DISTRIBUTED LEARNING UNDER AUTONOMOUS CLIENT PARTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The common assumption in distributed learning (DL) is that the central parameter server dictates client participation to optimize its training efficiency, while clients do not actively affect participation decisions beyond sending their updates when requested. However, this does not reflect the clients' real-world behavior, where they act autonomously based on self-interest, privacy concerns, and local model performance. In this paper, we introduce a theoretical framework that models autonomous state-dependent client participation in DL as a Markov chain, and analyze learning convergence in case client availability is dynamic and correlated across rounds. We analyze both homogeneous and heterogeneous data environments, as well as various participation rules that model client behavior. We demonstrate, both theoretically and empirically, that once autonomy exceeds a critical threshold, the Markov chain fails to mix and the learning process stagnates. We further show that, in heterogeneous settings, autonomous dropout of clients from the learning process induces a bias, which creates a feedback loop that steers the global model to an equilibrium that favors the dominant data distribution. Finally, we consider the setting in which the parameter server can incentivize the participation of a uniform set of clients using a limited budget. We develop a policy-optimization method for the allocation of this budget, and show it ensures convergence even in highly autonomous distributed networks.

1 INTRODUCTION

Federated learning (FL), first proposed by Google researchers (Konečný et al., 2016), (McMahan et al., 2017), is a distributed machine learning framework in which multiple clients collaborate to train a shared global model by periodically sending their model updates to a central server. The server then aggregates these updates and sends the global model back to the clients. This approach enables collaborative learning across a centralized server, while maintaining other desirable properties such as data privacy, as the clients' data is kept localized (Li et al., 2021).

As distributed systems evolve from a single server in isolation to *multiple* servers, the interaction between the model providers and data owners becomes increasingly competitive. This raises the theoretical problem of *learning under competition*, which has been studied in different contexts (Ginart et al., 2021). For example, Mansour et al. (2018) considered competition between two multi-armed bandit algorithms faced with the same bandit instance. Dean et al. (2024) studied the dynamics of populations apportioning themselves among services, and services that choose predictors based on their observed user population. They showed that when clients are trying to minimize their risk, servers tend to specialize in their subpopulations of client data. Ben-Porat & Tennenholtz (2017) considered a regression task faced by two players, where the payoff of each player is the proportion of points that she predicts more accurately than the other player. They proved that finding a best response linear predictor can be computed in polynomial time in the number of examples, when the features domain is fixed. Jagadeesan et al. (2023) defined a model of competition for classification tasks. They showed that such competition can fundamentally alter the behavior of the known scaling trends of a single model-provider in isolation. In the case of multi-model providers, competition can cause overall predictive accuracy across users to be non-monotonic or even to decrease with scale.

Despite this shift toward analyzing multiple servers competing for clients, the FL literature on client selection still mostly assumes the central server controls which clients participate in each training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

round, and can select them to improve its performance, e.g., higher model accuracy and faster convergence rates (Nishio & Yonetani, 2019; Fu et al., 2023; Cho et al., 2020). For example, Zhao & Zhang (2015) introduced an importance sampling (IS) approach, in which the server uses IS to select clients with data samples that are far from the global model. It was shown that this improves the convergence rate, by reducing stochastic variance when used in prox-SGD (stochastic gradient descent) and prox-SDCA (stochastic dual coordinate ascent). Subsequently, Wang et al. (2023) introduced DELTA, which converges faster than IS, by characterizing client diversity and local variance, and by sampling representative clients with valuable information for global model updates.

Nevertheless, in real-world distributed networks, clients are autonomous. Their participation is affected not only by the properties of their devices such as battery status and network connectivity, but also by their self-interest in the model. Moreover, nowadays, clients have multiple choices of service providers, for example, different large language models (LLMs) (Chang et al., 2024), or various social media platforms. Their decision whether to participate in a learning process can thus be changed based on factors such as model performance, privacy policies and computational efficiency. Moreover, it can also change dynamically across different rounds of the learning process, and might depend on the decisions made in previous rounds. In this spirit, Wang & Ji (2022) considered arbitrary client participation and showed convergence bounds for different participation patterns. Then, Sun et al. (2025) studied such FL model with correlated participation across rounds, where every client wait a minimum of R rounds before he can re-participate. Other works assume that at least one client participate in each round, and that the clients' participation rule has no correlation with the learning process. We refer the reader to Appendix A for a detailed discussion of related work.

In this paper, we address this gap from realistic client participation scenarios, and formulate a distributed learning setting with autonomous, round-dependent, client participation. We go beyond passive client compliance, and model the clients' participation as a Markov chain driven by the clients' participation rules, and a behavior parameter. The participation probability then depends on the global model and the client's knowledge history. Our main *contributions* are as follows:

- In Section 2, we introduce a general framework for distributed learning with autonomous clients, modeled by a participation rule that can be correlated with past decisions and data.
- In Section 3, we consider homogeneous data distributions, and analyze learning convergence by modeling the state-dependent participation decisions as a Markov chain. We derive the convergence error bounds and establish a *critical autonomy threshold*.
- In Section 4, we take the server's perspective, and utilize the derived bound to derive a budget-constrained server policy optimization problem. We propose an algorithm that guarantees the learning stability for the server even in highly autonomous client environments.
- In Section 5, we extend our framework to heterogeneous data distributions. We demonstrate that autonomous behavior induces statistical bias, creating a feedback loop that steers the global model toward the most dominant data distributions, dependent on the model's initialization point.

In Appendix E, we complement the above, and provide extensive numerical simulations that validate our theoretical convergence bounds, empirically confirm the critical autonomy threshold, and demonstrate our proposed server budget optimization.

2 Problem Formulation

We consider a distributed learning setting comprising a single central server and N_c clients. Each client $i \in [N_c] := \{0, 1, \ldots, N_c - 1\}$ has a dataset drawn from a distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^n$ is the features domain and $\mathcal{Y} \subset \mathbb{R}$ is the label domain. The server's goal is to minimize a global objective function, which is an average over client's objective functions, defined by

$$F(\theta) = \mathbb{E}_{i\sim\mathcal{P}}[F_i(\theta)] = \mathbb{E}_{i\sim\mathcal{P}}\mathbb{E}_{\xi\sim\mathcal{D}_i}[f_i(\theta, \xi)], \tag{1}$$

where $f_i(\theta, \xi)$ denotes the loss function evaluated on a data sample ξ drawn from distribution $\mathcal{D}_i$, and $\mathcal{P}$ is a uniform distribution over the clients.

Clients generally expect to utilize the data of the other clients in order to improve their own predictions. However, acting in self-interest, clients may not always be willing to participate in the learning process, for various reasons, such as privacy policies, energy consumption, limited communication or, to explore other competitive servers. To model this behavior, we define a participation rule.

At each round $r+1$, clients autonomously decide whether to participate based on their individual participation rule $\phi_c(k^r, h_i^r, b_i) \in \{0,1\}$, where 1 denotes a choice to participate in the round and 0 otherwise. This rule is dictated by three key parameters. First, k^r represents the common public knowledge of the server and all clients. It includes the server model from all previous rounds $(\theta^{(0)}, \theta^{(1)} \dots \theta^{(r-1)})$, it can include also global system metadata, such as the number of participating clients in previous rounds. Second, h_i^r denotes the personal history of the client i in round r. It represents the participation records of the client and all local data generated in past rounds from the client's local dataset and the common public knowledge. For example, using SGD, the client history can include all local gradients calculated based on their dataset and the server models $\{g_i^{(r)}\}_{k\in(0,1\dots r-1)}$. For LLMs, it can be the updated server models after editing based on the client's private dataset (Sinitsin et al., 2020). Finally, $b_i \in \mathbb{R}$ is a parameter that models the autonomous level of client i. For example, b_i can represent the client's incentive to explore different servers, or their reluctance to participate when the server's model diverges from their local data distribution. While this could be represented by a vector of parameters, we define it as a scalar, for simplicity.

Let $\mathcal{F}^{(r)}$ denote the entire history of the system up to round r, including the public knowledge k^r and all personal histories $\{h_i^r\}_{i=1}^{N_c}$, and the autonomous levels $\{b_i\}_{i=1}^{N_c}$. Given $\mathcal{F}^{(r)}$, we assume that this binary decision is probabilistic and governed by a state-dependent probability distribution $p^{(r+1)} = (p_1^{(r+1)}, \dots, p_{N_c}^{(r+1)})$, which serves as the concrete realization of the clients' autonomous parameters b_i and the current global state $\mathcal{F}^{(r)}$. This yields an active subset of participating clients

$$\mathcal{S}^{(r+1)} := \{i \in [N_c] : \phi_c(k^r, h_i^r, b_i) = 1\}. \tag{2}$$

We denote the server's aggregation function of the clients' gradients by $\mathrm{Agg}(\cdot)$. Assuming, for concreteness, that the clients run a single local iteration of SGD, the local update computed by client i is defined as $\Delta_i^{(r+1)} = -\eta\, g_i(\theta^{(r)}, \xi_i^{(r)})$, where η is the learning rate, assumed for simplicity to be equal across all clients. We may further assume that $\mathrm{Agg}(\cdot)$ is a standard uniform average of the local updates from the active set. The server's global update step is therefore defined as:

$$\theta^{(r+1)} = \theta^{(r)} + \mathrm{Agg}(\{\Delta_i^{(r+1)}\}_{i\in\mathcal{S}^{(r+1)}}) = \theta^{(r)} + \frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{i\in\mathcal{S}^{(r+1)}} \Delta_i^{(r+1)} = \theta^{(r)} - \eta\,\widehat{g}^{(r)}, \tag{3}$$

where $\widehat{g}^{(r)} := \frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{i\in\mathcal{S}^{(r+1)}} g_i(\theta^{(r)}, \xi_i^{(r)})$ is the aggregated stochastic gradient.

3 Homogeneous Data Analysis

In this section, we analyze the distributed learning problem formulated in Section 2 assuming homogeneous data. We assume all N_c clients sample data independently and identically distributed (i.i.d.) from the same global distribution, such that $\mathcal{D}_i = \mathcal{D}$ for all $i \in [N_c]$. The global objective function therefore simplifies to $F(\theta) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\theta, \xi)]$. Analyzing this homogeneous setting before the more challenging heterogeneous data setting, allows us to isolate the effect of clients' participation choices on the global model. We show that the effective number of rounds is reduced due to rounds without any client participation, and the variance is increased due to partial participation.

3.1 Assumptions and Aggregated Gradient Properties

The following assumptions that we make are common, and we next state them here for completeness:

Assumption 1 (Convexity and Smoothness)**.** *$F(\theta)$ is convex and L-smooth.*

Assumption 2 (Unbiased Stochastic Gradient)**.** *The stochastic gradients computed by the clients are unbiased estimators of the true gradient:* $\mathbb{E}_{\xi_i^{(r)}\sim\mathcal{D}}\left[g_i(\theta^{(r)}, \xi_i^{(r)})\right] = \nabla F_i(\theta^{(r)}) = \nabla F(\theta^{(r)})$.

Assumption 3 (Bounded Variance)**.** *The variance of the stochastic gradients is uniformly bounded by a constant $\sigma^2 > 0$:* $\mathbb{E}_{\xi_i^{(r)}\sim\mathcal{D}}\left[\left\|g_i(\theta^{(r)}, \xi_i^{(r)}) - \nabla F(\theta^{(r)})\right\|^2\right] \le \sigma^2$.

Based on these assumptions, we first establish simple statistical properties of the aggregated gradient estimator $\widehat{g}^{(r)}$, conditioned on at least one client participating (i.e., $|\mathcal{S}^{(r+1)}| > 0$).

Lemma 1 (Properties of the Aggregated Gradient)**.** *Given that the active set $\mathcal{S}^{(r+1)}$ is non-empty, the aggregated gradient estimator $\widehat{g}^{(r)}$ is unbiased, satisfying: $\mathbb{E}[\widehat{g}^{(r)} \mid \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0] = \nabla F(\theta^{(r)})$, and its variance is bounded by the expected inverse size of the active set:*

$$\mathbb{E}\left[\|\widehat{g}^{(r)} - \nabla F(\theta^{(r)})\|^2 \,\middle|\, \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0\right] \leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|} \,\middle|\, |\mathcal{S}^{(r+1)}| > 0\right] \sigma^2. \tag{4}$$

The simple proof is in Appendix B.1. Recent work on stochastic optimization under heterogeneous and random worker compute times (Maranjyan et al., 2025) establishes analogous unbiasedness and variance bounds. They investigate the problem of mini-batch SGD, which in the homogeneous case is similar to our DL model. However, their model does not account for the possibility of an empty learning round. Instead, their gradient estimator scales the global update by a fixed expected batch size, $B = \sum p_i B_i$, where B_i is the number of trials the server assigns to worker i and p_i is the probability of success dictated by the server's chosen time limit. In this way, they can assume that the server can guarantee that B is strictly larger than zero. In our setting, autonomous client participation means we cannot guarantee a non-empty round, our estimator necessitates conditioning on $|\mathcal{S}^{(r+1)}| > 0$, which consequently affects the variance bound and convergence dynamics.

3.2 Convergence Analysis

Following our assumption that client participation is probabilistic, it is possible for a training round to result in no global model update if no clients choose to participate. Let $\mathbb{I}^{(r+1)} = \mathbf{1}_{|\mathcal{S}^{(r+1)}|>0}$ be the indicator function for a valid update, and let $\mathcal{F}^{(r)}$ be the history up to round r. The effective update rule is therefore $\theta^{(r+1)} = \theta^{(r)} - \eta \mathbb{I}^{(r+1)} \widehat{g}^{(r)}$.

Since we can encounter rounds with no update, we must take into account the overall participation probability, ρ, which is applicable for all underlying participation rules and directly affects the effective convergence time. Therefore, we introduce two important parameters that dictate the convergence dynamics. Let $\rho^{(r+1)} := \mathbb{P}(|\mathcal{S}^{(r+1)}| > 0 \mid \mathcal{F}^{(r)})$ be the *participation probability*, the probability of a successful round given $\mathcal{F}^{(r)}$, and let $m_{\text{eff}}^{(r+1)} := \mathbb{E}^{-1}[\frac{1}{|\mathcal{S}^{(r+1)}|} \,|\, |\mathcal{S}^{(r+1)}| > 0]$ be the *effective sample size*, the expected size of the active set given that an update actually occurs.

When evaluating expectations over sets that may be empty, as in the term $\mathbb{E}\left[\frac{\mathbb{I}^{(r+1)}}{|\mathcal{S}^{(r+1)}|}\right]$, we adopt the convention that $0/0 = 0$. We now bound the expected distance to the optimal solution $\theta^\star$.

Theorem 1 (General One-Step Descent Bound)**.** *Under Assumptions1,2 and 3, for any learning rate $\eta > 0$, the expected distance to the optimal model $\theta^\star$ satisfies:*

$$\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 \leq \mathbb{E}\|\theta^{(r)} - \theta^\star\|^2 - 2\eta(1-\eta L)\mathbb{E}[\rho^{(r+1)} \cdot (F(\theta^{(r)}) - F(\theta^\star))] + \eta^2\sigma^2\mathbb{E}\left[\frac{\mathbb{I}^{(r+1)}}{|\mathcal{S}^{(r+1)}|}\right].$$

This theorem serves as a meta-bound for our analysis, and subsequent concrete convergence bounds in this paper will be derived from this theorem. The proof of Theorem 1 appears in Appendix B.2.

