. We adopt their equal-quality configuration with per-round bids i.i.d. $U(a, b)$.

Cost is quadratic in participation.

For a fixed n, with equal quality, ranking by quality-to-bid coincides with ranking by bid, so the critical price is $b_{(n+1)}$ and the round cost of admitting n clients is $C(n) = n \cdot b_{(n+1)}$. For $U(a, b)$, $E[b_{(k)}] = a + (b - a)k/(N + 1)$, giving

$$E[C(n)] = n \cdot [a + (b - a)(n + 1)/(N + 1)]$$

quadratic in n with positive leading coefficient $(b-a)/(N+1)$: the marginal price of each additional client strictly increases. For $N = 10$, $T = 100$, bids $U(0.5, 1.5)$, matching FedUCA's 53% participation costs a total budget of roughly 520-680, and reaching 90% costs roughly 1270, so 1.8x the participation requires 2.4x the budget. The table below captures this in detail.

Table 11: Monetary budget required to buy the Participation

Participation	n	Total budget for 100 rounds
10%	1	68
30%	3	259
50%	5	523
60%	6	682
80%	8	1055
90%	9	1268

Full participation is excluded by design:

The critical price is set by the first excluded client, so a losing bidder must always remain; the mechanism saturates at $(N - 1)/N$ (90% for $N = 10$) at any budget. Purchased participation is capped structurally, not merely by spend.

A.17 Additional Experiments

A.17.1 Comparison with Personalized baselines

Table 12: Performance comparison of personalized baselines with Dirichlet(0.5,1.0).

	CIFAR-10		**FMNIST**	
Method	Accuracy	Participation (%)	Accuracy	Participation (%)
pFedMe [34]	$22.75_{\pm 0.04}$	11	$47.30_{\pm 0.20}$	42
Ditto [26]	$27.45_{\pm 0.06}$	8	$56.75_{\pm 0.30}$	37
Per-Fedavg [10]	$19.25_{\pm 0.03}$	1	$36.15_{\pm 0.02}$	1
FedUCA	$66.89_{\pm 1.56}$	53	$83.18_{\pm 0.40}$	53

In Table 12, we compare personalized baselines under the identical rationality check and protocol applied to all baselines. On CIFAR-10, all three personalized methods sustain low shared-model participation (Per-FedAvg 1%, Ditto 8%, pFedMe 11%) and correspondingly low shared-model accuracy (19 to 27%), versus FedUCA's 53% and 66.89%. On the easier F-MNIST, the effect is milder (Ditto 37%, pFedMe 42%) but still below FedUCA (53%, 83.18%). This is not a deficiency of these methods but a consequence of their objective: each client's personalized model raises its acceptance bar for the global model, so clients reject the shared model and it receives too few updates to train; this is consistent with the participation collapse in Figure 1.

A.17.2 Experiments with cubic utility fit and loss as utility

Table 13: Performance on the CIFAR-10 dataset with Dirichlet partitioning $(0.5, 1.0)$. Results are reported as mean_{std}.

Method	**Accuracy (%)**
Loss as utility and Quadratic fit	$66.10_{1.30}$
Accuracy as utility and Cubic Fit	$66.50_{1.10}$
Accuracy as utility and Quadratic Fit (Default)	$66.89_{1.56}$

Regarding admissible metrics: not every metric is admissible, only those whose measured curve is concave in the interpolation weight $r \in [0, 1]$. The reasoning stems from the interpolation path itself. At $r = 0$, the client's own model is excluded from the aggregate, and at $r = 1$, the rest of the federation is excluded. Because both endpoints discard critical information, the curve naturally peaks at an interior point for any metric that reflects overall model quality. Validation accuracy satisfies this concavity condition, as shown by our plots in Section A.4, and negative validation loss works similarly well. To further validate our approach, we performed an experiment fitting a 3rd-order (cubic) polynomial. Because the empirical data exhibits diminishing returns, even the 3rd-order polynomial behaves concavely within the region of interest $[0, 1]$. As verified in the table 13, the overall performance remains largely unaffected whether we use negative validation loss as the utility metric or apply a cubic fit.

A.18 Proofs of Propositions

Proposition A.1. *Suppose $\epsilon = 0$, and all utilities $u_k(.)$ are concave then the solution to optimization 7 satisfies either $\sum_{i=1}^{n_s} s_i u_k(r_k^i) = \gamma_k - t_k$, with $t_k > 0$ or $\sum_{i=1}^{n_s} s_i u_k(r_k^i) \geq \gamma_k$, with $t_k = 0$. In the former case, the constraint is satisfied when $u_k(r_k^i) = \frac{s_i}{\sum_{j=1}^{n_s} s_j^2}(\gamma_k - t_k)$. If we further assume that $u_k(.)$ is strictly concave, the optimal solution $\mathbf{R}^*, \mathbf{t}^*$ is unique, if the former holds $\forall k \in [N]$.*

Proof. We form the lagrangian $L(\mathbf{t}, \mathbf{R})$ as

$$L(\mathbf{t}, \mathbf{R}, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \beta_k^i, \xi_k, \mu_k) = \frac{1}{2}\|\mathbf{t}\|_2^2 + \boldsymbol{\lambda_1}^T(\mathbf{R}\mathbf{1}_N - \mathbf{1}_{n_s}) + \boldsymbol{\lambda_2}^T(\mathbf{R}^T\mathbf{1}_{n_s} - \frac{n_s}{N}\mathbf{1}_N) \\ - \sum_{i=1,k=1}^{n_s,N} \beta_k^i r_k^i - \sum_{k=1}^{N} \xi_k (\sum_{i=1}^{n} s_i u_k(r_k^i) - \gamma_k + t_k) - \sum_{k=1}^{N} \mu_k t_k \tag{9}$$

by setting $\nabla_{\mathbf{t}} L = 0$ we have the following

$$t_k = \mu_k + \xi_k \tag{10}$$

by setting $\nabla_{r_k^i} L = 0$ we have the following, $\forall i \in [n_s], \forall k \in [N]$

$$\boldsymbol{\lambda}_1(i) + \boldsymbol{\lambda}_2(k) = \beta_k^i + \xi_k s_i \nabla_{r_k^i} u_k(r_k^i) \tag{11}$$

By the feasibility of equality constraints, we have

$$\mathbf{R}\mathbf{1}_N = \mathbf{1}_{n_s} \tag{12}$$

$$\mathbf{R}^T \mathbf{1}_{n_s} = \frac{n_s}{N} \mathbf{1}_N \tag{13}$$

By the feasibility of inequality constraints, we have $\forall k \in [N]$

$$\sum_{i=1}^{n} s_i u_k(r_k^i) \geq \gamma_k - t_k \tag{14}$$

By the complementary slackness of the inequality constraints, we have $\forall k \in [N]$

$$(\sum_{i=1}^{n} s_i u_k(r_k^i) - \gamma_k + t_k)\xi_k = 0 \tag{15}$$

$$\mu_k t_k = 0 \tag{16}$$

And $\forall k \in [N], \forall i \in [n_s]$

$$r_k^i \beta_k^i = 0 \tag{17}$$

From Eq. 16 we see that when $t_k > 0$ the Eq. 10 will simplify to $t_k = \xi_k$. Since we have assumed $t_k > 0$ and so is ξ_k.

But the slackness assumption in Eq. 15 implies that

$$\sum_{i=1}^{n} s_i u_k(r_k^i) = \gamma_k - t_k \tag{18}$$

If $\xi_k = 0$, then this will imply from the condition in. 15 we have the following

$$\sum_{i=1}^{n} s_i u_k(r_k^i) \geq \gamma_k - t_k \tag{19}$$

But From the condition in. 10 we have $t_k = \mu_k$ but the slackness also implies that $t_k \mu_k = 0$. This implies that $t_k = \mu_k = \xi_k = 0$.

The Eq. 19 simplifies to the following

$$\sum_{i=1}^{n} s_i u_k(r_k^i) \geq \gamma_k \tag{20}$$

From Eq. 20 and. 18 we conclude the first part of the proof.

The second part of the proof can be established by writing as system of equations with unknowns in $u_k(r_k^i)$. It is easy to verify that the condition $u_k(r_k^i) = \frac{s_i}{\sum_{j=1}^{n_s} s_j^2}(\gamma_k - t_k)$ satisfies the condition $\sum_{i=1}^{n} s_i u_k(r_k^i) = \gamma_k - t_k$.

Finally, we can prove this uniqueness by contradiction. Suppose there are two distinct optimal strategy matrices, $\mathbf{R}_1^*$ and $\mathbf{R}_2^*$, that yield the same optimal (minimal) slack $\boldsymbol{t}^*$. Owing to the strict concavity of the utility function u_k considered for a client k, let us form a convex combination

$$\mathbf{R} = \alpha \mathbf{R}_1^* + (1 - \alpha)\mathbf{R}_2^*, \forall \alpha \in (0, 1).$$

Evaluating the utility constraint under this mixed strategy yields:

$$\sum_{i=1}^{n_s} s_i u_k(r_k^i) > \alpha \sum_{i=1}^{n_s} s_i u_k(r_{1,k}^{*i}) + (1 - \alpha) \sum_{i=1}^{n_s} s_i u_k(r_{2,k}^{*i})$$

Finally we have

$$\sum_{i=1}^{n_s} s_i u_k(r_k^i) > \alpha(\gamma_k - t_k^*) + (1-\alpha)(\gamma_k - t_k^*) = \gamma_k - t_k^*$$

This demonstrates that under the combined strategy $\mathbf{R}$, the utility strictly exceeds the required threshold bounded by t_k^*. Consequently, the slack t_k^* can be further reduced while still satisfying the individual rationality constraint. This directly contradicts the premise that $\boldsymbol{t}^*$ is the optimal (minimal) slack. Thus, the strategy solution r_k^i generated by the server must be unique.