3.3 Markovian Dynamics Model

In this section, we focus on a dynamic participation regime, in which the selection probability $p_i^{(r+1)}$ depends directly on the active subset $\mathcal{S}^{(r)}$ of the previous round. This temporal dependency induces a Markov chain over the state space of all possible client subsets. Let $\Omega = 2^{[N_c]}$ denote the state space encompassing all possible subsets of clients. We define the state of the system at round r as the random variable $\mathcal{S}^{(r)} \in \Omega$. We further assume here that the clients make their participation decisions independently in a given round based on the shared probabilities $p(S) = \{p_1(S), \ldots, p_{N_c}(S)\}$, where the transition probability from state S to state S' is:

$$P(\mathcal{S}^{(r+1)} = S' \mid \mathcal{S}^{(r)} = S) = \prod_{i \in S'} p_i(S) \cdot \prod_{j \notin S'} (1 - p_j(S)). \tag{5}$$

Chain Properties Provided that the participation policy ensures $p_i(S) \in (0, 1)$ for all valid states, $P(S' \mid S) > 0$ for all $S, S' \in \Omega$. Consequently, the chain is irreducible and aperiodic. By the Convergence Theorem for Markov chains (Levin & Peres, 2017, Corollary 1.17), the chain converges

to a unique stationary distribution π over Ω such that $\pi P = \pi$. Under this stationary distribution, the algorithm implicitly minimizes the objective $F(\theta)$. The convergence of the learning algorithm is characterized by the following parameters of this Markov chain:

1. **Mixing Time** (τ_{mix}): Let $P^r(S_0, \cdot)$ be the distribution of the chain at round r starting from initial state S_0. The mixing time is the smallest integer such that the total variation (TV) distance to the stationary distribution is bounded:
$$\sup_{S_0 \in \Omega} \|P^r(S_0, \cdot) - \pi(\cdot)\|_{TV} \leq \frac{1}{4}. \tag{6}$$
As well known, this guarantees geometric convergence to the stationary distribution, that is, $\|P^r(S_0, \cdot) - \pi(\cdot)\|_{TV} \leq c_0 \exp(-r/\tau_{\text{mix}})$ for some constant $c_0 > 0$.
2. **Stationary Participation Probability** (ρ_π): The probability of a non-empty update under stationary distribution:
$$\rho_\pi := \mathbb{P}_{S\sim\pi}(|S| > 0) = \sum_{S\in\Omega, |S|>0} \pi(S). \tag{7}$$
3. **Stationary Effective Sample Size** ($m_{\text{eff},\pi}$): The harmonic mean of the sample size under stationary distribution:
$$(m_{\text{eff},\pi})^{-1} := \mathbb{E}_{S\sim\pi}\left[|S|^{-1} \big| |S| > 0\right]. \tag{8}$$

3.4 General Convergence Bound under Dynamic Participation

Unless initialized at the stationary distribution, the distribution of the active set $\mathcal{S}^{(r)}$ drifts toward the stationary distribution π, so the participation probability $\rho^{(r)}$ and the effective sample size $m_{\text{eff}}^{(r)}$ continuously evolve. We now bound the expected loss by incorporating the chain's mixing time.

Theorem 2 (Dynamic Convergence Bound)**.** *Let τ_{mix} be the mixing time of the client participation Markov chain. Under Assumptions1,2 and 3, the expected steady-state error after T rounds satisfies:*
$$\mathbb{E}[F(\bar{\theta}^{(T)}) - F(\theta^\star)] \leq \frac{\|\theta^{(0)} - \theta^\star\|^2}{\eta\rho_\pi T} + \frac{\eta\sigma^2}{\rho_\pi m_{eff,\pi}} + \frac{c_0\tau_{mix}}{T\rho_\pi}\left(\eta\sigma^2 + \frac{L}{2}\|\theta^{(0)} - \theta^\star\|^2\right). \tag{9}$$

The proof of Theorem 2 appears in Appendix B.5. Theorem 2 decomposes the convergence dynamics into three effects. First, the effective total number of rounds is reduced to $\rho_\pi T$, and determined by the stationary participation probability. Second, the asymptotic variance is inversely proportional to the stationary effective sample size, $m_{\text{eff},\pi}$. Interestingly, it is an harmonic mean, dominated by events of low client participation. Third, a transient penalty term captures the error accumulated during the Markov chain's relaxation phase, of length τ_{mix}. This captures the effect of client autonomy on the loss. A highly autonomous participation rule will prolong the mixing time, and reduces the participation probability. This diverges the error terms and actively prevents convergence.

A particular case of this theorem is the *static* participation: The participation probabilities p_i remain constant, and the parameters ρ and m_{eff} are time-invariant, and the mixing time is zero. The resulting convergence bound under static participation is detailed in Theorem 3, Appendix B.3.

3.5 Autonomous Client Selection Analysis (Case 1)

We next propose a concrete client participation rule that models autonomous behavior. We analyze its mixing time and stationary properties, and use these and Theorem 2 to bound the loss. We focus on the setting of a mixed participation environment, consisting of both dynamic and uniform clients (Case 1). We extend this theoretical analysis to two additional environments in Appendices B.7 and B.8: Fully Dynamic Participation (Case 2), where all clients are autonomous and stability relies on a regularization parameter β, and Independent Dynamic Participation (Case 3), where clients make completely independent decisions, allowing the system to mix almost instantly.

We assume that the total of N_c clients are partitioned into two disjoint sets: (1) *Dynamic Set ($\mathcal{S}_d$)*: N_d clients that choose to participate based on a autonomy parameter b and the total number of clients who participated in the last round. (2) *Uniform Set ($\mathcal{S}_u$)*: N_u clients that choose to participate with fixed probability, p_u, acting as a baseline to ensure stability. We further model the transition

probability for the participation of client i as:

$$p_i(S) = \begin{cases} \frac{1}{1+\exp(b/|S|)}, & \text{if } i \in \mathcal{S}_d \text{ and } |S| > 0 \\ 0, & \text{if } i \in \mathcal{S}_d \text{ and } |S| = 0 \\ p_u, & \text{if } i \in \mathcal{S}_u \end{cases}. \tag{10}$$

The primary motivation for this model is that autonomous clients might be more inclined to participate in the learning process when the active set size $|S|$ is large. This is practically motivated by various reasons. For example, for privacy reasons, as the individual client's gradient is more diluted during server aggregation when $|S|$ is larger, or for efficiency reasons, as participating requires local resources such as battery and bandwidth, and a client may choose to withhold these resources during rounds with low participation, as a small $|S|$ yields less improvement to the global model.

Under this assumption, we define a simplified Markov chain where the state at round r is the active subset size $Y_r = |\mathcal{S}^{(r)}|$, reducing the state space size from 2^{N_c} to N_c+1. The probability distribution conditioned on state $Y_r = n$ is governed by the sum of two independent Binomial variables, $M_s \sim \text{Bin}(N_d, p_d(n))$ and $M_u \sim \text{Bin}(N_u, p_u)$. So, the transition matrix $Q \in [0,1]^{(N_c+1)\times(N_c+1)}$ is defined by the convolution:

$$\begin{aligned} Q_{nm} &= P(Y_{r+1} = m \mid Y_r = n) \\ &= \sum_{j=0}^{m} \left[\binom{N_d}{j} p_d(n)^j (1-p_d(n))^{N_d-j} \right] \left[\binom{N_u}{m-j} p_u^{m-j} (1-p_u)^{N_u-(m-j)} \right]. \end{aligned} \tag{11}$$

To analyze the stationary parameters, we define the expected mean-field drift of the active set size given the current state n as $\mu(n) = N_d p_d(n) + N_u p_u$.

Proposition 1 (Approximated Stationary Parameters)**.** *In the large population limit ($N_c \to \infty$), let $n^\star$ be the fixed point of the expected mean-field drift, satisfying $n^\star = N_d p_d(n^\star) + N_u p_u$, and let $\mu'(n^\star)$ be the derivative of the drift at this fixed point, where we treat the discrete state n as a continuous variable for the continuous relaxation of the mean-field approximation. Assuming the system is strictly stable ($|\mu'(n^\star)| < 1$), the stationary distribution of the active set size is approximately Gaussian, $\pi \approx \mathcal{N}(n^\star, \sigma_\pi^2)$, with asymptotic variance:*

$$\sigma_\pi^2 = \frac{N_d p_d(n^\star)(1-p_d(n^\star)) + N_u p_u (1-p_u)}{1-(\mu'(n^\star))^2}. \tag{12}$$

Denote by $\Phi(\cdot)$ the standard normal cumulative distribution function. The stationary participation probability ρ_π and the expected inverse effective sample size $1/m_{\mathit{eff},\pi}$ are closely approximated by:

$$\rho_\pi \approx \Phi\left(\frac{n^\star}{\sigma_\pi}\right), \quad \frac{1}{m_{\mathit{eff},\pi}} \approx \frac{1}{n^\star} + \frac{\sigma_\pi^2}{(n^\star)^3}, \tag{13}$$

Proposition 2 (Mixing Time Bound via Path Coupling)**.** *Let $L_p = \max_k |p_d(k+1) - p_d(k)|$ be the maximum absolute change in the dynamic clients' participation probability between adjacent state sizes. Under the condition that the contraction factor $\gamma := N_d L_p < 1$, the mixing time of the system's active set size is upper bounded by:*

$$\tau_{\mathit{mix}} \leq \left\lceil \frac{\log(N_c)+\log(4)}{\log\left(\frac{1}{\gamma}\right)} \right\rceil. \tag{14}$$

The proofs of Propositions 1 and 2 are in Appendices B.6.1 and B.6.2, respectively.

The Critical Autonomy Threshold The convergence characteristics of the proposed distributed learning framework relies on two conditions. First, the Markov chain governing client participation must rapidly approach stationarity relative to the total number of training rounds ($\tau_{\text{mix}} < T$). Second, the resulting stationary distribution must guarantee a strictly positive expected participation probability ($\rho_\pi > 0$). However, as the autonomy parameter b increases, clients exhibit a growing reluctance to transmit their local updates. To formally characterize the boundary of system viability, we define the *Critical Autonomy Threshold* (b_{critical}). When client autonomy exceeds this critical value ($b > b_{\text{critical}}$), the autonomous clients stop participating, and the distributed learning process relies only on the set of uniform clients. This leads to a high variance and steady-state error is higher.

Proposition 3 (Critical Threshold)**.** *Consider the mixed participation regime where the equilibrium active set size is $n^\star = x + N_u p_u$, with $x := N_d p_d(n^\star)$ denoting the expected number of participating*

dynamic (autonomous) clients. The system collapses when the contraction factor satisfies $\mu'(n^\star) = 1$*, and this occurs at a critical autonomy threshold:*

$$b_{critical} = x\left(1 - \frac{x}{N_d}\right)\left[\ln\left(\frac{N_d}{x} - 1\right)\right]^2. \tag{15}$$

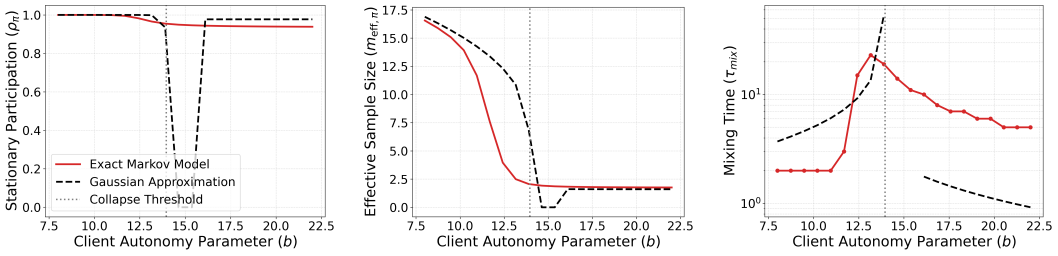
Furthermore, for a large distributed system near collapse ($x \ll N_d$)*, the required uniform stabilizing set* $N_u p_u$ *admit the following second-order asymptotic approximations:*

$$N_u p_u = x \ln\left(\frac{N_d}{x}\right) - x - \frac{x^2}{N_d}\left(\ln\left(\frac{N_d}{x}\right) + 1\right) + \mathcal{O}\left(\frac{x^3}{N_d^2}\right). \tag{16}$$

The complete proof is deferred to Appendix B.9.1. Later on, in Section 4, we will use $N_u p_u$ to establish a baseline stabilizing set for a server budget allocation algorithm. The derivations for the critical collapse thresholds (b_{critical}) under the Fully Dynamic (Case 2) and Independent Dynamic (Case 3) participation regimes are detailed in Appendices B.9.2 and B.9.3.

In Case 2, we remove the uniform set ($N_u = 0$) and introduce a regularization parameter β to ensure stability. As in Case 1, as the client autonomy parameter b increases, the system's contraction factor degrades until it collapses ($\mu'(n^\star) = 1$), causing the mixing time to diverge. Alternatively, in Case 3, each client makes a completely independent decision, and the Markov chain mixes almost instantly. Under this regime, exceeding the critical autonomy threshold causes the participation probability to drop to zero ($\rho_\pi \to 0$), entirely halting the learning process despite the rapid mixing time.

Figure 1 illustrates this collapse phenomenon. It displays the stationary metrics of the DL process as a function of the autonomy parameter b. As b approaches the critical threshold ($b_{\text{critical}} \approx 14$ for this configuration), the system undergoes a sharp phase transition. Both the stationary participation rate (ρ_π) and the effective sample size (m_{eff}) plummet, starving the global model of updates and increasing the gradient variance. Simultaneously, the mixing time (τ_{mix}) experiences a massive spike. This visualizes how exceeding b_{critical} violates the convergence requirements established in Theorem 2, preventing the Markov chain from reaching equilibrium. As observed, the Gaussian approximation is aligned with the exact Markov model when the system operates far from the critical threshold. However, as b approaches b_{critical}, the approximation predictably diverges since the linear noise approximation relies on a local linearization that breaks down when the derivative $\mu'(n^\star) \to 1$. Further empirical analysis, including finite-size scaling as $N_c \to \infty$, is detailed in Appendix E.1.



(a) Participation Rate (b) Effective Sample Size (c) Mixing Time

Figure 1: Case 1 participation stationary metrics as a function of the client autonomy parameter b.

4 Server Policy Optimization Problem

We have shown in Section 3 that low participation strongly affects the convergence, both via reduced effective number of rounds and effective sample size. Focusing on Case 1 analyzed therein, a uniform stabilizing set ($\mathcal{S}_u$) is necessary to prevent starvation of model updates under high client autonomy. Indeed, if all clients act autonomously, the system might get stuck in the Markov state $|S| = 0$, which will stop the learning process. In a real-world distributed systems, this can be solved by an intervention of the server, though this incurs a cost from the server to encourage participation.

Therefore, in this section, we define a server policy optimization problem where the server operates under a strict total budget, B, representing, e.g., financial costs or computational overheads. The server then faces the problem of determining an optimal resource allocation strategy among to

clients. The budget B is divided into individual rewards w for each of the uniform clients N_u. The server's challenge is the trade-off between querying a large number of uniform clients with small incentives, risking high variance and longer mixing time, or fewer clients with large incentives, which will guarantee their participation, yet reduce the effective sample size.

The individual reward w directly determines the participation probability, p_u, of the queried uniform clients. For concreteness, we model this response to incentives using a diminishing returns function:

$$p_u(w) = \frac{w}{1+w}. \tag{17}$$

This reflects a rational agent model, for which initial incentives significantly boost participation probability p_u, but the marginal utility of additional incentives diminishes as $p_u \to 1$.

The server's objective is to minimize the global steady-state convergence error $E(N_u, w)$ established in Theorem 2, subject to its budget constraints, which leads to the following objective:

$$E(N_u, w) := \frac{\|\theta^{(0)} - \theta^\star\|^2}{\eta\rho_\pi(N_u, w)T} + \frac{\eta\sigma^2}{\rho_\pi(N_u, w) \cdot m_{\text{eff},\pi}(N_u, w)} + \frac{c_0\tau_{\text{mix}}(N_u, w)\left(\eta\sigma^2 + \frac{L}{2}\|\theta^{(0)} - \theta^\star\|^2\right)}{T\rho_\pi(N_u, w)}, \tag{18}$$

and then to the constrained optimization problem:

$$\min_{N_u \in [1, N_c - N_d], w \geq 0} E(N_u, w) \quad \text{s.t.} \quad N_u \cdot w \leq B, \tag{19}$$

where $p_u(w) = \frac{w}{1+w}$, and the stationary metrics $\rho_\pi(N_u, w)$, $m_{\text{eff},\pi}(N_u, w)$, and $\tau_{\text{mix}}(N_u, w)$ are evaluated under the mixed participation Markov chain defined in Section 3.3. We assume that the transition matrix Q, which defines the state transition probabilities, and the gradient variance σ^2 are known to the PS. If these parameters are not piorly known, the PS can estimate them dynamically. Specifically, the server can observe the sequence of active set sizes $Y_r = |\mathcal{S}^{(r)}|$ over initial rounds to empirically estimate the transition matrix $\widehat{Q}$. Similarly, the server can approximate the variance $\widehat{\sigma}^2$ by computing the sample variance of the local updates received from the participating clients.

To solve the server's optimization problem, we leverage the objective function monotonicity as a function of the individual reward w. As w is increased, the participation probability of uniform clients is increased, and the global convergence error decreases. Consequently, a rational server will exhaust its budget to minimize the error, and the budget constraint will be tight, $N_u \cdot w = B$. Then, by substituting $w = \frac{B}{N_u}$ into the global steady-state convergence error $E(N_u, w)$, we reduce the 2-dimensional optimization problem into a 1-dimensional discrete optimization problem over the number of the participating uniform clients, $N_u \in [1, N_c - N_d]$. The details are in Appendix C.

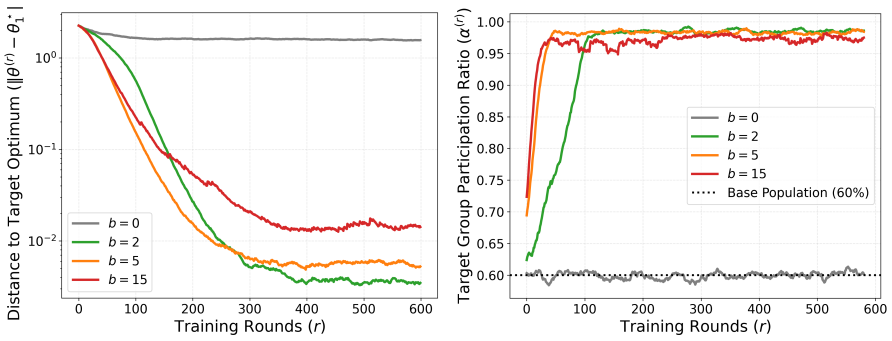
The objective function involves complex, non-differentiable stationary Markov expectations and spectral mixing times, so a closed-form solution for the optimal $N_u^\star$ is intractable. However, since N_u is bounded by the finite population of available uniform clients, $N_u^\star$ can be efficiently found via a 1-dimensional line search. Algorithm 1 in Appendix C, details this search process. To evaluate the convergence metrics at each step, the algorithm can be configured to use either our closed-form continuous approximations, detailed in Section 3.5, or the exact discrete Markov chain evaluations for smaller systems using the exact transition matrix from equation 11, or the estimated matrix $\widehat{Q}$. Once the optimal uniform client count $N_u^\star$ is identified, the optimal individual reward is trivially recovered as $w^\star = B/N_u^\star$. Empirical validation of this budget allocation strategy, illustrating the trade-off between client coverage and Markov mixing penalties, is detailed in Appendix E.2.

5 Extension to Heterogeneous Data

Our analysis thus far has isolated the effects of autonomous participation under homogeneous data, yet realistic distributed networks exhibit data heterogeneity. To model this, we extend our framework by dividing the client population into two disjoint groups: a primary target group G_1 (sampling from $\mathcal{D}_1$ with objective F_1) and a secondary group G_2 (sampling from $\mathcal{D}_2$ with objective F_2). At each round, let $m_1^{(r)}$ denote the number of participating clients from G_1, and $m^{(r)}$ denote the total number of active clients across both groups. Given that a round is successful ($m^{(r)} > 0$), the expected proportion of participating target clients is defined as $\alpha^{(r)} = \mathbb{E}[m_1^{(r)}/m^{(r)} \mid m^{(r)} > 0]$.

As we formally prove in the extended heterogeneous analysis (detailed in Appendix D), data heterogeneity introduces a steady-state convergence bias proportional to $(1-\alpha)^2\zeta_{1,2}$, where $\zeta_{1,2}$ bounds the gradient divergence ($\|\nabla F_2(\theta) - \nabla F_1(\theta)\|^2 \leq \zeta_{1,2}$) between the two groups. Under standard static participation, a low α permanently traps the global model at a high error floor.

However, under dynamic autonomous participation, $\alpha^{(r)}$ actively evolves in a powerful feedback loop. When clients employ a participation rule that reduces their participation probability as their local gradient norms diverge from the global model, the system can debias itself. If the global model begins to optimize toward the target objective F_1, the local losses and gradient norms for clients in G_2 naturally inflate. This divergence causes autonomous G_2 clients to drop out, rapidly skewing the participation ratio $\alpha^{(r)}$ toward 1.0, as illustrated in Figure 2b. This shift accelerates the global model's alignment with F_1, effectively purging the $\zeta_{1,2}$ bias penalty from the learning process.



(a) Distance to Target Optimum

(b) Target Group Participation Ratio

Figure 2: Impact of the autonomy parameter (b) on convergence with a $60/40$ population majority.

Despite this self-debiasing mechanism, the system remains bounded by the critical autonomy threshold. Figure 2a shows that moderate autonomy ($b = 2$) successfully filters out divergent updates and minimizes the target error, yet hyper-sensitive autonomy ($b = 15$) injects severe stochastic noise. When the selection threshold is too strict, even highly aligned clients from G_1 drop out due to minor stochastic variations in their local batches. This starvation severely reduces the effective sample size m_{eff}, inflating the variance penalty and raising the overall convergence error floor. Further empirical analysis detailing the impact of model initialization on this dynamic is provided in Appendix E.

6 Conclusion

In this paper, we bridged the gap between idealized distributed learning models and real-world network dynamics by analyzing how autonomous client behavior alters global model convergence. While we introduced a Markovian framework to formalize state-dependent client participation, our contribution lies in characterizing the theoretical limits and practical trade-offs of learning under client autonomy. By replacing the standard assumption of passive client compliance, we established new convergence error bounds under homogeneous and heterogeneous data distributions. Our analysis theoretically and empirically proves the existence of a critical autonomy threshold, beyond which the Markov chain fails to mix and the learning process suffers from a higher variance error. Furthermore, we demonstrated how autonomous dropout in heterogeneous environments creates a feedback loop that biases the global model, and we proposed a server budget allocation algorithm to guarantee system stability in highly autonomous networks.

In this work we focused on essential basic scenarios of learning dynamics of autonomous client participation. Real-world distributed networks offer much more to explore, including fully heterogeneous clients, more complex selection rules, and the simultaneous strategic selection by both clients and servers. As a concrete step toward this broader scope, in Appendix F, we extend our problem formulation to a competitive multi-server framework, where clients dynamically evaluate and choose between competing parameter servers. Extending our theoretical analysis to this multi-server setting appears to be challenging, but remains a highly compelling direction for future research.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, and assisting in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results, and assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, create or modify scientific figures or images, format references, and edit the research paper to improve readability. We have reviewed all AI-assisted work. Specifically, LLM-generated software code and simulation scripts were verified and tested for correctness by the authors and AI-assisted proof transcriptions and text edits were reviewed to ensure mathematical rigor and semantic precision. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Omer Ben-Porat and Moshe Tennenholtz. Best response regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Omer Ben-Porat and Moshe Tennenholtz. Regression equilibrium. In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*, pp. 73–91, 2019.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(3):1–45, 2024.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

Sarah Dean, Mihaela Curmei, Lillian J. Ratliff, Jamie Morgenstern, and Maryam Fazel. Emergent specialization from participation dynamics and multi-learner retraining. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Ohad Einav and Nir Rosenfeld. A market for accuracy: Classification under competition. In *International Conference on Machine Learning (ICML)*, 2025.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

Lei Fu, Huanle Zhang, Ge Gao, Mi Zhang, and Xin Liu. Client selection in federated learning: Principles, challenges, and opportunities. *IEEE Internet of Things Journal*, 10(24):21846–21865, 2023.

Antonio A. Ginart, Eva Zhang, Yongchan Kwon, and James Zou. Competing ai: How does competition feedback affect machine learning? In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pp. 1693–1701. PMLR, 2021.

Guy Horowitz, Yonatan Sommer, Moran Koren, and Nir Rosenfeld. Classification under strategic self-selection. In *International Conference on Machine Learning (ICML)*, 2024.

Meena Jagadeesan, Michael I. Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved bayes risk can yield reduced social welfare under competition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, second edition, 2017.

Alexandra L'Heureux, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021.

Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018.

Artavazd Maranjyan, Omar Shaikh Omar, and Peter Richtárik. Mindflayer sgd: Efficient parallel sgd in the presence of heterogeneous and random worker compute times. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2025.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications (ICC)*, pp. 1–7, 2019.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Zhenyu Sun, Ziyang Zhang, Zheng Xu, Gauri Joshi, Pranay Sharma, and Ermin Wei. Debiasing federated learning with correlated client participation. In *International Conference on Learning Representations (ICLR)*, 2025.

Lin Wang, Yongxin Guo, Tao Lin, and Xiaoying Tang. Delta: Diverse client sampling for fasting federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. In *International Conference on Machine Learning (ICML)*, pp. 1–9, 2015.

A Related Work

Client participation in FL. The common assumption in FL is that a central parameter server (PS) coordinates client participation to optimize training efficiency (Konečný et al., 2016; McMahan et al., 2017). To improve convergence rates and reduce stochastic variance under this centralized paradigm, various non-uniform sampling techniques have been proposed. These include importance sampling (Zhao & Zhang, 2015), diversity-driven client selection (Wang et al., 2023), and power-of-choice strategies (Cho et al., 2020). To account for the unreliability of mobile edge networks (Nishio & Yonetani, 2019), recent studies have also analyzed convergence under arbitrary or worst-case client participation (Wang & Ji, 2022).

Closest to our participation model, Sun et al. (2025) recognized that practical client availability is correlated across rounds. They introduced a framework that models client participation as a Markov chain to analyze FL convergence under system-imposed resting periods (e.g., devices requiring minimum intervals between active rounds). However, existing works still treat participation as an external constraint or a server-controlled policy (Fu et al., 2023). They do not reflect the clients' real-world behavior, where they act autonomously based on self-interest. In this paper, we extend the Markovian modeling of Sun et al. (2025) to formalize autonomous, state-dependent participation. Rather than relying on fixed hardware availability statistics, our Markov chain is driven dynamically by the clients' evaluation of the evolving global model.

Learning under competition. Our framework aligns with the growing literature on learning under competition, which abandons the assumption of a passive, static data distribution (L'Heureux et al., 2017). As distributed systems evolve into a reality of multiple service providers, the interaction between algorithms and data owners becomes explicitly competitive (Ginart et al., 2021). In multi-armed bandit scenarios, Mansour et al. (2018) analyzed the dynamics of algorithms directly competing for the same arms. These strategic, best-response behaviors have also been modeled in regression tasks (Ben-Porat & Tennenholtz, 2017; 2019). Furthermore, Einav & Rosenfeld (2025) and Horowitz et al. (2024) defined formal models of competition for classification tasks, demonstrating how users strategically self-select into different decision systems based on their own utility. Jagadeesan et al. (2023) demonstrated that in competitive classification, optimizing for a single model's individual performance (Bayes risk) can actually lead to reduced overall social welfare for the users. We extend these concepts of strategic self-selection directly into the distributed training phase. Similar to how Dean et al. (2024) observed emergent model specialization arising from users dynamically apportioning themselves among different learners, we demonstrate that when clients autonomously drop out due to local gradient divergence, it creates a feedback loop that steers the global model toward the most dominant data distributions.

Optimization under dynamic sampling. Robust optimization techniques are required when autonomous behavior introduces high variance and arbitrary dropouts into the learning process. At the algorithmic level, our setting shares mathematical similarities with optimization under heterogeneous resource constraints. Recent work by Maranjyan et al. (2025) establishes unbiasedness and variance bounds for parallel SGD in the presence of random and heterogeneous worker compute times. However, their gradient estimator scales the global update by a fixed expected batch size, relying on the assumption that the server can guarantee a non-empty round. In our setting, autonomous client participation means we cannot guarantee an active subset. Our estimator necessitates conditioning on strictly positive client participation ($|\mathcal{S}^{(r+1)}| > 0$ in the notation of this paper), which consequently alters the fundamental variance bounds. This critical distinction is what allows us to establish the existence of a critical autonomy threshold, beyond which the dynamic sampling fails to mix and the learning process stagnates.