□

Proposition A.2. *Let the expected Jensen's gap for client k be defined as $\bar{\mathcal{J}}_k(\delta) \coloneqq \mathbb{E}_{\mathbf{s}\sim Dir(\delta)}[\mathcal{J}_k(\mathbf{s})]$ for a concentration parameter $\delta > 0$. The instantaneous gap is given by $\mathcal{J}_k(\mathbf{s}) \coloneqq u_k\big(\sum_{i=1}^{n_s} s_i r_k^i\big) - \sum_{i=1}^{n_s} s_i u_k(r_k^i)$, where $u_k(\cdot)$ is a strictly concave utility function and the aggregation weights r_k^i are obtained by solving Optimization 7, where the constraint is binding. The expected gap $\bar{\mathcal{J}}_k(\delta)$ vanishes at the extreme concentration regimes, specifically when $\delta \leq \delta_{min}$ or $\delta \geq \delta_{max}$ (where $\delta_{min} \ll 1$ and $\delta_{max} \gg 1$). Furthermore, $\bar{\mathcal{J}}_k(\delta)$ is a continuous function of δ on the open interval $(\delta_{min}, \delta_{max})$.*

Proof. We establish that $\bar{\mathcal{J}}_k(\delta)$ is well-defined and continuous on $(\delta_{\min}, \delta_{\max})$.

The aggregation matrix $\mathbf{R}^*(\mathbf{s})$ is obtained from the parametric optimization problem in (7). The feasible set $\mathcal{R}$ is a non-empty compact polytope defined by linear equality and inequality constraints, and is independent of $\mathbf{s} \in \Delta$.

Define

$$V(\mathbf{R}, \mathbf{s}) = \frac{1}{2}\|\mathbf{t}(\mathbf{R}, \mathbf{s})\|_2^2, \qquad t_k(\mathbf{R}, \mathbf{s}) = \max\Big(0, \gamma_k - \sum_{i=1}^{n_s} s_i u_k(r_k^i)\Big).$$

For fixed $\mathbf{s}$, $V(\mathbf{R}, \mathbf{s})$ is continuous in $\mathbf{R}$ because u_k is continuous and the max operator is continuous. For fixed $\mathbf{R}$, $V(\mathbf{R}, \mathbf{s})$ is continuous, hence measurable, in $\mathbf{s}$. Thus, V is a Carathéodory function. Since the feasible correspondence is constant, compact-valued, and measurable, the Measurable Maximum Theorem. (Refer to Theorem 18.19 of [3]) implies that the argmin correspondence admits a Lebesgue-measurable selection $\mathbf{R}^*(\mathbf{s})$.

Each coordinate $r_k^i(\mathbf{s})$ is a coordinate projection of $\mathbf{R}^*(\mathbf{s})$. Since coordinate projections are continuous, each $r_k^i(\mathbf{s})$ is measurable. Therefore, because u_k is continuous, the mappings $u_k(r_k^i(\mathbf{s}))$ and

$$u_k\left(\sum_{i=1}^{n_s} s_i r_k^i(\mathbf{s})\right)$$

are measurable. Hence,

$$\mathcal{J}_k(\mathbf{s}) = u_k\left(\sum_{i=1}^{n_s} s_i r_k^i(\mathbf{s})\right) - \sum_{i=1}^{n_s} s_i u_k(r_k^i(\mathbf{s}))$$

is measurable.

Moreover, since $r_k^i(\mathbf{s}) \in [0,1]$ and u_k is continuous on the compact interval $[0,1]$, there exists $M_u < \infty$ such that $|u_k(r)| \leq M_u$ for all $r \in [0,1]$. Thus,

$$|\mathcal{J}_k(\mathbf{s})| \leq \left|u_k\left(\sum_i s_i r_k^i(\mathbf{s})\right)\right| + \sum_i s_i |u_k(r_k^i(\mathbf{s}))| \leq 2M_u.$$

Let $M_J = 2M_u$.

Now write

$$\bar{\mathcal{J}}_k(\delta) = \int_\Delta \mathcal{J}_k(\mathbf{s}) p(\mathbf{s}; \delta)\, d\mathbf{s}, \qquad p(\mathbf{s}; \delta) = \frac{1}{B(\delta)} \prod_{i=1}^{n_s} s_i^{\delta - 1}.$$

Let $\delta_n \to \delta \in (\delta_{\min}, \delta_{\max})$. Since $\{\delta_n\}$ converges, there exist constants $0 < a < b < \infty$ such that $\delta_n, \delta \in [a, b]$ for all sufficiently large n. For $\mathbf{s}$ in the interior of the simplex,

$$p(\mathbf{s}; \delta_n) \to p(\mathbf{s}; \delta)$$

pointwise. Boundary points have Lebesgue measure zero on the simplex, so this convergence holds almost everywhere.

Since $B(\cdot)$ is continuous and strictly positive on $[a, b]$, there exists

$$M_B = \max_{\eta \in [a,b]} \frac{1}{B(\eta)} < \infty.$$

Also, for $s_i \in (0, 1)$ and $\delta_n \geq a$,

$$s_i^{\delta_n - 1} \leq s_i^{a-1}.$$

Therefore,

$$|\mathcal{J}_k(\mathbf{s}) p(\mathbf{s}; \delta_n)| \leq M_J M_B \prod_{i=1}^{n_s} s_i^{a-1} \coloneqq G(\mathbf{s}).$$

The function G is integrable over Δ because $a > 0$ and

$$\int_\Delta \prod_{i=1}^{n_s} s_i^{a-1} \, d\mathbf{s} = B(a) < \infty.$$

Hence, by the Dominated Convergence Theorem,

$$\lim_{n \to \infty} \bar{\mathcal{J}}_k(\delta_n) = \int_\Delta \mathcal{J}_k(\mathbf{s}) p(\mathbf{s}; \delta) \, d\mathbf{s} = \bar{\mathcal{J}}_k(\delta).$$

Thus, $\bar{\mathcal{J}}_k(\delta)$ is continuous on $(\delta_{\min}, \delta_{\max})$.

We now consider the extreme concentration regimes.

As $\delta \to 0$, only one of $s_i \approx 1$ and the remaining $s_j \approx 0, \forall j \neq i$. This amounts to using the single strategy, which is a deterministic case. In this case, we have the following.

$$u_k(\sum_{i=1}^{n_s} s_i r_k^i) \approx u_k(r_k^i) \tag{21}$$

$$\sum_{i=1}^{n_s} s_i u_k(r_k^i) \approx u_k(r_k^i) \tag{22}$$

These two equations, Eq. 21 and Eq. 22, imply the zero Jensen's gap.

As $\delta \to \infty$, all the strategies are $s_i \approx \frac{1}{n_s}$. From the proposition A.1 we know that the solution r_k^i admits the following equality

$$u_k(r_k^i) = \frac{s_i}{\sum_{j=1}^{n_s} s_j^2} (\gamma_k - t_k) \tag{23}$$

Since all the s_i admits to the same value the above equation simplifies to

$$u_k(r_k^i) = \gamma_k - t_k$$

This is independent of i, Since t_k is the slack, its adjusted so that the solution exists and satisfies the constraint $\sum_{i=1}^{n_s} r_k^i = \frac{n_s}{N}$. Since they all are equal we have $r_k^i = \frac{1}{N}$.

We have $u_k(\sum_{i=1}^{n_s} s_i r_k^i) = u_k(\frac{1}{N})$ and $\sum_{i=1}^{n_s} s_i u_k(r_k^i) = u_k(\frac{1}{N})$. Resulting in the zero Jensen's gap.