B Homogeneous Data Analysis

B.1 Properties of the Aggregated Gradient

Lemma 1 (Properties of the Aggregated Gradient)**.** *Given that the active set $\mathcal{S}^{(r+1)}$ is non-empty, the aggregated gradient estimator $\widehat{g}^{(r)}$ is unbiased, satisfying:* $\mathbb{E}[\widehat{g}^{(r)} \mid \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0] =$

$\nabla F(\theta^{(r)})$*, and its variance is bounded by the expected inverse size of the active set:*

$$\mathbb{E}\left[\|\widehat{g}^{(r)} - \nabla F(\theta^{(r)})\|^2 \;\middle|\; \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0\right] \leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|} \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \sigma^2. \tag{4}$$

Proof of Lemma 1. We first prove the unbiasedness of the estimator. By the law of total expectation, we condition on the active client set $\mathcal{S}^{(r+1)}$. Leveraging the assumption that clients sample independent and identically distributed data from the global distribution, the expected value of the aggregated gradient is:

$$\begin{aligned}
&\mathbb{E}[\widehat{g}^{(r)} \mid \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0] \\
&= \mathbb{E}_{\mathcal{S}^{(r+1)}}\left[\mathbb{E}_{\xi}\left[\frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{j\in\mathcal{S}^{(r+1)}} g_j(\theta^{(r)}, \xi_j^{(r)}) \;\middle|\; \theta^{(r)}, \mathcal{S}^{(r+1)}\right] \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}_{\mathcal{S}^{(r+1)}}\left[\frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{j\in\mathcal{S}^{(r+1)}} \mathbb{E}_{\xi}\left[g_j(\theta^{(r)}, \xi_j^{(r)}) \mid \theta^{(r)}\right] \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right].
\end{aligned}$$

Because each local gradient evaluated on an i.i.d. and non-empty batch is an unbiased estimator of the global gradient (i.e., $\mathbb{E}_{\xi}[g_j(\theta^{(r)}, \xi_j^{(r)}) \mid \theta^{(r)}] = \nabla F(\theta^{(r)})$), the inner expectation simplifies:

$$\begin{aligned}
&\mathbb{E}[\widehat{g}^{(r)} \mid \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0] \\
&= \mathbb{E}_{\mathcal{S}^{(r+1)}}\left[\frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{j\in\mathcal{S}^{(r+1)}} \nabla F(\theta^{(r)}) \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}_{\mathcal{S}^{(r+1)}}\left[\frac{|\mathcal{S}^{(r+1)}|}{|\mathcal{S}^{(r+1)}|}\nabla F(\theta^{(r)}) \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \nabla F(\theta^{(r)}).
\end{aligned}$$

For the variance bound, using a similar conditioning approach and the independence of the clients' data samples (which zeroes out the cross-terms in the squared norm), we have:

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{g}^{(r)} - \nabla F(\theta^{(r)})\|^2 \;\middle|\; \theta^{(r)}, |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{|\mathcal{S}^{(r+1)}|}\sum_{j\in\mathcal{S}^{(r+1)}}\left(g_j(\theta^{(r)}, \xi_j^{(r)}) - \nabla F(\theta^{(r)})\right)\right\|^2 \;\middle|\; \theta^{(r)}, \mathcal{S}^{(r+1)}\right] \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|^2}\sum_{j\in\mathcal{S}^{(r+1)}} \mathbb{E}\left[\|g_j(\theta^{(r)}, \xi_j^{(r)}) - \nabla F(\theta^{(r)})\|^2 \mid \theta^{(r)}, \mathcal{S}^{(r+1)}\right] \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&\leq \mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|^2}\sum_{j\in\mathcal{S}^{(r+1)}} \sigma^2 \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}\left[\frac{|\mathcal{S}^{(r+1)}|\sigma^2}{|\mathcal{S}^{(r+1)}|^2} \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right] \\
&= \mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|} \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right]\sigma^2.
\end{aligned}$$

□

B.2 General One-Step Descent Bound

Theorem 1 (General One-Step Descent Bound)**.** *Under Assumptions1,2 and 3, for any learning rate $\eta > 0$, the expected distance to the optimal model $\theta^\star$ satisfies:*

$$\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 \leq \mathbb{E}\|\theta^{(r)} - \theta^\star\|^2 - 2\eta(1-\eta L)\mathbb{E}[\rho^{(r+1)} \cdot (F(\theta^{(r)}) - F(\theta^\star))] + \eta^2\sigma^2\mathbb{E}\left[\frac{\mathbb{I}^{(r+1)}}{|\mathcal{S}^{(r+1)}|}\right].$$

Proof. Expanding the squared distance yields:

$$\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 = \mathbb{E}\|\theta^{(r)} - \theta^\star\|^2 - 2\eta\mathbb{E}\langle \mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle + \eta^2\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2. \tag{20}$$

Using the definition of the indicator function and the law of total expectation used on the complementing events $|\mathcal{S}^{(r+1)}| > 0$ and $|\mathcal{S}^{(r+1)}| = 0$. Then, using the unbiasedness of the aggregated gradient (Lemma 1) and the convexity of F, we obtain:

$$\begin{aligned}
&\mathbb{E}\langle \mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle \mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \mid \mathcal{F}^{(r)}\right]\right] \\
&= \mathbb{E}\left[\rho^{(r+1)}\mathbb{E}\left[\langle \widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \mid \mathcal{F}^{(r)}, |\mathcal{S}^{(r+1)}| > 0\right]\right] \\
&= \mathbb{E}\left[\rho^{(r+1)}\langle \nabla F(\theta^{(r)}), \theta^{(r)} - \theta^\star\rangle\right] \\
&\geq \mathbb{E}\left[\rho^{(r+1)}(F(\theta^{(r)}) - F(\theta^\star))\right].
\end{aligned}$$

Next, we bound the second moment of the effective gradient update using L-smoothness:

$$\begin{aligned}
&\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}^{(r+1)}(\|\nabla F(\theta^{(r)})\|^2 + \|\widehat{g}^{(r)} - \nabla F(\theta^{(r)})\|^2) \mid \mathcal{F}^{(r)}, \mathcal{S}^{(r+1)}\right]\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\rho^{(r+1)}\|\nabla F(\theta^{(r)})\|^2\right] + \sigma^2\mathbb{E}\left[\rho^{(r+1)}\mathbb{E}\left[\frac{1}{|\mathcal{S}^{(r+1)}|} \;\middle|\; |\mathcal{S}^{(r+1)}| > 0\right]\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\rho^{(r+1)}\left(2L(F(\theta^{(r)}) - F(\theta^\star))\right)\right] + \sigma^2\mathbb{E}\left[\frac{\mathbb{I}^{(r+1)}}{|\mathcal{S}^{(r+1)}|}\right],
\end{aligned}$$

where (a) follows since Lemma 1, and (b) follows from the L-smoothness, and the law of total expectation used on the complementing events $|\mathcal{S}^{(r+1)}| > 0$ and $|\mathcal{S}^{(r+1)}| = 0$.

Substituting these bounds back into equation 20 completes the proof. □

B.3 Static Participation Convergence

Before analyzing complex state-dependent dynamic participation, we first establish the convergence under static participation policies as a baseline. In this regime, the participation probabilities p_i remain constant across all rounds. Consequently, the participation probability and the effective sample size are time-invariant ($\rho^{(r+1)} = \rho$ and $m_{\text{eff}}^{(r+1)} = m_{\text{eff}}$). Applying these constants yields the following steady-state bound.

Theorem 3 (Convergence under Static Participation)**.** *Assume the participation probabilities p_i are static over time, yielding constant participation probability ρ and effective sample size m_{eff}. For a step size $\eta \leq \frac{1}{2L}$, the expected suboptimality of the averaged model $\bar{\theta}^{(T)} = \frac{1}{T}\sum_{r=0}^{T-1}\theta^{(r)}$ after T rounds is bounded by:*

$$\mathbb{E}[F(\bar{\theta}^{(T)}) - F(\theta^\star)] \leq \frac{\|\theta^{(0)} - \theta^\star\|^2}{\eta\rho T} + \frac{\eta\sigma^2}{m_{\text{eff}}}. \tag{21}$$

Proof. Substituting the time-invariant constants ρ and m_{eff} into Theorem 1 and summing over T rounds yields a telescoping sum:

$$\sum_{r=0}^{T-1} 2\eta\rho(1-\eta L)\mathbb{E}[F(\theta^{(r)}) - F(\theta^\star)] \leq \|\theta^{(0)} - \theta^\star\|^2 - \mathbb{E}\|\theta^{(T)} - \theta^\star\|^2 + \sum_{r=0}^{T-1}\frac{\eta^2\sigma^2\rho}{m_{\text{eff}}}. \tag{22}$$

By choosing a step size $\eta \leq \frac{1}{2L}$, we guarantee $2(1-\eta L) \geq 1$. Dropping the negative distance term $-\mathbb{E}\|\theta^{(T)} - \theta^\star\|^2$, dividing both sides by $\eta\rho T$, and applying Jensen's inequality for the averaged model $\bar{\theta}^{(T)}$ gives the final result. □

The precise formulations of ρ and m_{eff} depend on the underlying static participation distribution. We highlight two specific cases of static participation (Binomial and Poisson-Binomial) that explicitly instantiate these parameters in Appendix B.4.

B.4 Static Participation Examples

Case 1: Homogeneous Static Participation. If each client participates independently with an identical probability p, the subset size follows a standard Binomial distribution $|\mathcal{S}| \sim \text{Bin}(N_c, p)$.

- Participation probability: $\rho_{\text{bin}} = 1 - (1-p)^{N_c}$.
- Effective sample size: For large $N_c p$, the effective sample size is well-approximated by the expected value $m_{\text{eff}} \approx N_c p$. The exact harmonic mean formulation is:

$$\frac{1}{m_{\text{eff}}} = \frac{1}{\rho_{\text{bin}}} \sum_{k=1}^{N_c} \frac{1}{k} \binom{N_c}{k} p^k (1-p)^{N_c-k}. \tag{23}$$

Case 2: Heterogeneous Static Participation. If clients possess fixed but distinct participation probabilities p_i, the active subset size follows a Poisson-Binomial distribution $|\mathcal{S}| \sim \text{PB}(p_1, \ldots, p_{N_c})$.

- Participation probability: $\rho_{\text{pb}} = 1 - \prod_{i=1}^{N_c}(1-p_i)$.
- Effective sample size:

$$\frac{1}{m_{\text{eff}}} = \frac{1}{\rho_{\text{pb}}} \sum_{k=1}^{N_c} \frac{1}{k} \mathbb{P}(|\mathcal{S}|_{\text{PB}} = k). \tag{24}$$

When each client participates independently with the same probability p, the subset size of participating clients at each round follows the Binomial distribution $|\mathcal{S}| \sim \text{Bin}(N_c, p)$. In this case, as the total number of clients N_c increases, the probability of an empty round vanishes exponentially ($\rho_{\text{bin}} \to 1$) and the effective sample size approaches the expected value ($m_{\text{eff}} \approx N_c p$). This effectively reduces the convergence bound to that of standard mini-batch SGD. Conversely, when clients possess distinct participation probabilities p_i, the subset size follows a Poisson-Binomial distribution $|\mathcal{S}| \sim \text{PB}(p_1, \ldots, p_{N_c})$. In this scenario, the convergence dynamics depend not only on the average participation probability but also on its variance. A higher variance across the individual p_i values yields a lower effective sample size m_{eff}, which directly elevates the steady-state variance error floor. Furthermore, the effective convergence speed, ρT, is impeded when a large proportion of clients have a low p_i.

B.5 Dynamic Convergence Bound

Theorem 2 (Dynamic Convergence Bound)**.** *Let τ_{mix} be the mixing time of the client participation Markov chain. Under Assumptions1,2 and 3, the expected steady-state error after T rounds satisfies:*

$$\mathbb{E}[F(\bar{\theta}^{(T)}) - F(\theta^\star)] \leq \frac{\|\theta^{(0)} - \theta^\star\|^2}{\eta\rho_\pi T} + \frac{\eta\sigma^2}{\rho_\pi m_{\mathit{eff},\pi}} + \frac{c_0\tau_{mix}}{T\rho_\pi}\left(\eta\sigma^2 + \frac{L}{2}\|\theta^{(0)} - \theta^\star\|^2\right). \tag{9}$$

Proof. We define two functions on the state space: the participation probability indicator $f_\rho(S) = \mathbf{1}_{|S|>0}$, and the efficiency indicator $h_m(S) = \frac{1}{|S|}\mathbf{1}_{|S|>0}$. Because the chain converges to its stationary distribution in total variation, the variational characterization of the total variation distance guarantees that the expectation of any bounded function ϕ approaches its stationary expectation as $|\mathbb{E}[\phi(S_r)] - \mathbb{E}_\pi[\phi(S)]| \leq \|\phi\|_\infty c_0 e^{-r/\tau_{\text{mix}}}$.

Applying this to $h_m(S)$, we bound the expected inverse sample size:

$$\mathbb{E}\left[\frac{\mathbb{I}^{(r+1)}}{|S^{(r+1)}|}\right] = \mathbb{E}_\pi[h_m(S_{r+1})] + |\mathbb{E}[h_m(S_{r+1})] - \mathbb{E}_\pi[h_m(S)]|$$
$$\leq \frac{1}{m_{\text{eff},\pi}} + \|h_m\|_\infty d_{TV}(P^{r+1}, \pi) \leq \frac{1}{m_{\text{eff},\pi}} + c_0 e^{-(r+1)/\tau_{\text{mix}}}.$$

Similarly, for the participation probability $\rho^{(r+1)} = \mathbb{E}[f_\rho(S_{r+1})]$:

$$\rho^{(r+1)} \geq \rho_\pi - \|f_\rho\|_\infty d_{TV}(P^{r+1}, \pi) \geq \rho_\pi - c_0 e^{-(r+1)/\tau_{\text{mix}}}.$$

We substitute these dynamic bounds into Theorem 1. Setting $\eta \leq \frac{1}{2L}$ yields:

$$\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 \leq \mathbb{E}\|\theta^{(r)} - \theta^\star\|^2 - \eta\left(\rho_\pi - c_0 e^{-(r+1)/\tau_{\text{mix}}}\right)\mathbb{E}[F(\theta^{(r)}) - F(\theta^\star)]$$
$$+ \eta^2\sigma^2\left(\frac{1}{m_{\text{eff},\pi}} + c_0 e^{-(r+1)/\tau_{\text{mix}}}\right).$$

Rearranging and summing from $r = 0$ to $T-1$, we bound the geometric series sum $\sum_{r=1}^{T} e^{-r/\tau_{\text{mix}}} \leq \tau_{\text{mix}}$. We note that under L-smoothness, the expected suboptimality at any round r is bounded by the initial gap: $\mathbb{E}[F(\theta^{(r)}) - F(\theta^\star)] \leq F(\theta^{(0)}) - F(\theta^\star) \leq \frac{L}{2}\|\theta^{(0)} - \theta^\star\|^2$. Substituting these bounds, we obtain:

$$\eta\rho_\pi \sum_{r=0}^{T-1} \mathbb{E}[F(\theta^{(r)}) - F(\theta^\star)] \leq \|\theta^{(0)} - \theta^\star\|^2 + \frac{T\eta^2\sigma^2}{m_{\text{eff},\pi}} + \eta^2\sigma^2 c_0\tau_{\text{mix}} + \eta c_0\tau_{\text{mix}}\frac{L}{2}\|\theta^{(0)} - \theta^\star\|^2.$$

Dividing by $\eta\rho_\pi T$ and applying Jensen's Inequality for $\bar{\theta}^{(T)}$ yields the final bound. □

B.6 Case 1: Mixed Dynamic and Uniform participation

B.6.1 Gaussian Approximation

Proposition 1 (Approximated Stationary Parameters)**.** *In the large population limit ($N_c \to \infty$), let $n^\star$ be the fixed point of the expected mean-field drift, satisfying $n^\star = N_d p_d(n^\star) + N_u p_u$, and let $\mu'(n^\star)$ be the derivative of the drift at this fixed point, where we treat the discrete state n as a continuous variable for the continuous relaxation of the mean-field approximation. Assuming the system is strictly stable ($|\mu'(n^\star)| < 1$), the stationary distribution of the active set size is approximately Gaussian, $\pi \approx \mathcal{N}(n^\star, \sigma_\pi^2)$, with asymptotic variance:*

$$\sigma_\pi^2 = \frac{N_d p_d(n^\star)(1 - p_d(n^\star)) + N_u p_u(1 - p_u)}{1 - (\mu'(n^\star))^2}. \tag{12}$$

Denote by $\Phi(\cdot)$ the standard normal cumulative distribution function. The stationary participation probability ρ_π and the expected inverse effective sample size $1/m_{\text{eff},\pi}$ are closely approximated by:

$$\rho_\pi \approx \Phi\left(\frac{n^\star}{\sigma_\pi}\right), \quad \frac{1}{m_{\text{eff},\pi}} \approx \frac{1}{n^\star} + \frac{\sigma_\pi^2}{(n^\star)^3}, \tag{13}$$

Proof. **Mean-Field Approximation:** To derive closed-form bounds on τ_{mix} in the large N_c limit, we apply a Mean-Field approximation (Ethier & Kurtz, 1986, Chapter 11). The expected drift function given current state $Y_r = n$ is:

$$\mu(n) = \mathbb{E}[Y_{r+1} \mid Y_r = n] = N_d p_d(n) + N_u p_u. \tag{25}$$

The stationary state is characterized by the fixed point

$$n^\star = N_d p_d(n^\star) + N_u p_u. \tag{26}$$

Around this fixed point, the variance is

$$\sigma^2(n^\star) = N_d p_d(n^\star)(1 - p_d(n^\star)) + N_u p_u(1 - p_u). \tag{27}$$

Linear Noise Approximation: We linearize the system dynamics around $n^\star$ using the perturbation $\delta Y_r = Y_r - n^\star$, yielding $\delta Y_{r+1} \approx \mu'(n^\star)\delta Y_r + \xi_r$, where $\xi_r \sim \mathcal{N}(0, \sigma^2(n^\star))$ (Ethier & Kurtz, 1986, Chapter 11). This behaves as a discrete-time Ornstein–Uhlenbeck process, or an $AR(1)$ autoregressive model.

Under this stability regime, the stationary distribution is approximately Gaussian $\pi \approx \mathcal{N}(n^\star, \sigma_\pi^2)$ with asymptotic variance $\sigma_\pi^2 = \sigma^2(n^\star)/(1-(\mu'(n^\star))^2)$.

We can approximate the target stationary statistics as:

$$\rho_\pi \approx 1 - \Phi\left(\frac{-n^\star}{\sigma_\pi}\right) = \Phi\left(\frac{n^\star}{\sigma_\pi}\right) \tag{28}$$

$$\frac{1}{m_{\text{eff},\pi}} \approx \frac{1}{n^\star} + \frac{\sigma_\pi^2}{(n^\star)^3} \tag{29}$$

where $\Phi(\cdot)$ is the standard normal CDF.

The mixing time is proportional to the relaxation time of the $AR(1)$ process given by $\tau_{\text{mix}} = \mathcal{O}\left(-1/\ln|\mu'(n^\star)|\right)$. □

B.6.2 Mixing Time Bound via Path Coupling

Proposition 2 (Mixing Time Bound via Path Coupling). *Let $L_p = \max_k |p_d(k+1) - p_d(k)|$ be the maximum absolute change in the dynamic clients' participation probability between adjacent state sizes. Under the condition that the contraction factor $\gamma := N_d L_p < 1$, the mixing time of the system's active set size is upper bounded by:*

$$\tau_{mix} \leq \left\lceil \frac{\log(N_c)+\log(4)}{\log\left(\frac{1}{\gamma}\right)} \right\rceil. \tag{14}$$

Proof. Let $L_p = \max_k |p_d(k+1) - p_d(k)|$ be the Lipschitz constant of the dynamic probability. According to the Path Coupling Theorem (Levin & Peres, 2017, Theorem 14.6), if we can define a joint distribution (X_{r+1}, Y_{r+1}) for any adjacent states $X_r = k$ and $Y_r = k+1$ such that the expected distance strictly decreases, this contraction generalizes to all pairs.

In the next round, $X_{r+1} = U_X + D_X$ and $Y_{r+1} = U_Y + D_Y$, where uniform clients are drawn identically $U \sim \text{Bin}(N_u, p_u)$, and dynamic clients are $D_X \sim \text{Bin}(N_d, p_d(k))$ and $D_Y \sim \text{Bin}(N_d, p_d(k+1))$. The expected distance between the chains is:

$$\mathbb{E}[d(X_{r+1}, Y_{r+1}) \mid X_r = k, Y_r = k+1] = \mathbb{E}[|D_X - D_Y|] = N_d|p_d(k) - p_d(k+1)|. \tag{30}$$

Taking the maximum over all possible adjacent states yields the contraction factor $\gamma := N_d L_p$. Assuming $\gamma < 1$, the path coupling theorem extends this linearly to any two arbitrary initial states $x, y \in \Omega$ such that $\mathbb{E}[d(X_t, Y_t)] \leq \gamma^t N_c$.

By the Coupling Lemma and Markov's Inequality, the total variation distance is bounded by the expected distance: $d_{TV}(P^t, \pi) \leq \gamma^t N_c$. Setting $d_{TV} \leq \frac{1}{4}$ yields $\gamma^t \leq \frac{1}{4N_c}$. Solving for t provides the strict upper bound:

$$\tau_{\text{mix}} \leq \left\lceil \frac{\log(N_c) + \log(4)}{\log\left(\frac{1}{\gamma}\right)} \right\rceil. \tag{31}$$

□

B.7 Case 2: Fully Dynamic Participation

If all clients operate under the dynamic policy ($N_c = N_d$ and $N_u = 0$), the uniform stabilizing set is removed. To prevent the system from collapsing into an absorbing state when no clients participate (i.e., resolving the singularity at $|S| = 0$), we introduce a regularization parameter $\beta > 0$ to the gradient normalization. The transition probability simplifies to:

$$p_i(S) = \frac{1}{1 + \exp\left(\frac{b}{|S|+\beta}\right)} := p_d(|S|). \tag{32}$$

By introducing β, $p_d(0) = (1+\exp(b/\beta))^{-1} > 0$. The state $|S| = 0$ is no longer absorbing, guaranteeing an irreducible and aperiodic chain. The expected drift is $\mu(n) = N_c \cdot p_d(n)$, and the derivative is:

$$\mu'(n^\star) = \frac{b}{(n^\star+\beta)^2} N_c p_d(n^\star)(1 - p_d(n^\star)). \tag{33}$$

By the same Path Coupling Theorem, provided $\gamma := N_c L_p < 1$, the mixing time is bounded by:

$$\tau_{\text{mix}} \leq \left\lceil \frac{\log(N_c) + \log(4)}{\log\left(\frac{1}{\gamma}\right)} \right\rceil. \tag{34}$$

As $b \to b_{\text{critical}}$, the maximum derivative L_p scales such that $N_c L_p \to 1$, causing the mixing time to diverge.

B.8 Case 3: Independent Dynamic Participation

In this regime, each client's participation depends exclusively on its own state in the previous round. Let $X_i^{(r)} \in \{0, 1\}$ be the indicator variable that client i participates in round r. This forms an independent 2-state Markov chain for each client with transition probabilities

$$\mathbb{P}(X_i^{(r+1)} = 1 \mid X_i^{(r)} = 0) = \epsilon$$
$$\mathbb{P}(X_i^{(r+1)} = 0 \mid X_i^{(r)} = 1) = \delta.$$

The N_c chains are independent, leading to a unique stationary distribution for a single client $\pi_{i,1} = \frac{\epsilon}{\epsilon+\delta}$. Let $p^\star = \frac{\epsilon}{\epsilon+\delta}$ be this stationary probability. The active set size follows a Binomial distribution $|S| \sim \text{Bin}(N_c, p^\star)$. The stationary statistics are:

$$\rho_\pi = 1 - \left(\frac{\delta}{\epsilon + \delta}\right)^{N_c} \tag{35}$$

$$\frac{1}{m_{\text{eff},\pi}} = \frac{1}{\rho_\pi} \sum_{k=1}^{N_c} \frac{1}{k} \binom{N_c}{k} (p^\star)^k (1 - p^\star)^{N_c - k} \approx \frac{\epsilon + \delta}{N_c \epsilon}. \tag{36}$$

To bound the mixing time, we use coupling Levin & Peres (2017). We consider two independent copies of the N_c-client Markov chain, $\mathbf{X}^{(r)}$ and $\mathbf{Y}^{(r)}$, starting from arbitrary initial states. We define a coupling such that each pair of coordinates $(X_i^{(r)}, Y_i^{(r)})$ evolves independently according to the same transition matrix.

The coalescence time τ_{couple} is the first time the two global chains meet. Since clients are independent, $\tau_{\text{couple}} = \max\{T_1, \ldots, T_{N_c}\}$, where T_i is the coalescence time of pair i. The second eigenvalue of the 2-state transition matrix is $\lambda_2 = 1 - \epsilon - \delta$, yielding a spectral gap $\gamma = \epsilon + \delta$. The probability that a specific pair fails to couple by round r is $\mathbb{P}(T_i > r) = (1 - \gamma)^r$.

By Corollary 5.5 in Levin & Peres (2017), the total variation distance $d_{TV}(r)$ is bounded by the maximum probability that the chains have not coupled. Applying a union bound:

$$d_{TV}(r) \leq \mathbb{P}\left(\bigcup_{i=1}^{N_c} \{T_i > r\}\right) \leq \sum_{i=1}^{N_c} \mathbb{P}(T_i > r) = N_c (1 - \gamma)^r. \tag{37}$$

By definition, the mixing time requires $d_{TV}(\tau_{\text{mix}}) \leq \frac{1}{4}$.

$$N_c (1 - \gamma)^r \leq \frac{1}{4} \implies r \geq \frac{\log(N_c) + \log(4)}{\epsilon + \delta}. \tag{38}$$

$$\tau_{\text{mix}} \leq \left\lceil \frac{\log(N_c) + \log(4)}{\epsilon + \delta} \right\rceil. \tag{39}$$

B.9 Critical Autonomy Threshold

B.9.1 Case 1: Mixed Dynamic and Uniform Participation

Proposition 3 (Critical Threshold)**.** *Consider the mixed participation regime where the equilibrium active set size is $n^\star = x + N_u p_u$, with $x := N_d p_d(n^\star)$ denoting the expected number of participating dynamic (autonomous) clients. The system collapses when the contraction factor satisfies $\mu'(n^\star) =$*

1*, and this occurs at a critical autonomy threshold:*

$$b_{critical} = x\left(1 - \frac{x}{N_d}\right)\left[\ln\left(\frac{N_d}{x} - 1\right)\right]^2. \tag{15}$$

Furthermore, for a large distributed system near collapse ($x \ll N_d$), the required uniform stabilizing set $N_u p_u$ admit the following second-order asymptotic approximations:

$$N_u p_u = x\ln\left(\frac{N_d}{x}\right) - x - \frac{x^2}{N_d}\left(\ln\left(\frac{N_d}{x}\right) + 1\right) + \mathcal{O}\left(\frac{x^3}{N_d^2}\right). \tag{16}$$

Proof. In Markov chain theory, the mixing time is dictated by the spectral gap (the difference between the largest eigenvalue, which is always 1, and the second-largest eigenvalue, λ_2). The smaller the gap, the longer it takes to mix (Levin & Peres, 2017, Chapter 12). When we approximate the Markov chain as a dynamical system around its fixed point, the derivative $\mu'(n^\star)$ acts as the equivalent of that second eigenvalue ($\lambda_2 \approx \mu'(n^\star)$) . It represents the contraction factor of the system. Therefore, to find the approximate b_{critical}, we evaluate the derivative of the drift function at the fixed point and set it equal to 1:

$$\mu'(n^\star) = \left.\frac{\partial \mu}{\partial n}\right|_{n^\star} = \frac{b}{(n^\star)^2} N_d p_d(n^\star)(1 - p_d(n^\star)) = 1. \tag{40}$$

We define $x := N_d p_d(n^\star)$ as the expected number of autonomous clients that choose to participate when the system reaches equilibrium.

By definition, the total expected participation at the fixed point must equal the sum of expected autonomous and uniform clients participating:

$$n^\star = x + N_u p_u. \tag{41}$$

We first express b in terms of the fixed point state x. From the logistic definition

$$p_d(n^\star) = \left(1 + \exp\left(\frac{b}{n^\star}\right)\right)^{-1},$$

and we get:

$$b = (x + N_u p_u)\ln\left(\frac{N_d}{x} - 1\right). \tag{42}$$

To find the collapse threshold, we substitute $n^\star = x + N_u p_u$ and $N_d p_d(n^\star) = x$ into the critical derivative condition $1 = \mu'(n^\star)$, yielding:

$$1 = \frac{b}{(x + N_u p_u)^2} x\left(1 - \frac{x}{N_d}\right). \tag{43}$$

We isolate the term $(x + N_u p_u)$ from our expression for b as $(x + N_u p_u) = \frac{b}{\ln\left(\frac{N_d}{x} - 1\right)}$ and substitute it into the denominator of the derivative equation:

$$1 = \frac{b}{\left(\frac{b}{\ln(N_d/x - 1)}\right)^2} x\left(1 - \frac{x}{N_d}\right) = \frac{[\ln(N_d/x - 1)]^2}{b} x\left(1 - \frac{x}{N_d}\right). \tag{44}$$

Multiplying both sides by b yields an exact, closed-form parametric equation for the critical threshold:

$$b_{\text{critical}} = x\left(1 - \frac{x}{N_d}\right)\left[\ln\left(\frac{N_d}{x} - 1\right)\right]^2. \tag{45}$$

To extract a tighter closed-form approximation for a large distributed system near collapse ($x \ll N_d$), we apply a second-order Taylor expansion. Expanding the logarithmic term $\ln(1 - z) = -z + \mathcal{O}(z^2)$:

$$\ln\left(\frac{N_d}{x} - 1\right) = \ln\left(\frac{N_d}{x}\left(1 - \frac{x}{N_d}\right)\right) = \ln\left(\frac{N_d}{x}\right) - \frac{x}{N_d} + \mathcal{O}\left(\frac{x^2}{N_d^2}\right). \tag{46}$$

Squaring this expanded term and substituting it back into the exact equation for b_{critical} gives:

$$b_{\text{critical}} = \left(x - \frac{x^2}{N_d}\right)\left[\ln^2\left(\frac{N_d}{x}\right) - 2\frac{x}{N_d}\ln\left(\frac{N_d}{x}\right) + \mathcal{O}\left(\frac{x^2}{N_d^2}\right)\right]. \tag{47}$$

Dropping terms of order $\mathcal{O}(x^3/N_d^2)$ and higher, we obtain the second-order asymptotic approximation:

$$b_{\text{critical}} = x\ln^2\left(\frac{N_d}{x}\right) - \frac{x^2}{N_d}\ln\left(\frac{N_d}{x}\right)\left[\ln\left(\frac{N_d}{x}\right) + 2\right] + \mathcal{O}\left(\frac{x^3}{N_d^2}\right) \tag{48}$$

Finally, we connect this threshold back to the expected uniform stabilizing set $N_u p_u$. Substituting the exact b into the identity $(x + N_u p_u) = \frac{b}{\ln(N_d/x-1)}$ and applying the same Taylor expansion yields the formal relation:

$$N_u p_u = x\ln\left(\frac{N_d}{x}\right) - x - \frac{x^2}{N_d}\left(\ln\left(\frac{N_d}{x}\right) + 1\right) + \mathcal{O}\left(\frac{x^3}{N_d^2}\right). \tag{49}$$

□

B.9.2 Case 2: Fully Dynamic Participation

Proof. When the uniform set is removed ($N_u = 0$), stability is entirely reliant on the regularization parameter β. The expected drift is $\mu(n) = N_c p_d(n)$, and the fixed point satisfies $n^\star = \frac{N_c}{1+\exp\left(\frac{b}{n^\star+\beta}\right)}$. The derivative is:

$$\mu'(n^\star) = \frac{b}{(n^\star+\beta)^2} N_c p_d(n^\star)(1 - p_d(n^\star)).$$