Hence, the expected gap $\bar{\mathcal{J}}_k(\delta)$ vanishes in both extremes. □

Proposition A.3. *Suppose $u_k(\cdot)$ is concave, $\mathcal{J}_k(s) \geq 0$, and $t \leq \mathop{\mathbb{E}}_{\mathbf{s}\sim Dir(\delta)}[\mathcal{J}_k(\mathbf{s})]$ and $n_s \leq N$ then $Pr[\mathcal{J}_k(\mathbf{s}) \geq t] \geq \frac{(\mathbb{E}[\mathcal{J}_k(\mathbf{s})]-t)^2}{{M_J}^2}$, for some M_J and $M_J \leq 1$, if the utility is accuracy.*

Proof. By Paley-Zygmund inequality we have the following for any positive random variable X and a constant β we have the following

$$Pr[X \geq \beta\mathbb{E}[X]] \geq (1-\beta)^2 \frac{\mathbb{E}^2[X]}{\mathbb{E}[X^2]} \tag{24}$$

Since this is a Jensen's gap and its non negative and its a random variable, and by assumption, we have the $\frac{t}{\mathbb{E}[\mathcal{J}_k(\mathbf{s})]} \leq 1$

Now set $X = \mathcal{J}_k(\mathbf{s})$ and $\beta = \frac{t}{\mathbb{E}[\mathcal{J}_k(\mathbf{s})]}$ and substitute in the Eq. 24.

$$Pr[\mathcal{J}_k(\mathbf{s}) \geq t] \geq \frac{(\mathbb{E}[\mathcal{J}_k(\mathbf{s})] - t)^2}{\mathbb{E}[\mathcal{J}_k^2(\mathbf{s})]} \tag{25}$$

We note that minimum value of $\mathcal{J}_k(\mathbf{s})^2$ is 0 and from the previous proposition it can be seen that $\mathcal{J}_k(\mathbf{s}) \leq M_J$. The reason we claim $M_J \leq 1$ is since, if the utillity is accuracy the the max value is 1 and a lower quantity is subtracted from the same.

Plugging the above in the Eq. 25 gives the desired result.

□

A.19 Assumptions for Convergence analysis

We state our assumptions as below; these assumptions are commonly made in the FL literature such as [16, 30]

Assumption A.4. $\|\nabla f_k(\mathbf{x}) - \nabla f_k(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$
This assumption implies that the loss function of the client k, i.e, $f_k(.)$, is Lipschitz Smooth.

Assumption A.5. $\mathbb{E}_\xi\|\nabla f_k(\mathbf{w}, \xi) - \nabla f_k(\boldsymbol{\theta})\|^2 \leq \sigma^2$ for all $k \in [N]$ and $\mathbb{E}_\xi[\nabla f_k(\boldsymbol{\theta}, \xi)] = \nabla f_k(\boldsymbol{\theta})$
This implies that the variance of the gradients is bounded and the gradients are unbiased.

Assumption A.6. $\|\nabla f_k(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|^2 \leq \sigma_g^2$.
This implies that the dissimilarity between local and global gradients is bounded, commonly referred to as the bounded gradient dissimilarity assumption.

Assumption A.7. $\|\nabla f_k(\boldsymbol{\theta})\|^2 \leq G$ for all $k \in [N]$.
This assumption implies that local gradients are bounded. This is the same as loss functions being Lipschitz continuous and differentiable [30].

Assumption A.8. The staleness of the client model at the server is bounded by a maximum delay of ν_{max}.

Assumption A.9. Let S_t denote the number of clients participating in the round t. The we assume $1 \leq S \leq |S_t| \leq N$.

The above assumption implies that while the exact subset of participating clients dynamically changes based on the state-dependent mechanism, the number of active clients per round is strictly bounded below by S.

A.20 Update rules for the FedUCA

The global objective function is $f(\boldsymbol{\theta}) = \sum_{k=1}^{N} \beta_k f_k(\boldsymbol{\theta})$.

We further define the following two notions:

$$\text{Averaged stochastic Gradient:} \quad \mathbf{d}_k^t = \frac{1}{\tau}\sum_{r=0}^{\tau-1} \nabla f_k(\boldsymbol{\theta}_k^{(t,r)}, \xi_k^{t,r}), \tag{26}$$

$$\text{Averaged Gradient:} \quad \mathbf{h}_k^t = \frac{1}{\tau}\sum_{r=0}^{\tau-1} \nabla f_k(\boldsymbol{\theta}_k^{(t,r)}). \tag{27}$$

$$\text{Client Update:} \quad \boldsymbol{\theta}_k^{t,\tau} = \boldsymbol{\theta}^{t,0} - \eta_c \tau \mathbf{d}_k^t \tag{28}$$

Then, the update of the global model between two rounds is as follows:

$$\boldsymbol{\theta}^{t+1,0} = \boldsymbol{\theta}^{t,0} - \eta\tau \left[\sum_{k\in S_t} \alpha_k^t \mathbf{d}_k^t + \sum_{k\in S_t^c} \alpha_k^t \mathbf{y}_k^t\right] \tag{29}$$

We further define the total variance σ_{total}^2 as

$$\sigma_{total}^2 \coloneqq \sigma^2 + \sigma_g^2 + G \tag{30}$$

In the above Eq. 29 $\mathbf{y}_k^t \coloneqq \mathbf{d}_k^{t-\nu_k^t}$ and $\eta \coloneqq \eta_s\eta_c$. In the ν_k^t is the stale update of the client k relative to the current round t.

Our analysis utilizes the proof techniques of the FedBuff [30].

A.21 Convergence Analysis

Lemma A.10 (Expected-Aggregation Discrepancy)**.** *Under the constraints of Optimization 7, if the strategies s are drawn from a symmetric Dirichlet distribution $Dir(\delta \mathbf{1}_{n_s})$, the expected aggregation discrepancy is strictly bounded by:*

$$\mathbb{E}_s||\frac{1}{N}\mathbf{1}_N - \boldsymbol{\alpha}^t||^2 \leq \frac{n_s - 1}{N(n_s\delta + 1)}, \tag{31}$$

Further $\mathcal{E}_{agg} \leq C\frac{n_s-1}{S(n_s\delta+1)}$, where $\mathcal{E}_{agg} = \frac{12KNG}{S}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2\right)$.

Proof. From the constraints of Optimization 7, the strategy matrix $\mathbf{R}$ satisfies the column-sum constraint $\mathbf{1}_{n_s}^T\mathbf{R} = \frac{n_s}{N}\mathbf{1}_N^T$. Taking the transpose yields $\mathbf{R}^T\mathbf{1}_{n_s} = \frac{n_s}{N}\mathbf{1}_N$. Since the mean of a symmetric Dirichlet distribution is $\mathbb{E}[\mathbf{s}] = \frac{1}{n_s}\mathbf{1}_{n_s}$, mapping this expected strategy through any feasible matrix $\mathbf{R}$ deterministically recovers the uniform target:

$$\mathbf{R}^T\mathbb{E}[\mathbf{s}] = \mathbf{R}^T\left(\frac{1}{n_s}\mathbf{1}_{n_s}\right) = \frac{1}{n_s}\left(\frac{n_s}{N}\mathbf{1}_N\right) = \frac{1}{N}\mathbf{1}_N = \frac{1}{N}\mathbf{1}_N \tag{32}$$

Consequently, the realized aggregation weights $\boldsymbol{\alpha}^t = \mathbf{R}^T s$ can be expressed as a zero-mean perturbation around the ideal uniform target:

$$\boldsymbol{\alpha}^t = \mathbf{R}^T\mathbf{s} = \mathbf{R}^T(\mathbb{E}[\mathbf{s}] + (\mathbf{s} - \mathbb{E}[\mathbf{s}])) = \frac{1}{N}\mathbf{1}_N + \mathbf{R}^T(\mathbf{s} - \mathbb{E}[\mathbf{s}]) \tag{33}$$

Thus, the discrepancy norm simplifies exactly to $||\frac{1}{N}\mathbf{1}_N - \boldsymbol{\alpha}^t||_2^2 = ||\mathbf{R}^T(\mathbf{s} - \mathbb{E}[\mathbf{s}])||_2^2$.

Applying the sub-multiplicative property of induced matrix norms, we bound the discrepancy:

$$||\mathbf{R}^T(\mathbf{s} - \mathbb{E}[\mathbf{s}])||_2^2 \leq ||\mathbf{R}^T||_2^2 \cdot ||\mathbf{s} - \mathbb{E}[\mathbf{s}]||_2^2 \tag{34}$$

Because $\mathbf{R}$ contains strictly non-negative elements, we bound its spectral norm ($||\mathbf{R}^T||_2^2$) using the inequality $||\mathbf{A}||_2^2 \leq ||\mathbf{A}||_1||\mathbf{A}||_\infty$:

- $||\mathbf{R}^T||_1 = \max_i \sum_j \mathbf{R}_{i,j} = \max(\text{row sums of } \mathbf{R}) = 1.$
- $||\mathbf{R}^T||_\infty = \max_j \sum_i \mathbf{R}_{i,j} = \max(\text{col sums of } \mathbf{R}) = \frac{n_s}{N}.$

Therefore, for *any* feasible optimization outcome, the spectral norm is deterministically bounded by $||\mathbf{R}^T||_2^2 \leq \frac{n_s}{N}$.

Taking the expectation over the Dirichlet distribution $\mathbf{s}$:

$$\mathbb{E}_{\mathbf{s}}||\frac{1}{N}\mathbf{1}_N - \boldsymbol{\alpha}^t||_2^2 \leq \frac{n_s}{N}\mathbb{E}_{\mathbf{s}}||\mathbf{s} - \mathbb{E}[\mathbf{s}]||_2^2 \tag{35}$$

The expectation term is exactly the trace of the Dirichlet covariance matrix Σ_s. For $\mathbf{s} \sim Dir(\delta\mathbf{1}_{n_s})$, the trace is known analytically:

$$\text{Tr}(\Sigma_s) = \frac{n_s - 1}{n_s(n_s\delta + 1)} \tag{36}$$

Substituting this into the inequality yields the final bound:

$$\mathbb{E}_s||\frac{1}{N}\mathbf{1}_N - \boldsymbol{\alpha}^t||_2^2 \leq \frac{n_s}{N}\left(\frac{n_s - 1}{n_s(n_s\delta + 1)}\right) = \frac{n_s - 1}{N(n_s\delta + 1)} \tag{37}$$

Finally substituting this, in $\mathcal{E}_{agg}$ gives the bound

$\mathcal{E}_{agg} \leq C\frac{n_s - 1}{S(n_s\delta+1)}$, where $C = 12KG$. □

Lemma A.11 (Point-wise Aggregation Discrepancy Bound)**.** *Under the constraints of Optimization. 7 with $\beta_k = 1/N$, $n_s > 1$ the FedUCA aggregation weights satisfy*

$$\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2 \leq \frac{n_s - 1}{N} \quad \textit{almost surely, for every } t. \tag{38}$$

In particular, $\mathcal{E}_{agg} \leq O\left(\frac{n_s - 1}{S}\right)$, where $\mathcal{E}_{agg} \coloneqq \frac{12KNG}{S}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2\right)$

Proof. For any $x \in [0, M]$ and any $\mu \in [0, M]$, we have

$$(x - \mu)^2 = x^2 - 2\mu x + \mu^2.$$

Since $0 \leq x \leq M$, it follows that $x^2 \leq Mx$. Hence,

$$(x - \mu)^2 \leq Mx - 2\mu x + \mu^2 = (M - 2\mu)x + \mu^2.$$

Taking $M = \frac{n_s}{N}$ and $\mu = \frac{1}{N}$, we obtain

$$\left(\alpha_k - \frac{1}{N}\right)^2 \leq \left(\frac{n_s}{N} - \frac{2}{N}\right)\alpha_k + \frac{1}{N^2} = \frac{n_s - 2}{N}\alpha_k + \frac{1}{N^2}.$$

Summing over $k \in [N]$ and using $\sum_{k=1}^N \alpha_k = 1$ gives

$$\sum_{k=1}^N\left(\alpha_k - \frac{1}{N}\right)^2 \leq \frac{n_s - 2}{N}\sum_{k=1}^N\alpha_k + \frac{N}{N^2} = \frac{n_s - 2}{N} + \frac{1}{N} = \frac{n_s - 1}{N}.$$

This implies that

$$\mathcal{E}_{\text{agg}} \leq 12KG\left(\frac{n_s - 1}{S}\right).$$

□

Lemma A.12. *The gradients of the client's loss $\nabla f_k(\boldsymbol{\theta}_k^{t,r}; \xi_k^{t,r})$ satisfy the bound*

$$\mathbb{E}\left\|\nabla f_k(\boldsymbol{\theta}_k^{t,r}; \xi_k^{t,r})\right\|^2 \leq 3\sigma_{total}^2$$

Proof. The proof is provided in [30] □

Lemma A.13. *The iterates generated by FedUCA $\boldsymbol{\theta}^t$ satisfy the following*

$$\mathbb{E}_{\xi_t}\left\|\nabla f_k(\boldsymbol{\theta}^t) - \mathbf{h}_k^t\right\|^2 \leq 3\eta_c^2L^2\sigma_{total}^2\tau(\tau - 1)$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\xi_t}\left\|\nabla f_k(\boldsymbol{\theta}^t) - \mathbf{h}_k^t\right\|^2 &= \mathbb{E}_{\xi_t}\left\|\frac{1}{\tau}\sum_{r=0}^{\tau-1}\nabla f_k(\boldsymbol{\theta}^t) - \nabla f_k(\boldsymbol{\theta}_k^{(t,r)})\right\|^2 \\
&\leq \frac{1}{\tau}\sum_{r=0}^{\tau-1}\mathbb{E}_{\xi_t}\left\|\nabla f_k(\boldsymbol{\theta}^t) - \nabla f_k(\boldsymbol{\theta}_k^{(t,r)})\right\|^2 \\
&\overset{a}{\leq} \frac{L^2}{\tau}\sum_{r=0}^{\tau-1}\mathbb{E}_{\xi_t}\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^{(t,r)}\right\|^2 \\
&= \frac{L^2\eta_c^2}{\tau}\sum_{r=0}^{\tau-1}\mathbb{E}_{\xi_t}\left\|\sum_{m=0}^{r-1}\nabla f_k(\boldsymbol{\theta}_k^{t,m};\xi_k^{t,m})\right\|^2 \\
&\overset{b}{\leq} \frac{L^2\eta_c^2}{\tau}\sum_{r=0}^{\tau-1} r\sum_{m=0}^{r-1}\mathbb{E}_{\xi_t}\left\|\nabla f_k(\boldsymbol{\theta}_k^{t,m};\xi_k^{t,m})\right\|^2 \\
&\overset{c}{\leq} \frac{3L^2\eta_c^2}{\tau}\sum_{r=0}^{\tau-1} r^2(\sigma^2+\sigma_g^2+G) \\
&\leq 3L^2\eta_c^2\sigma_{total}^2(\tau-1)\tau
\end{aligned}$$

Here a follows from the Assumption. A.4. b follows from the client update. 28, finally c is via the Lemma. A.12. □

Because the client's participation decision I_k^t is evaluated deterministically based on the Optimization parameters and rationality thresholds broadcasted in the history H_t, it is formally H_t-measurable. Consequently, I_k^t behaves as a constant under the conditional expectation $\mathbb{E}[\cdot \mid H_t]$, rendering it independent of the specific realization α_k^t for that round. This measurability allows us to treat the active cohort as a fixed set for the round t analysis, cleanly isolating the stochasticity purely to the server's draw.

Lemma A.14. *The iterates generated by FedUCA $\boldsymbol{\theta}^t$ satisfy the following*

$$\mathbb{E}\left\|\sum_{k=1}^{N}\alpha_k^t(I_k^t\mathbf{h}_k^t)\right\|^2 \leq 9L^2\eta_c^2\sigma_{total}^2(\tau-1)\tau + 3\sigma_g^2 + 3\mathbb{E}\left[\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2\right]$$

Proof. Let $V_{act} = \sum_{k=1}^{N}\alpha_k^t(I_k^t\mathbf{h}_k^t)$. Because the broadcasted weights reside on the probability simplex ($\alpha_k^t \geq 0$, $\sum_{k=1}^{N}\alpha_k^t = 1$), we exploit the convexity of the squared L_2-norm. Applying Jensen's Inequality pushes the norm inside the summation:

$$\mathbb{E}\|V_{act}\|^2 \leq \mathbb{E}\left[\sum_{k=1}^{N}\alpha_k^t\left\|I_k^t\mathbf{h}_k^t\right\|^2\right] \tag{39}$$

Since $I_k^t \in \{0,1\}$, we have $\|I_k^t\mathbf{h}_k^t\|^2 = I_k^t\|\mathbf{h}_k^t\|^2$. By using the relaxed triangle inequality, we decompose this into the following

$$\begin{aligned}
\mathbb{E}\|V_{act}\|^2 \leq\ & 3\,\underbrace{\mathbb{E}\left[\sum_{k=1}^{N}\alpha_k^t I_k^t\left\|\mathbf{h}_k^t - \nabla f_k(\boldsymbol{\theta}^{(t,0)})\right\|^2\right]}_{F_1} + 3\,\underbrace{\mathbb{E}\left[\sum_{k=1}^{N}\alpha_k^t I_k^t\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2\right]}_{F_2} \\
&+ 3\,\underbrace{\mathbb{E}\left[\sum_{k=1}^{N}\alpha_k^t I_k^t\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2\right]}_{F_3}
\end{aligned} \tag{40}$$

$$
\begin{aligned}
F_1 &= \mathbb{E}\left[\sum_{k=1}^{N} I_k^t \alpha_k^t \left\| \mathbf{h}_k^t - \nabla f_k(\boldsymbol{\theta}^{(t,0)}) \right\|^2\right] \\
&= \mathbb{E}_{H_t}\left[\sum_{k=1}^{N} \mathbb{E}[I_k^t \alpha_k^t] \mathbb{E}\left[\left\| \mathbf{h}_k^t - \nabla f_k(\boldsymbol{\theta}^{(t,0)}) \right\|^2 \;\middle|\; H_t\right]\right] \\
&\leq \left(3L^2\eta_c^2\sigma_{total}^2(\tau-1)\tau\right) \times \mathbb{E}\left[\sum_{k=1}^{N} I_k^t \alpha_k^t\right] \leq 3L^2\eta_c^2\sigma_{total}^2(\tau-1)\tau
\end{aligned} \tag{41}
$$

The first equality follows via the conditional expectation. The second holds because the SGD path variance is independent of the realization given H_t. The final inequality holds due to the simplex constraint, which guarantees $\sum I_k^t \alpha_k^t \leq \sum \alpha_k^t = 1$.