Similar to Case 1, we first express b in terms of the fixed point state $n^\star$. From the logistic definition $p_d(n^\star) = \frac{n^\star}{N_c} = \left(1 + \exp\left(\frac{b}{n^\star+\beta}\right)\right)^{-1}$, we isolate the exponential term:

$$\exp\left(\frac{b}{n^\star+\beta}\right) = \frac{N_c}{n^\star} - 1$$

Taking the natural logarithm yields the exact expression for b:

$$b = (n^\star+\beta)\ln\left(\frac{N_c}{n^\star} - 1\right).$$

To find the absolute threshold, we enforce the boundary of the contraction condition $1 = \mu'(n^\star)$. By substituting $N_c p_d(n^\star) = n^\star$ into the derivative, we get:

$$1 = \frac{b}{(n^\star+\beta)^2} n^\star\left(1 - \frac{n^\star}{N_c}\right).$$

We isolate $(n^\star + \beta)$ from our expression for b as $(n^\star + \beta) = \frac{b}{\ln(N_c/n^\star-1)}$ and substitute it into the denominator of the derivative equation:

$$1 = \frac{b}{\left(\frac{b}{\ln(N_c/n^\star-1)}\right)^2} n^\star\left(1 - \frac{n^\star}{N_c}\right) = \frac{[\ln(N_c/n^\star-1)]^2}{b} n^\star\left(1 - \frac{n^\star}{N_c}\right).$$

Multiplying both sides by b yields the exact, closed-form parametric equation for the critical threshold:

$$b_{\text{critical}} = n^\star\left(1 - \frac{n^\star}{N_c}\right)\left[\ln\left(\frac{N_c}{n^\star} - 1\right)\right]^2.$$

To extract a tighter closed-form approximation for a large distributed system near collapse ($n^\star \ll N_c$), we apply a second-order Taylor expansion. Expanding the logarithmic term $\ln(1-z) = -z + \mathcal{O}(z^2)$:

$$\ln\left(\frac{N_c}{n^\star} - 1\right) = \ln\left(\frac{N_c}{n^\star}\left(1 - \frac{n^\star}{N_c}\right)\right) = \ln\left(\frac{N_c}{n^\star}\right) - \frac{n^\star}{N_c} + \mathcal{O}\left(\frac{(n^\star)^2}{N_c^2}\right).$$

Squaring this expanded term and substituting it back into the exact equation for b_{critical} yields:

$$b_{\text{critical}} = \left(n^\star - \frac{(n^\star)^2}{N_c}\right)\left[\ln^2\left(\frac{N_c}{n^\star}\right) - 2\frac{n^\star}{N_c}\ln\left(\frac{N_c}{n^\star}\right) + \mathcal{O}\left(\frac{(n^\star)^2}{N_c^2}\right)\right].$$

Dropping terms of order $\mathcal{O}((n^\star)^3/N_c^2)$ and higher, we obtain the second-order asymptotic approximation:

$$b_{\text{critical}} = n^\star \ln^2\left(\frac{N_c}{n^\star}\right) - \frac{(n^\star)^2}{N_c}\ln\left(\frac{N_c}{n^\star}\right)\left[\ln\left(\frac{N_c}{n^\star}\right) + 2\right] + \mathcal{O}\left(\frac{(n^\star)^3}{N_c^2}\right).$$

Finally, we connect this threshold back to the regularization parameter β. Substituting the exact b into the identity $(n^\star + \beta) = \frac{b}{\ln(N_c/n^\star - 1)}$ and applying the same Taylor expansion gives the formal relation:

$$\beta = n^\star \ln\left(\frac{N_c}{n^\star}\right) - n^\star - \frac{(n^\star)^2}{N_c}\left(\ln\left(\frac{N_c}{n^\star}\right) + 1\right) + \mathcal{O}\left(\frac{(n^\star)^3}{N_c^2}\right).$$

□

B.9.3 Case 3: Independent Dynamic Participation

This case models a scenario where each client makes completely independent decisions. Each client has an independent two-state Markov chain governed by an entry rate, ϵ, and an exit rate, δ.

To capture autonomous behavior, we assume these transition rates are functions of the autonomy parameter $b \geq 0$. As b increases, the perceived cost of participating (e.g., privacy leakage, computational wear) outweighs the benefits. Therefore, we assume the entry rate $\epsilon(b)$ is a monotonically decreasing function of b, and the exit rate $\delta(b)$ is a monotonically increasing function of b.

To instantiate these dynamics, let $\epsilon_0, \delta_0 \in (0, 1)$ represent the baseline transition rates of a purely altruistic client ($b = 0$). We can model the state-dependent transition probabilities using rational decay functions:

$$\epsilon(b) = \frac{\epsilon_0}{1 + b}, \tag{50}$$

$$\delta(b) = \frac{\delta_0 + b}{1 + b}. \tag{51}$$

Proof. For an independent two-state Markov chain, the spectral gap is defined as the sum of its transition probabilities: $\gamma = \epsilon(b) + \delta(b)$. Substituting our specific rate functions yields:

$$\gamma = \frac{\epsilon_0 + \delta_0 + b}{1 + b}. \tag{52}$$

Because $\delta(b) \to 1$ for highly autonomous clients, the spectral gap strictly approaches unity ($\gamma \to 1$). Recall that the mixing time is bounded by the inverse of the spectral gap ($\tau_{\text{mix}} \propto 1/\gamma$). Because the gap approaches 1, the Markov chain mixes almost instantly. Thus, the mathematical stability bottleneck ($\tau_{\text{mix}} < T$) that causes divergence in Cases 1 and 2 is easily satisfied here even for high autonomy parameter.

Instead, the convergence bottleneck in this case is the second condition: a strictly positive expected participation probability ($\rho_\pi > 0$). If the system stabilizes quickly, but stabilizes into a state of total starvation, the global optimization process still collapses.

We want to enforce $\mathbb{E}[|\mathcal{S}|] \geq 1$. The total expected active subset size at stationarity is the sum of their individual probabilities: $\mathbb{E}[|\mathcal{S}|] = N_c p^\star(b)$, where the probability for a single independent client to participate is the ratio of its entry rate to its total transition rate:

$$p^\star(b) = \frac{\epsilon(b)}{\epsilon(b) + \delta(b)} = \frac{\epsilon_0}{\epsilon_0 + \delta_0 + b}. \tag{53}$$

This leads to the following condition:

$$\frac{N_c \epsilon_0}{\epsilon_0 + \delta_0 + b} \geq 1. \tag{54}$$

Rearranging this inequality to isolate the autonomy parameter b yields the critical threshold:

$$b_{\text{critical}} = N_c\epsilon_0 - (\epsilon_0 + \delta_0). \tag{55}$$

If the client autonomy parameter b exceeds this linear threshold, the stationary distribution collapses to the empty set. The expected participation probability drops to zero ($\rho_\pi \to 0$), completely halting the distributed learning process. □

C Server Policy Optimization Solution

To solve the optimization problem, we first observe the monotonicity of the objective function. The global error $E(N_u, w)$ strictly decreases as the participation probability ρ_π and the effective sample size $m_{\text{eff},\pi}$ increase, and as the mixing time τ_{mix} decreases. Because higher individual rewards w strictly increase the participation probability $p_u(w)$, the objective function is monotonically decreasing with respect to w.

Consequently, the rational server will entirely exhaust its available resources to minimize the error, meaning the inequality budget constraint is always tight (active) at the optimum:

$$N_u \cdot w = B \implies w = \frac{B}{N_u} \tag{56}$$

By substituting the tight budget constraint into the participation probability, we eliminate the continuous variable w:

$$p_u\left(\frac{B}{N_u}\right) = \frac{B/N_u}{1 + B/N_u} = \frac{B}{N_u + B} \tag{57}$$

This substitution formalizes the trade-off between the expected number of participating uniform clients:

$$N_u p_u = \frac{N_u B}{N_u + B}, \tag{58}$$

against the probability that no uniform client is participating:

$$(1 - p_u)^{N_u} = \left(1 - \frac{B}{N_u + B}\right)^{N_u} = \left(\frac{N_u}{N_u + B}\right)^{N_u} = \left(1 + \frac{B}{N_u}\right)^{-N_u}. \tag{59}$$

By projecting the continuous variable out of the system, we successfully reduce the complex two-dimensional optimization space into a 1-dimensional discrete optimization problem solely over N_u:

$$N_u^\star = \underset{N_u \in [1, N_c - N_d]}{\arg\min} \; E\left(N_u, \frac{B}{N_u}\right) \tag{60}$$

Algorithm 1: Optimal Budget Allocation via 1D Grid Search

Input: Total budget B, Total number of clients N_c, Number of dynamic clients N_d, Constants $(\eta, \sigma^2, L, T, c_0)$, Mode `Exact` or `Approx`

Output: Optimal number of uniform clients $N_u^\star$, Optimal individual reward $w^\star$

Initialize minimum error $E_{\min} \leftarrow \infty$;
Initialize $N_u^\star \leftarrow 0$, $w^\star \leftarrow 0$;
for $N_u \leftarrow 1$ **to** $N_c - N_d$ **do**
  `// Assume the entire budget is actively distributed`
  $w \leftarrow \frac{B}{N_u}$;
  $p_u \leftarrow \frac{B}{N_u + B}$;
  `// Compute Stationary Metrics based on selected Mode`
  if *Mode* $==$ `Approx` **then**
    Calculate stationary participation probability $\rho_\pi(N_u, p_u)$;
    Calculate effective sample size $m_{\text{eff},\pi}(N_u, p_u)$;
    Calculate mixing time bound $\tau_{\text{mix}}(N_u, p_u)$;
  else
    Construct transition matrix $Q_d \in \mathbb{R}^{(N_d+1)\times(N_d+1)}$ using exact Binomial convolutions;
    Find exact stationary distribution π_d via power iteration ($\pi_d \leftarrow \pi_d Q_d$);
    $\rho_\pi \leftarrow 1 - \pi_d(0) \cdot (1 - p_u)^{N_u}$;
    $\frac{1}{m_{\text{eff},\pi}} \leftarrow \frac{1}{\rho_\pi} \sum_{y=0}^{N_d} \sum_{u=0}^{N_u} \pi_d(y) \binom{N_u}{u} p_u^u (1-p_u)^{N_u - u} \left(\frac{1}{y+u}\right) \mathbf{1}_{\text{y+u>0}}$;
    Simulate exact τ_{mix} via Total Variation distance threshold;
  `// Evaluate the Global Convergence Error Objective`
  $E(N_u) \leftarrow \frac{\|\theta^{(0)} - \theta^\star\|^2}{\eta \rho_\pi T} + \frac{\eta \sigma^2}{\rho_\pi m_{\text{eff},\pi}} + \frac{c_0 \tau_{\text{mix}}}{T \rho_\pi} \left(\eta \sigma^2 + \frac{L}{2} \|\theta^{(0)} - \theta^\star\|^2\right)$;
  `// Update global optimum`
  if $E(N_u) < E_{\min}$ **then**
    $E_{\min} \leftarrow E(N_u)$;
    $N_u^\star \leftarrow N_u$;
    $w^\star \leftarrow w$;
return $N_u^\star, w^\star$

D Extension to Heterogeneous Data

Building on the concepts presented in the main text, this section provides the full technical details of our system under heterogeneous data distributions. Here, we extend our single-server baseline to account for heterogeneous data distributions and analyze how biased, autonomous client participation impacts global convergence.

D.1 Problem Formulation

To model data heterogeneity, we divide the client population of size N_c into two disjoint groups, G_1 and G_2. Clients in group G_1 sample their local data independent and identically distributed (i.i.d.) from distribution $\mathcal{D}_1$, while clients in group G_2 sample from distribution $\mathcal{D}_2$. This induces two distinct local objective functions, $F_1(\theta)$ and $F_2(\theta)$. Without loss of generality, we assume the server's primary goal is to minimize the expected risk of the first group, $F_1(\theta)$, which represents the target objective according to the population balance between the two groups and the initialization state of the server.

At each round r, let $m_1^{(r)}$ and $m_2^{(r)}$ denote the number of participating clients from G_1 and G_2, respectively. The total active subset size is $m^{(r)} = m_1^{(r)} + m_2^{(r)}$. Because client participation is state-dependent and autonomous, the server cannot guarantee a balanced dataset. Given that a round

is successful ($m^{(r)} > 0$), we define the expected target participation ratio as:

$$\alpha^{(r)} = \mathbb{E}\left[\frac{m_1^{(r)}}{m^{(r)}} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]. \tag{61}$$

D.2 Assumptions and Aggregated Gradient Properties

To establish convergence guarantees in this heterogeneous environment, we adapt our previous assumptions to account for group-specific distributions and bound the degree of data divergence.

Assumption 4 (Heterogeneous Unbiased Gradients)**.** *The stochastic gradients computed by the clients are unbiased estimators only for their respective group's objective function:*

$$\mathbb{E}_{\xi_i^{(r)} \sim \mathcal{D}_j}\left[g_i(\theta^{(r)}, \xi_i^{(r)})\right] = \nabla F_j(\theta^{(r)}), \quad \forall i \in G_j,\ j \in \{1, 2\}. \tag{62}$$

Assumption 5 (Bounded Group Variance)**.** *The variance of the stochastic gradients computed by clients within each group is uniformly bounded by a group-specific constant $\sigma_j^2 > 0$:*

$$\mathbb{E}_{\xi_i^{(r)} \sim \mathcal{D}_j}\left[\left\|g_i(\theta^{(r)}, \xi_i^{(r)}) - \nabla F_j(\theta^{(r)})\right\|^2\right] \leq \sigma_j^2, \quad \forall i \in G_j. \tag{63}$$

Assumption 6 (Bounded Gradient Divergence)**.** *The heterogeneity between the two groups is strictly bounded. The squared L_2-norm of the difference between the two group objectives (the gradient bias) is uniformly bounded by a constant $\zeta_{1,2} > 0$:*

$$\|\nabla F_2(\theta^{(r)}) - \nabla F_1(\theta^{(r)})\|^2 \leq \zeta_{1,2}. \tag{64}$$

Based on these assumptions, we define the statistical properties of the aggregated gradient estimator $\widehat{g}^{(r)}$ under the effects of heterogeneous participation.

Lemma 2 (Properties of the Heterogeneous Aggregated Gradient)**.** *Given that the active set is non-empty ($m^{(r)} > 0$), the aggregated gradient estimator is a biased convex combination of the two group gradients, satisfying:*

$$\mathbb{E}\left[\widehat{g}^{(r)} \mid \theta^{(r)}, m^{(r)} > 0\right] = \alpha^{(r)} \nabla F_1(\theta^{(r)}) + (1 - \alpha^{(r)}) \nabla F_2(\theta^{(r)}). \tag{65}$$

and its variance is bounded by the weighted sum of the local group variances:

$$\mathbb{E}\left[\|\widehat{g}^{(r)} - \mathbb{E}[\widehat{g}^{(r)}]\|^2 \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \leq V_1^{(r)} \sigma_1^2 + V_2^{(r)} \sigma_2^2 + \sigma_\alpha^{(r)2} \zeta_{1,2}, \tag{66}$$

where the variance weights are defined as $V_1^{(r)} := \mathbb{E}\left[\frac{m_1^{(r)}}{(m^{(r)})^2} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]$ and $V_2^{(r)} := \mathbb{E}\left[\frac{m_2^{(r)}}{(m^{(r)})^2} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]$, and the variance of the participation ratio is $\sigma_\alpha^{(r)2} := \mathrm{Var}\left(\frac{m_1^{(r)}}{m^{(r)}} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right)$.