$$
F_2 = \mathbb{E}\left[\sum_{k=1}^{N} I_k^t \alpha_k^t \left\| \nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \nabla f(\boldsymbol{\theta}^{(t,0)}) \right\|^2\right] \leq \sigma_g^2 \mathbb{E}\left[\sum_{k=1}^{N} I_k^t \alpha_k^t\right] \leq \sigma_g^2 \tag{42}
$$

$$
F_3 = \mathbb{E}\left[\left\| \nabla f(\boldsymbol{\theta}^{(t,0)}) \right\|^2 \left(\sum_{k=1}^{N} I_k^t \alpha_k^t\right)\right] \leq \mathbb{E}\left\| \nabla f(\boldsymbol{\theta}^{(t,0)}) \right\|^2 \tag{43}
$$

Combining F_1, F_2, F_3 we conclude the proof. □

Lemma A.15.

$$
\sum_{k=1}^{N} \mathbb{E}\left\| \mathbf{h}_k^{t-\nu_k^t} \right\|^2 \leq 9NL^2\eta_c^2\sigma_{total}^2(\tau-1)\tau + 3N\sigma_g^2 + 3NG \tag{44}
$$

Proof.

$$
\begin{aligned}
\sum_{k=1}^{N} \mathbb{E}\left\| \mathbf{h}_k^{t-\nu_k^t} \right\|^2 &= \sum_{k=1}^{N} \mathbb{E}\Big\| \left(\mathbf{h}_k^{t-\nu_k^t} - \nabla f_k(\boldsymbol{\theta}^{(t-\nu_k^t,0)})\right) + \left(\nabla f_k(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) - \nabla f(\boldsymbol{\theta}^{(t-\nu_k^t,0)})\right) \\
&\quad + \nabla f(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) \Big\|^2 \\
&\leq 3\sum_{k=1}^{N} \mathbb{E}\left\| \mathbf{h}_k^{t-\nu_k^t} - \nabla f_k(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) \right\|^2 + 3\sum_{k=1}^{N} \mathbb{E}\left\| \nabla f_k(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) - \nabla f(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) \right\|^2 \\
&\quad + 3\sum_{k=1}^{N} \mathbb{E}\left\| \nabla f(\boldsymbol{\theta}^{(t-\nu_k^t,0)}) \right\|^2 \\
&\leq 3\sum_{k=1}^{N} \left(3L^2\eta_c^2\sigma_{total}^2(\tau-1)\tau\right) + 3\sum_{k=1}^{N} \left(\sigma_g^2\right) + 3\sum_{k=1}^{N} \left(G\right) \\
&= 9NL^2\eta_c^2\sigma_{total}^2(\tau-1)\tau + 3N\sigma_g^2 + 3NG
\end{aligned} \tag{45}
$$

□

Lemma A.16. *Let the global objective gradient be defined as $\nabla f(\boldsymbol{\theta}^{(t,0)}) = \sum_{k=1}^{N} \beta_k \nabla f_k(\boldsymbol{\theta}^{(t,0)})$. Under the assumption that the expected squared gradient norm is bounded ($\mathbb{E}\|\nabla f_k(\boldsymbol{\theta})\|^2 \leq G$), the iterates generated by the FedUCA mechanism satisfy the following bound on the directional drift:*

$$
\mathbb{E}\left\| \sum_{k=1}^{N} I_k^t \alpha_k^t \left(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^t\right) \right\|^2 \leq 3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1) \tag{46}
$$

Proof. Bounding (Local SGD Drift): Let M be the left hand side. Because the broadcasted weights reside on the probability simplex ($\alpha_k^t \geq 0$, $\sum_{k=1}^N \alpha_k^t = 1$) and $I_k^t \in \{0, 1\}$, it holds that $\sum_{k=1}^N I_k^t \alpha_k^t \leq 1$.

Exploiting the convexity of the squared L_2-norm, we apply Jensen's Inequality to push the norm inside the summation, completely avoiding the dimensional penalty of Cauchy-Schwarz:

$$M = \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \big(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^t\big)\right\|^2\right] \leq \mathbb{E}\left[\sum_{k=1}^N I_k^t \alpha_k^t \left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^t\right\|^2\right]$$

We evaluate the expectation by conditioning on the history filtration H_t. Because I_k^t is H_t-measurable, and the local SGD drift variance is conditionally independent of the specific realization given H_t:

$$M \leq \mathbb{E}_{H_t}\left[\sum_{k=1}^N I_k^t \alpha_k^t \mathbb{E}\left[\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^t\right\|^2 \;\middle|\; H_t\right]\right]$$

Substituting the standard bounded local SGD drift $3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)$:

$$M \leq \Big(3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)\Big)\mathbb{E}\left[\sum_{k=1}^N I_k^t \alpha_k^t\right] \leq 3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1) \tag{47}$$

The final inequality strictly holds because the sum of the filtered simplex weights is bounded by 1 ($\sum_{k=1}^N I_k^t \alpha_k^t \leq \sum_{k=1}^N \alpha_k^t = 1$). This concludes the proof. □

Lemma A.17. *Let the global objective gradient be defined as $\nabla f(\boldsymbol{\theta}^{(t,0)}) = \sum_{k=1}^N \beta_k \nabla f_k(\boldsymbol{\theta}^{(t,0)})$. Under the assumption that the expected squared gradient norm is bounded ($\mathbb{E}\|\nabla f_k(\boldsymbol{\theta})\|^2 \leq G$) the drift introduced by the stale trajectory satisfies:*

$$\begin{aligned}\mathbb{E}\Big\|\sum_{k=1}^N (1 - I_k^t)\alpha_k^t \left(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right)\Big\|^2 &\leq (N-S)\left(12L^2\eta^2\tau^2\nu_{max}^2 N\sigma_{total}^2\right) \\ &\quad + (N-S)\left(3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)\right)\end{aligned}$$

Proof. Let D be the left hand side, representing the Stale Trajectory Drift. We apply Cauchy-Schwarz strictly over the set of dropped clients $S_t^c = \{k : I_k^t = 0\}$. The physical size of this dropped set is $|S_t^c| = N - |S_t|$. Because the Optimization mechanism guarantees a minimum active set of $S \leq |S_t|$, the dropped set size is deterministically bounded from above by $N - S$:

$$\begin{aligned}D &= \mathbb{E}\left[\Big\|\sum_{k\in S_t^c} \alpha_k^t \left(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right)\Big\|^2\right] \\ &\leq \mathbb{E}\left[\underbrace{|S_t^c|}_{\leq N-S}\sum_{k=1}^N (1-I_k^t)(\alpha_k^t)^2\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2\right] \\ &\leq (N-S)\mathbb{E}_{H_t}\left[\sum_{k=1}^N \left[(\alpha_k^t)^2\right]\cdot \underbrace{\left[1-I_k^t\right]}_{=1-I_k^t\leq 1}\right. \\ &\qquad \left.\times\, \mathbb{E}\left[\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2 \;\middle|\; H_t\right]\right] \\ &\leq (N-S)\sum_{k=1}^N \mathbb{E}\left[(\alpha_k^t)^2\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2\right]\end{aligned} \tag{48}$$

Because the elements of the simplex strictly satisfy $\sum_{k=1}^{N}(\alpha_k^t)^2 \leq (\sum_{k=1}^{N}\alpha_k^t) = 1$, we cleanly decouple the maximum element:

$$\begin{aligned} D &\leq (N-S)\mathbb{E}\left[\max_k \left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2 \sum_{k=1}^{N}(\alpha_k^t)^2\right] \\ &\leq (N-S)\max_k \mathbb{E}\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2 \end{aligned}$$

To bound the inner norm, we expand $\mathbf{h}_k^{t-\nu_k^t}$ and split it into the global server delay drift (C_1) and local SGD drift (C_2):

$$\mathbb{E}\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2 \leq \frac{2L^2}{\tau}\sum_{r=0}^{\tau-1}\Bigg(\underbrace{\mathbb{E}\left\|\boldsymbol{\theta}^{(t,0)} - \boldsymbol{\theta}^{(t-\nu_k^t,0)}\right\|^2}_{C_1} + \underbrace{\mathbb{E}\left\|\boldsymbol{\theta}^{(t-\nu_k^t,0)} - \boldsymbol{\theta}_k^{t-\nu_k^t,r}\right\|^2}_{C_2}\Bigg) \tag{49}$$

Bounding C_1 (Server Delay Drift): We unroll the global trajectory over ν_k^t rounds, splitting the update into the active set S_m and the stale set S_m^c. Applying Cauchy-Schwarz to each set yields their respective true sizes $|S_m|$ and $(N-|S_m|)$, which seamlessly collapse to exactly N:

$$\begin{aligned} C_1 &= \mathbb{E}\left\|\sum_{m=t-\nu_k^t}^{t-1}\eta\tau\Big(\sum_{j\in S_m}\alpha_j^m\mathbf{h}_j^m + \sum_{j\in S_m^c}\alpha_j^m\mathbf{h}_j^{m-\delta_j}\Big)\right\|^2 \\ &\leq 2\eta^2\tau^2\nu_k^t\sum_{m=t-\nu_k^t}^{t-1}\Bigg(|S_m|\sum_{j\in S_m}\mathbb{E}[(\alpha_j^m)^2](3\sigma_{total}^2) \\ &\quad + (N-|S_m|)\sum_{j\in S_m^c}\mathbb{E}[(\alpha_j^m)^2](3\sigma_{total}^2)\Bigg) \\ &\leq 6\eta^2\tau^2\nu_k^t\sum_{m=t-\nu_k^t}^{t-1}\Bigg(N\sigma_{total}^2\sum_{j=1}^{N}\mathbb{E}[(\alpha_j^m)^2]\Bigg) \\ &\leq 6\eta^2\tau^2\nu_{max}^2 N\sigma_{total}^2 \end{aligned} \tag{50}$$