Proof. Because clients sample from different distributions, we first evaluate the expectation of the aggregated update using the law of total expectation, conditional on a successful round:

$$\begin{aligned}
&\mathbb{E}\left[\widehat{g}^{(r)} \mid \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\frac{1}{m^{(r)}} \sum_{i \in G_1^{(r)}} \mathbb{E}[g_i] + \frac{1}{m^{(r)}} \sum_{j \in G_2^{(r)}} \mathbb{E}[g_j] \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\frac{m_1^{(r)}}{m^{(r)}} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \nabla F_1(\theta^{(r)}) + \mathbb{E}\left[\frac{m_2^{(r)}}{m^{(r)}} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \nabla F_2(\theta^{(r)}) \\
&= \alpha^{(r)} \nabla F_1(\theta^{(r)}) + (1 - \alpha^{(r)}) \nabla F_2(\theta^{(r)}).
\end{aligned}$$

Next, we bound the variance of this estimator. We apply the law of total variance to decompose the total error into two distinct components: the noise from the stochastic gradient sampling given

a fixed set of active clients, and the noise caused by the varying ratio of clients drawn from each heterogeneous group. Because the cross-terms vanish in expectation, we get:

$$
\begin{aligned}
&\mathbb{E}\left[\|\widehat{g}^{(r)} - \mathbb{E}[\widehat{g}^{(r)}]\|^2 \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\left\|\left(\widehat{g}^{(r)} - \mathbb{E}[\widehat{g}^{(r)} \mid G_1^{(r)}, G_2^{(r)}]\right) + \left(\mathbb{E}[\widehat{g}^{(r)} \mid G_1^{(r)}, G_2^{(r)}] - \mathbb{E}[\widehat{g}^{(r)}]\right)\right\|^2 \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&= \underbrace{\mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{g}^{(r)} - \mathbb{E}[\widehat{g}^{(r)} \mid G_1^{(r)}, G_2^{(r)}]\right\|^2 \;\middle|\; \theta^{(r)}, G_1^{(r)}, G_2^{(r)}\right] \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]}_{\text{(a)}} \\
&\quad + \underbrace{\mathbb{E}\left[\left\|\mathbb{E}[\widehat{g}^{(r)} \mid G_1^{(r)}, G_2^{(r)}] - \mathbb{E}[\widehat{g}^{(r)}]\right\|^2 \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]}_{\text{(b)}}.
\end{aligned}
$$

Assuming the stochastic gradients g_i and g_j are independent across different clients, the expectations of the cross-products vanish. We can then upper bound the variance of each local gradient by the group variances σ_1^2 and σ_2^2:

$$
\begin{aligned}
\text{(a)} &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{m^{(r)}}\left(\sum_{i\in G_1^{(r)}}\left(g_i(\theta^{(r)}, \xi_i^{(r)}) - \nabla F_1(\theta^{(r)})\right)\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. + \sum_{j\in G_2^{(r)}}\left(g_j(\theta^{(r)}, \xi_j^{(r)}) - \nabla F_2(\theta^{(r)})\right)\right)\right\|^2 \;\middle|\; \theta^{(r)}, G_1^{(r)}, G_2^{(r)}\right] \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\frac{1}{(m^{(r)})^2}\left(\sum_{i\in G_1^{(r)}}\mathbb{E}\left[\|g_i(\theta^{(r)}, \xi_i^{(r)}) - \nabla F_1(\theta^{(r)})\|^2 \mid \theta^{(r)}, G_1^{(r)}, G_2^{(r)}\right]\right.\right. \\
&\qquad\qquad \left.\left. + \sum_{j\in G_2^{(r)}}\mathbb{E}\left[\|g_j(\theta^{(r)}, \xi_j^{(r)}) - \nabla F_2(\theta^{(r)})\|^2 \mid \theta^{(r)}, G_1^{(r)}, G_2^{(r)}\right]\right) \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&\le \mathbb{E}\left[\frac{1}{(m^{(r)})^2}\left(\sum_{i\in G_1^{(r)}}\sigma_1^2 + \sum_{j\in G_2^{(r)}}\sigma_2^2\right) \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\frac{m_1^{(r)}}{(m^{(r)})^2} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]\sigma_1^2 + \mathbb{E}\left[\frac{m_2^{(r)}}{(m^{(r)})^2} \;\middle|\; \theta^{(r)}, m^{(r)} > 0\right]\sigma_2^2 := V_1^{(r)}\sigma_1^2 + V_2^{(r)}\sigma_2^2.
\end{aligned}
$$

Because the total number of active clients is $m^{(r)} = m_1^{(r)} + m_2^{(r)}$, we can express the group 2 ratio as $\frac{m_2^{(r)}}{m^{(r)}} = 1 - \frac{m_1^{(r)}}{m^{(r)}}$. Similarly, $(1 - \alpha^{(r)})$ substitutes the expected ratio for group 2. Substituting these relationships allows us to factor out the difference between the two group gradients:

$$
\begin{aligned}
\text{(b)} &= \mathbb{E}\left[\left\|\left(\frac{m_1^{(r)}}{m^{(r)}}\nabla F_1(\theta^{(r)}) + \frac{m_2^{(r)}}{m^{(r)}}\nabla F_2(\theta^{(r)})\right) - \left(\alpha^{(r)}\nabla F_1(\theta^{(r)}) + (1-\alpha^{(r)})\nabla F_2(\theta^{(r)})\right)\right\|^2 \,\middle|\, \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\left\|\left(\frac{m_1^{(r)}}{m^{(r)}} - \alpha^{(r)}\right)\left(\nabla F_1(\theta^{(r)}) - \nabla F_2(\theta^{(r)})\right)\right\|^2 \,\middle|\, \theta^{(r)}, m^{(r)} > 0\right] \\
&= \mathbb{E}\left[\left(\frac{m_1^{(r)}}{m^{(r)}} - \alpha^{(r)}\right)^2 \,\middle|\, \theta^{(r)}, m^{(r)} > 0\right] \|\nabla F_1(\theta^{(r)}) - \nabla F_2(\theta^{(r)})\|^2 \\
&= \mathrm{Var}\left(\frac{m_1^{(r)}}{m^{(r)}} \,\middle|\, \theta^{(r)}, m^{(r)} > 0\right) \|\nabla F_1(\theta^{(r)}) - \nabla F_2(\theta^{(r)})\|^2 \\
&\leq \sigma_\alpha^{(r)2}\zeta_{1,2}.
\end{aligned}
$$

Adding the bounds for (a) and (b) yields the final inequality.

□

D.3 Convergence Analysis

Analogous to our analysis of the homogeneous setting in Theorem 1, we begin by establishing the single-step expected descent bound toward the target optimal model $\theta^\star$ for the heterogeneous regime.

Theorem 4 (One-Step Descent with Heterogeneous Data)**.** *Under Assumptions 1 and 4,5,6, for any learning rate $\eta > 0$ and any relaxation parameter $\lambda > 0$, the expected squared distance to the target optimal model $\theta^\star$ satisfies:*

$$
\begin{aligned}
\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 \leq \mathbb{E}\left[\left(1 + \frac{\eta\rho^{(r+1)}}{\lambda}\right)\|\theta^{(r)} - \theta^\star\|^2\right] - 2\eta(1-2\eta L)\mathbb{E}\left[\rho^{(r+1)}(F_1(\theta^{(r)}) - F_1(\theta^\star))\right] \\
+ \eta\mathbb{E}\left[\rho^{(r+1)}\left(\lambda(1-\alpha^{(r)})^2\zeta_{1,2} + \eta\left(V_1^{(r)}\sigma_1^2 + V_2^{(r)}\sigma_2^2 + \left(2(1-\alpha^{(r)})^2 + \sigma_\alpha^{(r)2}\right)\zeta_{1,2}\right)\right)\right].
\end{aligned} \tag{67}
$$

Proof. Expanding the expected squared distance to the optimum yields:

$$
\mathbb{E}\|\theta^{(r+1)} - \theta^\star\|^2 = \mathbb{E}\|\theta^{(r)} - \theta^\star\|^2 - 2\eta\mathbb{E}\langle\mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle + \eta^2\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2.
$$

Using the law of total expectation conditioned on the history $\mathcal{F}^{(r)}$ and a successful round ($m^{(r)} > 0$), we apply the expected aggregated gradient derived in Lemma 2 to the inner product term:

$$
\begin{aligned}
&\mathbb{E}\langle\mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \\
&= \mathbb{E}\left[\rho^{(r+1)}\mathbb{E}\left[\langle\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \mid \mathcal{F}^{(r)}, m^{(r)} > 0\right]\right] \\
&= \mathbb{E}\left[\rho^{(r+1)}\langle\alpha^{(r)}\nabla F_1(\theta^{(r)}) + (1-\alpha^{(r)})\nabla F_2(\theta^{(r)}), \theta^{(r)} - \theta^\star\rangle\right] \\
&= \mathbb{E}\left[\rho^{(r+1)}\langle\nabla F_1(\theta^{(r)}) + (1-\alpha^{(r)})(\nabla F_2(\theta^{(r)}) - \nabla F_1(\theta^{(r)})), \theta^{(r)} - \theta^\star\rangle\right].
\end{aligned}
$$

Applying the convexity of the target objective function F_1 (Assumption 1), we can bound the first component of the inner product:

$$
\begin{aligned}
\mathbb{E}\langle\mathbb{I}^{(r+1)}\widehat{g}^{(r)}, \theta^{(r)} - \theta^\star\rangle \geq \mathbb{E}\left[\rho^{(r+1)}(F_1(\theta^{(r)}) - F_1(\theta^\star))\right] \\
+ \mathbb{E}\left[\rho^{(r+1)}\langle(1-\alpha^{(r)})(\nabla F_2(\theta^{(r)}) - \nabla F_1(\theta^{(r)})), \theta^{(r)} - \theta^\star\rangle\right].
\end{aligned}
$$

To decouple the heterogeneity bias from the distance to the optimum, we apply Young's inequality, $-2\langle A, B\rangle \leq \lambda\|A\|^2 + \frac{1}{\lambda}\|B\|^2$, for a strictly positive constant $\lambda > 0$. Using the bounded gradient

divergence $\zeta_{1,2}$ (Assumption 6), we obtain:

$$
\begin{aligned}
&-2\eta\mathbb{E}\left[\rho^{(r+1)}\langle(1-\alpha^{(r)})(\nabla F_2(\theta^{(r)})-\nabla F_1(\theta^{(r)})),\theta^{(r)}-\theta^\star\rangle\right]\\
&\quad\le \eta\lambda\mathbb{E}\left[\rho^{(r+1)}\|(1-\alpha^{(r)})(\nabla F_2(\theta^{(r)})-\nabla F_1(\theta^{(r)}))\|^2\right]+\frac{\eta}{\lambda}\mathbb{E}\left[\rho^{(r+1)}\|\theta^{(r)}-\theta^\star\|^2\right]\\
&\quad\le \eta\lambda\mathbb{E}\left[\rho^{(r+1)}(1-\alpha^{(r)})^2\zeta_{1,2}\right]+\frac{\eta}{\lambda}\mathbb{E}\left[\rho^{(r+1)}\|\theta^{(r)}-\theta^\star\|^2\right].
\end{aligned}
$$

Next, we bound the expected second moment of the effective gradient update. We begin by conditioning on a successful communication round ($m^{(r)}>0$, which occurs with probability $\rho^{(r+1)}$) and applying the standard variance decomposition $\mathbb{E}[\|X\|^2]=\|\mathbb{E}[X]\|^2+\mathrm{Var}(X)$ alongside the properties established in Lemma 2:

$$
\begin{aligned}
&\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2\\
&=\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}^{(r+1)}\|\widehat{g}^{(r)}\|^2\mid\mathcal{F}^{(r)}\right]\right]\\
&=\mathbb{E}\left[\rho^{(r+1)}\mathbb{E}\left[\|\widehat{g}^{(r)}\|^2\mid\mathcal{F}^{(r)},m^{(r)}>0\right]\right]\\
&=\mathbb{E}\left[\rho^{(r+1)}\left(\left\|\mathbb{E}\left[\widehat{g}^{(r)}\mid\mathcal{F}^{(r)},m^{(r)}>0\right]\right\|^2+\mathbb{E}\left[\|\widehat{g}^{(r)}-\mathbb{E}[\widehat{g}^{(r)}]\|^2\mid\mathcal{F}^{(r)},m^{(r)}>0\right]\right)\right]\\
&\le\mathbb{E}\left[\rho^{(r+1)}\left(\|\alpha^{(r)}\nabla F_1(\theta^{(r)})+(1-\alpha^{(r)})\nabla F_2(\theta^{(r)})\|^2+V_1^{(r)}\sigma_1^2+V_2^{(r)}\sigma_2^2+\sigma_\alpha^{(r)2}\zeta_{1,2}\right)\right].
\end{aligned}
$$

To bound the squared norm of the mean gradient, we add and subtract $\nabla F_1(\theta^{(r)})$ inside the norm, and then apply the inequality $\|A+B\|^2\le 2\|A\|^2+2\|B\|^2$:

$$
\begin{aligned}
&\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2\\
&\le\mathbb{E}\left[\rho^{(r+1)}\left(\|\nabla F_1(\theta^{(r)})+(1-\alpha^{(r)})(\nabla F_2(\theta^{(r)})-\nabla F_1(\theta^{(r)}))\|^2+V_1^{(r)}\sigma_1^2+V_2^{(r)}\sigma_2^2+\sigma_\alpha^{(r)2}\zeta_{1,2}\right)\right]\\
&\le\mathbb{E}\left[\rho^{(r+1)}\left(2\|\nabla F_1(\theta^{(r)})\|^2+2(1-\alpha^{(r)})^2\|\nabla F_2(\theta^{(r)})-\nabla F_1(\theta^{(r)})\|^2+V_1^{(r)}\sigma_1^2+V_2^{(r)}\sigma_2^2+\sigma_\alpha^{(r)2}\zeta_{1,2}\right)\right].
\end{aligned}
$$

Next, we apply the L-smoothness of F_1, which guarantees $\|\nabla F_1(\theta^{(r)})\|^2\le 2L(F_1(\theta^{(r)})-F_1(\theta^\star))$, along with Assumption 6 for the bounded gradient divergence ($\zeta_{1,2}$). Substituting these bounds yields the complete second moment bound:

$$
\mathbb{E}\|\mathbb{I}^{(r+1)}\widehat{g}^{(r)}\|^2\le\mathbb{E}\left[\rho^{(r+1)}\left(4L(F_1(\theta^{(r)})-F_1(\theta^\star))+\left(2(1-\alpha^{(r)})^2+\sigma_\alpha^{(r)2}\right)\zeta_{1,2}+V_1^{(r)}\sigma_1^2+V_2^{(r)}\sigma_2^2\right)\right].
$$

Finally, substituting the decoupled inner product bound and the second moment bound back into the fundamental distance expansion, and grouping terms by their respective coefficients, completes the proof. □

Theorem 5 (Static Heterogeneous Convergence Bound)**.** *Assume the expected squared distance to the optimum is bounded by D^2, such that $\mathbb{E}\|\theta^{(r)}-\theta^\star\|^2\le D^2$ for all $r\ge 0$. For a learning rate $\eta\le\frac{1}{4L}$ and a free parameter $\lambda>0$, the expected suboptimality of the averaged model $\bar{\theta}^{(T)}=\frac{1}{T}\sum_{r=0}^{T-1}\theta^{(r)}$ with respect to the target objective F_1 after T rounds is bounded by:*

$$
\mathbb{E}[F_1(\bar{\theta}^{(T)})-F_1(\theta^\star)]\le\frac{D^2}{\eta\rho T}+\frac{D^2}{\lambda}+\lambda(1-\alpha)^2\zeta_{1,2}+\eta\left(V_1\sigma_1^2+V_2\sigma_2^2+\left(2(1-\alpha)^2+\sigma_\alpha^2\right)\zeta_{1,2}\right). \tag{68}
$$

Proof. We begin by rearranging the general one-step descent bound for heterogeneous data (Theorem 4) to isolate the objective suboptimality gap on the left-hand side:

$$
\begin{aligned}
&2\eta(1-2\eta L)\mathbb{E}\left[\rho^{(r+1)}(F_1(\theta^{(r)})-F_1(\theta^\star))\right] \\
&\leq \mathbb{E}\|\theta^{(r)}-\theta^\star\|^2-\mathbb{E}\|\theta^{(r+1)}-\theta^\star\|^2+\frac{\eta\rho^{(r+1)}}{\lambda}\mathbb{E}\|\theta^{(r)}-\theta^\star\|^2 \\
&+\eta\mathbb{E}\left[\rho^{(r+1)}\left(\lambda(1-\alpha^{(r)})^2\zeta_{1,2}+\eta\left(V_1^{(r)}\sigma_1^2+V_2^{(r)}\sigma_2^2+\left(2(1-\alpha^{(r)})^2+\sigma_\alpha^{(r)2}\right)\zeta_{1,2}\right)\right)\right].
\end{aligned}
$$

Because the client participation probabilities are time-invariant in the static regime, the subset statistics are identical at every round. We can substitute the round-dependent variables directly with their static constants: $\rho^{(r+1)}=\rho$, $\alpha^{(r)}=\alpha$, $V_1^{(r)}=V_1$, $V_2^{(r)}=V_2$ and $\sigma_\alpha^{(r)2}=\sigma_\alpha^2$.