Bounding C_2 (Local SGD Drift): Using standard local update bounds:

$$C_2 \leq 3\eta_c^2 r^2\sigma_{total}^2 \tag{51}$$

Substituting C_1 and C_2:

$$\begin{aligned} \mathbb{E}\left\|\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\right\|^2 &\leq \frac{2L^2}{\tau}\sum_{r=0}^{\tau-1}\Big(6\eta^2\tau^2\nu_{max}^2 N\sigma_{total}^2 + 3\eta_c^2 r^2\sigma_{total}^2\Big) \\ &\leq 12L^2\eta^2\tau^2\nu_{max}^2 N\sigma_{total}^2 + 3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1) \end{aligned} \tag{52}$$

Substituting (52) back into D concludes the proof. □

Lemma A.18.

$$\mathbb{E}\left\|\mathbf{e}_{stale}^t\right\|^2 \leq (N-S)N\frac{\sigma^2}{\tau} \tag{53}$$

Proof. Let $S_t^c = \{k : I_k^t = 0\}$ denote the set of dropped clients in round t. The size of this dropped cohort is $|S_t^c| = N - |S_t|$. Because the Optimization mechanism guarantees a minimum active cohort of $|S_t| \geq S$, the dropped set size is strictly bounded from above by $|S_t^c| \leq N - S$.

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{e}_{stale}^{t}\right\|^2 &= \mathbb{E}\left[\left\|\sum_{k\in S_t^c}\alpha_k^t(\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t})\right\|^2\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\underbrace{|S_t^c|}_{\leq N-S}\sum_{k\in S_t^c}(\alpha_k^t)^2\left\|\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(b)}{=} \mathbb{E}_{H_t}\left[|S_t^c|\sum_{k=1}^{N}(1-I_k^t)\mathbb{E}_{\alpha_k^t|H_t}\left[(\alpha_k^t)^2\right]\cdot\mathbb{E}\left[\left\|\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t}\right\|^2\;\middle|\;H_t\right]\right] \\
&\overset{(c)}{\leq} (N-S)\sum_{k=1}^{N}\mathbb{E}\left[(1-I_k^t)(\alpha_k^t)^2\left\|\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(d)}{\leq} (N-S)\sum_{k=1}^{N}\mathbb{E}\left[(1)\cdot(1)\cdot\left\|\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(e)}{\leq} (N-S)\sum_{k=1}^{N}\frac{\sigma^2}{\tau} \\
&= (N-S)N\frac{\sigma^2}{\tau}
\end{aligned}
\tag{54}
$$

(a) Cauchy-Schwarz Inequality over Dropped Set: We restrict the summation strictly to the dropped client set S_t^c and apply Cauchy-Schwarz. This seamlessly extracts the maximum possible dropped cohort size $(N-S)$ as a coefficient without dividing by dropout probabilities.

(b) Filtration and Measurability: Because the dropout indicator I_k^t is evaluated deterministically given the history H_t, it is H_t-measurable. Similarly, the past local SGD noise is already realized within the history filtration, allowing us to factor it conditionally.

(c) Tower Property and Set Re-expansion: We transition back to the unconditional expectation, re-expanding the summation to all N clients by using the measurable indicator $(1-I_k^t)$.

(d) Deterministic Upper Bounds: The indicator is strictly bounded by $(1-I_k^t)\leq 1$. Furthermore, to decouple the current weight α_k^t from the past gradient noise (which influenced the current model $\boldsymbol{\theta}^{(t,0)}$), we apply the deterministic simplex bound $\alpha_k^t\leq 1$, meaning $(\alpha_k^t)^2\leq 1$.

(e) Unconditional Local Variance: With the random variables decoupled, we apply the standard assumption for the variance of the local SGD updates: $\mathbb{E}\|\mathbf{d}_k^{t-\nu_k^t}-\mathbf{h}_k^{t-\nu_k^t}\|^2\leq\sigma^2/\tau$. Summing this constant upper bound over all N clients yields the final result. □

Lemma A.19. *Let* $\bar{\mathbf{v}}^t=\sum_{k=1}^{N}\left(I_k^t\alpha_k^t\mathbf{h}_k^t+(1-I_k^t)\alpha_k^t\mathbf{h}_k^{t-\nu_k^t}\right)$. *The unified drift satisfies:*

$$
\begin{aligned}
\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})-\bar{\mathbf{v}}^t\right\|^2 \leq 3\underbrace{\mathbb{E}\left\|\sum_{k=1}^{N}\left(\beta_k-\alpha_k^t\right)\nabla f_k(\boldsymbol{\theta}^{(t,0)})\right\|^2}_{\textit{Unified Bias}} &+9L^2\eta_c^2\sigma_{total}^2\tau(\tau-1) \\
&+3(N-S)\Big(12L^2\eta^2\tau^2\nu_{max}^2N\sigma_{total}^2+3L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)\Big)
\end{aligned}
\tag{55}
$$

Proof. We expand the total combined update inside the directional drift norm, adding and subtracting the ideal global gradients evaluated at the current server model $\nabla f_k(\boldsymbol{\theta}^{(t,0)})$. Applying the relaxed triangle inequality ($\|a+b+c\|^2\leq 3\|a\|^2+3\|b\|^2+3\|c\|^2$) separates the terms:

$$\mathbb{E}\Big\|\nabla f(\boldsymbol{\theta}^{(t,0)}) - \sum_{k=1}^{N} I_k^t \alpha_k^t \mathbf{h}_k^t - \sum_{k=1}^{N} (1 - I_k^t)\alpha_k^t \mathbf{h}_k^{t-\nu_k^t}\Big\|^2$$
$$\leq \underbrace{3\,\mathbb{E}\Big\|\sum_{k=1}^{N} \beta_k \nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \sum_{k=1}^{N} \Big(I_k^t + (1 - I_k^t)\Big)\alpha_k^t \nabla f_k(\boldsymbol{\theta}^{(t,0)})\Big\|^2}_{U_1 \text{ (Bias)}}$$
$$+ \underbrace{3\,\mathbb{E}\Big\|\sum_{k=1}^{N} I_k^t \alpha_k^t \Big(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^t\Big)\Big\|^2}_{U_2 \text{ (Active Drift)}} + \underbrace{3\,\mathbb{E}\Big\|\sum_{k=1}^{N} (1 - I_k^t)\alpha_k^t \Big(\nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \mathbf{h}_k^{t-\nu_k^t}\Big)\Big\|^2}_{U_3 \text{ (Stale Drift)}} \tag{56}$$

For the Unified Bias (U_1), we observe that the indicator variables for the active and dropped sets share the exact same ideal gradient evaluation. We factor it out conditioning on the history H_t:

$$U_1 = \mathbb{E}_{H_t}\left[\mathbb{E}\left[\Big\|\sum_{k=1}^{N} \beta_k \nabla f_k(\boldsymbol{\theta}^{(t,0)}) - \sum_{k=1}^{N} \Big(I_k^t + (1 - I_k^t)\Big)\alpha_k^t \nabla f_k(\boldsymbol{\theta}^{(t,0)})\Big\|^2 \;\Big|\; H_t\right]\right] \tag{57}$$

Because $I_k^t + (1 - I_k^t) = 1$ strictly, the participation indicators perfectly collapse. The sampling strategy reconstructs the full gradient space, leaving only the structural discrepancy between the objective weights β_k and the server's realization α_k^t:

$$U_1 = \mathbb{E}\Big\|\sum_{k=1}^{N} \left(\beta_k - \alpha_k^t\right) \nabla f_k(\boldsymbol{\theta}^{(t,0)})\Big\|^2 \tag{58}$$

For U_2, we substitute the bounded active local SGD drift from Lemma A.16. Because the coefficient is multiplied by 3, we obtain:

$$U_2 \leq 3\Big(3L^2\eta_c^2\sigma_{total}^2\tau(\tau - 1)\Big) = 9L^2\eta_c^2\sigma_{total}^2\tau(\tau - 1) \tag{59}$$

For U_3, we substitute the bounded stale trajectory drift from Lemma A.17, applying the multiplier:

$$U_3 \leq 3(N - S)\Big(12L^2\eta^2\tau^2\nu_{max}^2 N\sigma_{total}^2 + 3L^2\eta_c^2\sigma_{total}^2\tau(\tau - 1)\Big) \tag{60}$$

Combining the optimized bounds for U_1, U_2, and U_3 completes the proof. □

Proposition A.20. *Let the sequence of global server models be denoted by $\boldsymbol{\theta}^{(t,0)}$ for $t \in \{0, \ldots, T-1\}$. Assume the global objective is bounded below by f^*, with an initial suboptimality gap of $D_0 = f(\boldsymbol{\theta}^{(0,0)}) - f^*$. Let the target objective weights be defined as the vector $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]^T$, and the dynamically assigned active weights at round t be defined as $\boldsymbol{\alpha}^t = [\alpha_1^t, \ldots, \alpha_N^t]^T$. Given the learning rate condition $-\frac{\eta\tau}{4}(1 - 12L\eta\tau) \leq -\eta\tau\frac{S}{8K}$ (for some $K > 0$) and assuming locally bounded gradients ($\|\nabla f_k\|^2 \leq G$), the average expected gradient norm is bounded by five discrete error components:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big\|\nabla f(\boldsymbol{\theta}^{(t,0)})\Big\|^2 \leq \mathcal{E}_{init} + \mathcal{E}_{agg} + \mathcal{E}_{var} + \mathcal{E}_{drift} + \mathcal{E}_{sys} \tag{61}$$

here, $f(\boldsymbol{\theta}) = \sum_{k=1}^{N} \beta_k f_k(\boldsymbol{\theta})$, $\sum_{k=1}^{N} \beta_k = 1$ and the discrete error components are defined as follows: the optimization rate is $\mathcal{E}_{init} = \frac{8KD_0}{\eta\tau ST}$, the structural Optimization bias is $\mathcal{E}_{agg} = \frac{12KNG}{S}\big(\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2\big)$, the stochastic variance is $\mathcal{E}_{var} = \frac{8KL\eta\sigma^2}{S}\big(\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\alpha}^t\|^2\big)$, the temporal drift penalty is $\mathcal{E}_{drift} = \frac{8K}{\eta\tau S}\Omega$, and the irreducible system error floor is $\mathcal{E}_{sys} = \frac{8K}{\eta\tau S}\big(\Psi + \eta(N - S)N\sigma^2\big) + \frac{8KL\eta\sigma^2 N(N-S)}{S}$, with Ω representing the aggregate temporal drift penalties (local updates and server delays) and Ψ representing the structural data heterogeneity.

Proof. The following inequality follows from smoothness assumptions.

$$\mathbb{E}\left[f(\boldsymbol{\theta}^{(t+1,0)})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^{(t,0)})\right] \leq \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \boldsymbol{\theta}^{(t+1,0)} - \boldsymbol{\theta}^{(t,0)} \right\rangle\right] + \frac{L}{2}\mathbb{E}\left[\left\|\boldsymbol{\theta}^{(t+1,0)} - \boldsymbol{\theta}^{(t,0)}\right\|^2\right]$$
$$\leq -\tau\eta \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \mathbf{v}^t \right\rangle\right]}_{T_1} + L\tau^2\eta^2 \underbrace{\mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{d}_k^t\right\|^2\right]}_{T_2 \text{ (Fresh Variance)}} + L\tau^2\eta^2 \underbrace{\mathbb{E}\left[\left\|\sum_{k=1}^N (1 - I_k^t)\alpha_k^t \mathbf{d}_k^{t-\nu_k^t}\right\|^2\right]}_{T_3 \text{ (Stale Variance)}}$$

Where the total update is defined as $\mathbf{v}^t = \mathbf{e}_{fresh}^t + \mathbf{e}_{stale}^t + \bar{\mathbf{v}}^t$ here $\mathbf{e}_{fresh}^t = \sum_{k=1}^N I_k^t \alpha_k^t (\mathbf{d}_k^t - \mathbf{h}_k^t)$ and $\mathbf{e}_{stale}^t = \sum_{k=1}^N (1 - I_k^t)\alpha_k^t (\mathbf{d}_k^{t-\nu_k^t} - \mathbf{h}_k^{t-\nu_k^t})$. $\bar{\mathbf{v}} = \sum_{k=1}^N I_k^t \alpha_k^t \mathbf{h}_k^t + \sum_{k=1}^N (1 - I_k^t)\alpha_k^t \mathbf{h}_k^{t-\nu_k^t}$

The expectation is taken by conditioning over the randomness of history H_t up to time t or the Filtration up to time t.T be specific we repeatedly use $\mathbb{E}[.] = \mathbb{E}_{H_t}\mathbb{E}[.|H_t]$

We now bound $-\eta\tau T_1$

$$T_1 = \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \bar{\mathbf{v}}^t + \mathbf{e}_{fresh}^t + \mathbf{e}_{stale}^t \right\rangle\right] = \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \bar{\mathbf{v}}^t \right\rangle\right] + \underbrace{\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \mathbf{e}_{fresh}^t \right\rangle\right]}_{=\mathbf{0} \text{ (Tower Rule)}}$$
$$+ \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \mathbf{e}_{stale}^t \right\rangle\right]$$
$$= \frac{1}{2}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 + \frac{1}{2}\mathbb{E}\left\|\bar{\mathbf{v}}^t\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)}) - \bar{\mathbf{v}}^t\right\|^2 + \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \mathbf{e}_{stale}^t \right\rangle\right] \tag{62}$$

By Young's inequality we have the following

$$\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{\theta}^{(t,0)}), \mathbf{e}_{stale}^t \right\rangle\right] \geq -\frac{1}{4}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 - \mathbb{E}\left\|\mathbf{e}_{stale}^t\right\|^2$$

$$-\eta\tau T_1 \leq -\frac{\eta\tau}{4}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)}) - \bar{\mathbf{v}}^t\right\|^2$$
$$+ \eta\tau\mathbb{E}\left\|\mathbf{e}_{stale}^t\right\|^2 \tag{63}$$

We bound T_2 as below

$$T_2 = \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{d}_k^t\right\|^2\right] \overset{(a)}{=} \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{h}_k^t\right\|^2\right] + \sum_{k=1}^N \mathbb{E}\left[(I_k^t \alpha_k^t)^2 \left\|\mathbf{d}_k^t - \mathbf{h}_k^t\right\|^2\right]$$
$$\overset{(b)}{\leq} \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{h}_k^t\right\|^2\right] + \frac{\sigma^2}{\tau}\sum_{k=1}^N \mathbb{E}\left[I_k^t (\alpha_k^t)^2\right] \overset{(c)}{=} \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{h}_k^t\right\|^2\right] + \frac{\sigma^2}{\tau}\sum_{k=1}^N \mathbb{E}_{H_t}\left[\mathbb{E}[I_k^t \mid H_t](\alpha_k^t)^2\right]$$
$$\leq \mathbb{E}\left[\left\|\sum_{k=1}^N I_k^t \alpha_k^t \mathbf{h}_k^t\right\|^2\right] + \frac{\sigma^2}{\tau}\sum_{k=1}^N \mathbb{E}\left[(\alpha_k^t)^2\right] \tag{64}$$

(a) Exact Variance Decomposition: The fresh local sampling noise $\mathbf{e}_k^t = \mathbf{d}_k^t - \mathbf{h}_k^t$ is conditionally zero-mean given H_t and independent across clients, making all cross-terms evaluate strictly to zero. We can therefore separate the expected trajectory from the sampling variance.
(b) Unconditional Noise Bound: We substitute the standard local variance bound $\mathbb{E}[\|\mathbf{d}_k^t - \mathbf{h}_k^t\|^2] \leq \sigma^2/\tau$. Because $(I_k^t)^2$ is a binary indicator, it simplifies exactly to I_k^t.
(c) Tower Rule Filtration: We apply the Tower Rule ($\mathbb{E}[\cdot] = \mathbb{E}_{H_t}[\mathbb{E}[\cdot \mid H_t]]$) specifically to the noise sum. The algorithm's weight generation for the current round (α_k^t) is entirely H_t-measurable, so it factors outside the inner conditional expectation.

We now bound T_3 as below:

$$
\begin{aligned}
T_3 &= \mathbb{E}\left[\left\|\sum_{k=1}^{N}(1-I_k^t)\alpha_k^t \mathbf{d}_k^{t-\nu_k^t}\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{k\in S_t^c}\alpha_k^t \mathbf{d}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\underbrace{|S_t^c|}_{\leq N-S}\sum_{k\in S_t^c}(\alpha_k^t)^2\left\|\mathbf{d}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(b)}{=} \mathbb{E}_{H_t}\left[|S_t^c|\sum_{k=1}^{N}(1-I_k^t)\mathbb{E}_{\alpha_k^t|H_t}\left[(\alpha_k^t)^2\right]\mathbb{E}\left[\left\|\mathbf{d}_k^{t-\nu_k^t}\right\|^2 \;\middle|\; H_t\right]\right] \\
&\overset{(c)}{\leq} (N-S)\sum_{k=1}^{N}\mathbb{E}\left[(1-I_k^t)(\alpha_k^t)^2\left\|\mathbf{d}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(d)}{\leq} (N-S)\sum_{k=1}^{N}\mathbb{E}\left[(1)\cdot(1)\cdot\left\|\mathbf{d}_k^{t-\nu_k^t}\right\|^2\right] \\
&\overset{(e)}{\leq} (N-S)\sum_{k=1}^{N}\left(\frac{\sigma^2}{\tau}+\mathbb{E}\left\|\mathbf{h}_k^{t-\nu_k^t}\right\|^2\right) \\
&= (N-S)N\frac{\sigma^2}{\tau}+(N-S)\sum_{k=1}^{N}\mathbb{E}\left\|\mathbf{h}_k^{t-\nu_k^t}\right\|^2
\end{aligned} \tag{65}
$$

(a) Cauchy-Schwarz Inequality over Dropped Set: We restrict the summation strictly to the dropped client set $S_t^c = \{k : I_k^t = 0\}$ and apply Cauchy-Schwarz. Because the Optimization mechanism guarantees a minimum active cohort of S, the dropped set size is deterministically bounded from above by $|S_t^c| \leq N - S$. This cleanly extracts the coefficient without requiring division by dropout probabilities.

(b) Filtration and Measurability: We condition on the history H_t. The dropout indicator I_k^t and the active set size $|S_t^c|$ are evaluated deterministically given H_t, making them H_t-measurable. Furthermore, the past stale update $\mathbf{d}_k^{t-\nu_k^t}$ is fully realized within the history filtration. This allows us to cleanly decouple the expectation of the current weights α_k^t.

(c) Tower Property and Set Re-expansion: We transition back to the unconditional expectation, re-expanding the summation to all N clients by multiplying by the indicator $(1 - I_k^t)$, and applying the structural upper bound $(N - S)$.

(d) Deterministic Upper Bounds: The indicator is strictly bounded by $(1 - I_k^t) \leq 1$. To safely eliminate the weights, we apply the deterministic simplex bound $\alpha_k^t \leq 1$, which strictly implies $(\alpha_k^t)^2 \leq 1$.

(e) Local SGD Second Moment: We apply the standard variance decomposition for local SGD updates ($\mathbb{E}\|\mathbf{x}\|^2 = \text{Var}(\mathbf{x}) + \|\mathbb{E}[\mathbf{x}]\|^2$). The variance of the local update is bounded by σ^2/τ, and its expected drift is exactly $\mathbf{h}_k^{t-\nu_k^t}$. Summing these components yields the final bound.

By substituting T_1, T_2, T_3 we get the following

$$
\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{\theta}^{(t+1,0)})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^{(t,0)})\right] \leq &-\frac{\eta\tau}{4}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 + \frac{\eta\tau}{2}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)}) - \bar{\mathbf{v}}^t\right\|^2 + \eta\tau\mathbb{E}\left\|\mathbf{e}_{stale}^t\right\|^2 \\
&+ L\tau^2\eta^2\mathbb{E}\left[\left\|\sum_{k=1}^{N} I_k^t\alpha_k^t\mathbf{h}_k^t\right\|^2\right] + L\tau^2\eta^2(N-S)\sum_{k=1}^{N}\mathbb{E}\left\|\mathbf{h}_k^{t-\nu_k^t}\right\|^2 \\
&+ L\tau\eta^2\sigma^2\left(\sum_{k=1}^{N}\mathbb{E}\left[(\alpha_k^t)^2\right] + N(N-S)\right)
\end{aligned} \tag{66}
$$

Substituting the bounded components from Lemma A.18, Lemma A.19, Lemma A.14, and Lemma A.15 into the master descent inequality (66), we collect and group the terms by their functional source.

Extracting the global gradient norms yields the primary descent condition. The remaining terms strictly bound the variance injected by the Optimization mechanism, local client drift, and stale updates:

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{\theta}^{(t+1,0)})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^{(t,0)})\right] \leq & \underbrace{-\eta\tau\left(\frac{1}{4} - 3L\eta\tau\right)\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2}_{\text{Descent Condition}} \\ & + \underbrace{\frac{3\eta\tau}{2}\mathbb{E}\left\|\sum_{k=1}^{N}\left(\beta_k - \alpha_k^t\right)\nabla f_k(\boldsymbol{\theta}^{(t,0)})\right\|^2}_{\text{Optimization Structural Bias}} \\ & + \underbrace{\frac{9\eta\tau}{2}L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)\Big(1 + (N-S)\Big)}_{\text{Client Drift (Unified)}} \\ & + \underbrace{9L^3\tau^3\eta^2\eta_c^2\sigma_{total}^2(\tau-1)\Big(1 + N(N-S)\Big)}_{\text{Client Drift (Aggregate Norms)}} \\ & + \underbrace{18L^2\eta^3\tau^3\nu_{max}^2 N(N-S)\sigma_{total}^2}_{\text{Server Delay Penalty}} \\ & + \underbrace{3L\tau^2\eta^2\Big(\sigma_g^2\big[1 + N(N-S)\big] + N(N-S)G\Big)}_{\text{Heterogeneity Bound}} \\ & + \underbrace{\eta(N-S)N\sigma^2 + L\tau\eta^2\sigma^2\left(\sum_{k=1}^{N}\mathbb{E}\big[(\alpha_k^t)^2\big] + N(N-S)\right)}_{\text{Stochastic Noise Floor}} \end{aligned} \tag{67}$$

By using the learning rate condition $-\frac{\eta\tau}{4}(1 - 12L\eta\tau) \leq -\eta\tau\frac{S}{8K}$ to bound the descent step, and grouping the variance components into stochastic noise (Φ_{noise}^t), system heterogeneity (Φ_{het}), and stale delay penalties (Φ_{stale}), we obtain the following single-round descent bound:

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{\theta}^{(t+1,0)})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^{(t,0)})\right] \leq & -\eta\tau\frac{S}{8K}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 + \frac{3\eta\tau NG}{2}\sum_{k=1}^{N}\mathbb{E}\left[(\beta_k - \alpha_k^t)^2\right] \\ & + \Phi^t + \Psi + \Omega \end{aligned} \tag{68}$$

where the system variance components, stochastic noise (Φ^t), system heterogeneity (Ψ), and stale delay drift (Ω), are defined as:

$$\begin{aligned}\Phi^t &= \eta(N-S)N\sigma^2 + L\tau\eta^2\sigma^2\left(\sum_{k=1}^{N}\mathbb{E}\big[(\alpha_k^t)^2\big] + N(N-S)\right) \\ \Psi &= 3L\tau^2\eta^2\Big(\sigma_g^2\big[1 + N(N-S)\big] + N(N-S)G\Big) \\ \Omega &= \frac{9\eta\tau}{2}L^2\eta_c^2\sigma_{total}^2\tau(\tau-1)\Big(1 + (N-S)\Big) \\ &\quad + 9L^3\tau^3\eta^2\eta_c^2\sigma_{total}^2(\tau-1)\Big(1 + N(N-S)\Big) \\ &\quad + 18L^2\eta^3\tau^3\nu_{max}^2 N(N-S)\sigma_{total}^2 \end{aligned} \tag{69}$$

Assuming the global objective is bounded below by f^*, we define the initial suboptimality gap as $D_0 = f(\boldsymbol{\theta}^{(0,0)}) - f^*$. Telescoping Eq. 68 over T rounds and dividing by $T \cdot \eta\tau\frac{S}{8K}$, we obtain the final convergence bound:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 \leq \frac{8KD_0}{\eta\tau ST} + \frac{12KNG}{ST}\sum_{t=0}^{T-1}\sum_{k=1}^{N}\mathbb{E}\left[(\beta_k - \alpha_k^t)^2\right] + \frac{8K}{\eta\tau ST}\sum_{t=0}^{T-1}\left(\Phi^t + \Psi + \Omega\right) \tag{70}$$

By defining the probability weight vectors as $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]^T$ and $\boldsymbol{\alpha}^t = [\alpha_1^t, \ldots, \alpha_N^t]^T$, we can completely decompose the convergence bound of FedUCA into five interpretable driving forces:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\boldsymbol{\theta}^{(t,0)})\right\|^2 \leq \mathcal{E}_{\text{init}} + \mathcal{E}_{\text{agg}} + \mathcal{E}_{\text{var}} + \mathcal{E}_{\text{drift}} + \mathcal{E}_{\text{sys}} \tag{71}$$

where the discrete error components are defined as follows: the optimization rate is $\mathcal{E}_{init} = \frac{8KD_0}{\eta\tau ST}$, the structural Optimization bias is $\mathcal{E}_{agg} = \frac{12KNG}{S}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\alpha}^t\|^2\right)$, the stochastic variance is $\mathcal{E}_{var} = \frac{8KL\eta\sigma^2}{S}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\alpha}^t\|^2\right)$, the temporal drift penalty is $\mathcal{E}_{drift} = \frac{8K}{\eta\tau S}\Omega$, and the irreducible system error floor is $\mathcal{E}_{sys} = \frac{8K}{\eta\tau S}\left(\Psi + \eta(N-S)N\sigma^2\right) + \frac{8KL\eta\sigma^2 N(N-S)}{S}$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims made in the abstract and introduction are supported by experiments and theoretical analysis.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have clearly discussed the limitations in the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the assumptions required for the proofs are clearly stated, and the known lemmas are clearly referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the experiment details are clearly documented for reproducibility of the results presented in the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the experimental details are clearly documented for reproducibility of the results presented in the main paper. The code is released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training/hyperparameter details have been included to understand the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included the one sigma error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided this information in the supplementary.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: We have read and followed the mentioned ethics guidelines.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: Our work addresses FL with rational clients. It does not have any negative societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the resources used to develop the work are mentioned and acknowledged.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.