Setting $\eta\leq\frac{1}{4L}$ ensures that $2(1-2\eta L)\geq 1$, which allows us to bound the expected descent term coefficient as $2\eta(1-2\eta L)\geq\eta$. We also enforce the bounded domain assumption $\mathbb{E}\|\theta^{(r)}-\theta^\star\|^2\leq D^2$ to control the distance expansion term.

Substituting these constants and bounds into the inequality yields:

$$
\begin{aligned}
\eta\rho\mathbb{E}\left[F_1(\theta^{(r)})-F_1(\theta^\star)\right]&\leq\mathbb{E}\|\theta^{(r)}-\theta^\star\|^2-\mathbb{E}\|\theta^{(r+1)}-\theta^\star\|^2 \\
&+\frac{\eta\rho}{\lambda}D^2+\eta\rho\left(\lambda(1-\alpha)^2\zeta_{1,2}+\eta\left(V_1\sigma_1^2+V_2\sigma_2^2+\left(2(1-\alpha)^2+\sigma_\alpha^2\right)\zeta_{1,2}\right)\right).
\end{aligned}
$$

Summing this inequality over T rounds from $r=0$ to $T-1$, the first two terms on the right-hand side form a perfect telescoping sum, which is bounded by the initial distance $\mathbb{E}\|\theta^{(0)}-\theta^\star\|^2-\mathbb{E}\|\theta^{(T)}-\theta^\star\|^2\leq D^2$. The remaining constant terms scale linearly with T:

$$
\begin{aligned}
\eta\rho\sum_{r=0}^{T-1}\mathbb{E}[F_1(\theta^{(r)})-F_1(\theta^\star)]&\leq D^2+\frac{\eta\rho T}{\lambda}D^2 \\
&+\eta\rho T\left(\lambda(1-\alpha)^2\zeta_{1,2}+\eta\left(V_1\sigma_1^2+V_2\sigma_2^2+\left(2(1-\alpha)^2+\sigma_\alpha^2\right)\zeta_{1,2}\right)\right).
\end{aligned}
$$

Dividing the entire inequality by $\eta\rho T$ and applying Jensen's Inequality to the convex function F_1 for the averaged model $\bar{\theta}^{(T)}=\frac{1}{T}\sum_{r=0}^{T-1}\theta^{(r)}$ yields the final bound:

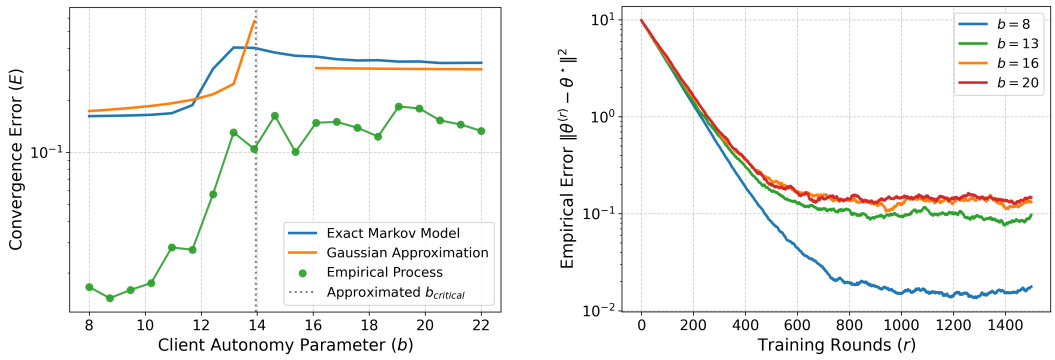
$$
\mathbb{E}[F_1(\bar{\theta}^{(T)})-F_1(\theta^\star)]\leq\frac{D^2}{\eta\rho T}+\frac{D^2}{\lambda}+\lambda(1-\alpha)^2\zeta_{1,2}+\eta\left(V_1\sigma_1^2+V_2\sigma_2^2+\left(2(1-\alpha)^2+\sigma_\alpha^2\right)\zeta_{1,2}\right).
$$

□

Theorem 5 shows that data heterogeneity leads to an additional bias term in the error bound, which is proportional to $(1-\alpha)^2\zeta_{1,2}$. As $\alpha\to 1$, the target group G_1 participates more, which reduces this bias. To extend this steady-state analysis to the fully dynamic Markovian regime, as we did in Section 3.3 for the homogeneous case, we must take into account the dependency of the client's participation probability on the global model's divergence from its local data. This creates a feedback loop. As the global model begins to optimize toward the target objective F_1, the local losses and gradient norms for clients in G_2 increase. Under an autonomous participation policy, this divergence causes G_2 clients a lower participation probability. As a result, $\alpha^{(r)}$ skews even further toward 1, which in turn accelerates the bias of the global model toward F_1. To understand these convergence dynamics, we will observe them empirically through the numerical simulations that can be found in Appendix E.3.

E Numerical Results and Simulations

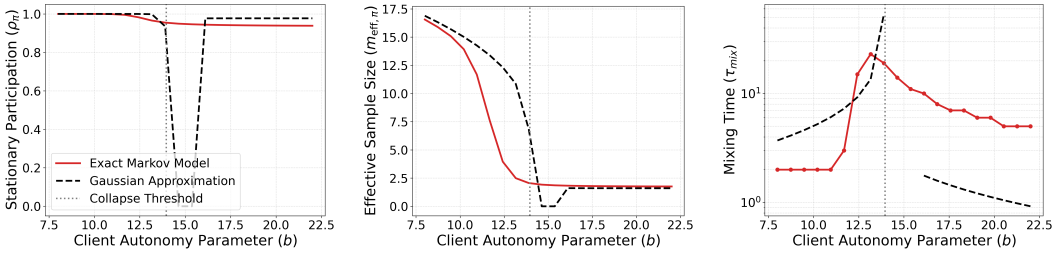
In this appendix, we present numerical experiments to illustrate our theoretical results. The following experiments are based on the autonomous client participation rule described in Case 1, which model a system where the total of N_c clients are divided into a dynamic set of autonomous clients and a uniform set of clients with a fixed probability to participate.



(a) Global convergence error at steady-state.

(b) Empirical training error over time.

Figure 3: Convergence behavior of the system comparing the theoretical bounds to the empirical SGD process.



(a) Participation Rate

(b) Effective Sample Size

(c) Mixing Time

Figure 4: Stationary metrics of the distributed learning process under Case 1 participation as a function of the client autonomy parameter b.

E.1 Homogeneous Data Simulation

First, we evaluate the homogeneous data model established in Section 3. Figure 3 illustrates the empirical learning processes for various client autonomy parameters b around the critical autonomy threshold. For our specific simulation parameters, Proposition 3 dictates a critical threshold of $b_{\text{critical}} \approx 14$.

When $b < b_{\text{critical}}$, we can see that the global model achieves a low steady-state convergence error. In this regime, a large proportion of autonomous clients actively participate, which maximizes the effective sample size m_{eff} and minimizes the Markov chain's mixing time, aligning with the theoretical bounds of Theorem 2.

Conversely, when $b > b_{\text{critical}}$, the convergence error stays constant at a higher level. This indicates that the N_d autonomous clients stopped participating, forcing the learning process to rely only on the updates from the N_u uniform clients. This behavior can be understood better looking at the stationary parameter metrics in Figure 1. As b approaches the critical threshold, both the stationary participation rate (ρ_π) and the effective sample size (m_{eff}) decrease. Simultaneously, the mixing time (τ_{mix}) experiences a massive spike demonstrating the system's struggle to reach equilibrium.

In Figure 3b, we validate these empirical findings against our theoretical convergence bounds. Notably, the Gaussian approximation closely tracks the exact Markov model in both the stable and collapsed regimes. Near the critical transition ($14 \leq b \leq 16$), the Gaussian approximation curve is not calculated. This gap occurs because the Linear Noise Approximation requires the system's contraction factor $\mu' < 1$. During the transition, the fixed point becomes unstable and $\mu' \geq 1$, rendering the asymptotic variance undefined. However, as b increases past 17, the autonomous clients completely stop to participate, and the system settles into a new stable equilibrium driven entirely by the uniform clients.

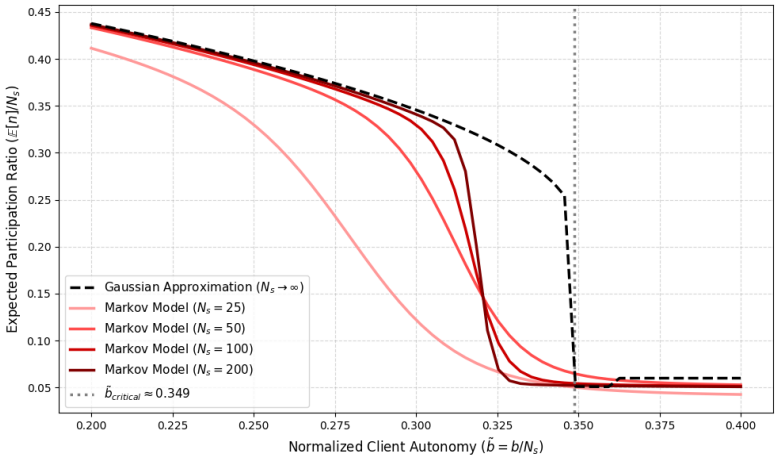


Figure 5: Finite-size scaling in Case 1. Increasing the population size N_d drives the exact Markov models to converge with the continuous mean-field limit, demonstrating a sharp collapse at the critical threshold $\tilde{b}_{\text{critical}}$.

Previous experiments showed results for $N_c = 50$ clients. In Figure 5, we compare different N_d values to the Gaussian approximation we used in Proposition 1. To calculate the expected number of participating clients, we built the exact transition matrix (Q) and compute the stationary distribution π via power iteration ($\pi^{(k+1)} = \pi^{(k)}Q$) for up to 20,000 steps until numerical convergence ($\|\pi^{(k+1)} - \pi^{(k)}\|_\infty < 10^{-12}$), capturing the long-term steady state of the Markov chain. Finally, we compute the expected number of active clients $\mathbb{E}[n] = \sum(n \times \pi_n)$ and divide it by N_d to plot the normalized expected participation ratio. As we can see, the empirical results increasingly align with the continuous Gaussian approximation as the dynamic client population N_d increases.

E.2 Empirical Validation of the Proposed Server Policy

In Section 4, we introduced an optimization problem where the server should decide how to distribute his budget B among the clients, in order to minimize the global steady-state convergence error, as defined in equation 19.

Figure 6 illustrates the empirical solution of this optimization problem. The server chooses whether to distribute its budget B across a large pool of clients, which yields low individual rewards and correspondingly low participation probabilities, or, alternatively, to concentrate the budget on fewer clients with higher rewards to strictly guarantee their participation. As observed in the graph, choosing too few uniform clients results in a high global convergence error. This occurs because the limited number of guaranteed participants reduces the effective sample size m_{eff} and thereby causes higher variance in the global gradient estimator.

Conversely, increasing the number of queried uniform clients N_u mitigates this variance but makes the mixing time of the system longer. Since each client reward w is smaller, the uniform set of clients has a smaller participation probability, making the Markov chain take longer to reach the stationary distribution. Furthermore, because the mixing time τ_{mix} scales as a discrete step function, these delays appear as spikes in the overall global error.

E.3 Heterogeneous Data Simulation

To analyze the impact of data heterogeneity, we introduce a new participation rule where autonomous client behavior depends on local gradient norms. To motivate this approach, note that while traditional methods (Zhao & Zhang, 2015) use importance sampling to preferentially select clients with large gradient norms to accelerate global convergence, our autonomous clients exhibit the opposite behavior. They are reluctant to share their data when their local gradient is exceptionally large, as it indicates their data diverges significantly from the current global model. We

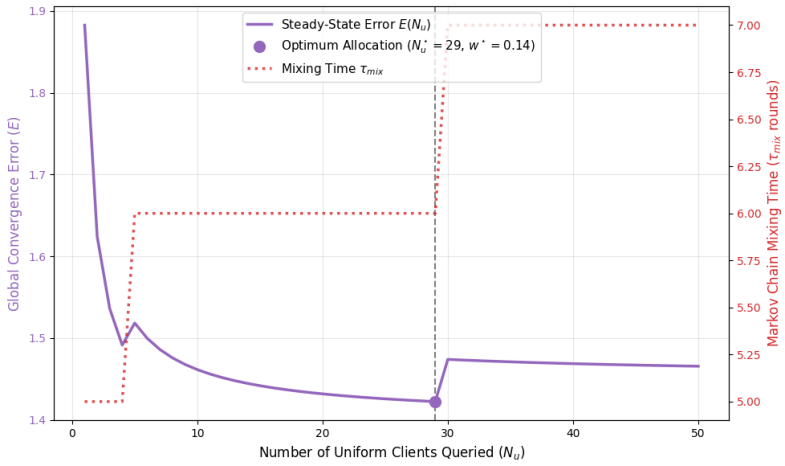


Figure 6: Server budget optimization. The global error minimizes at $N_u^\star = 29$, optimally balancing the number of participating clients against discrete penalties in the mixing time (τ_{mix}).

formalize this self-interested behavior by modeling the participation probability for client i via a logistic function parameterized by their relative gradient norm:

$$p_i^{(r+1)} = \frac{1}{1 + e^{b_i c_i^{(r)}}}, \qquad \text{where } c_i^{(r)} = \frac{\|g_i(\theta^{(r-1)}, \xi_i^{(r-1)})\|}{\frac{1}{|\mathcal{S}^{(r)}|}\sum_{k \in \mathcal{S}^{(r)}} \|g_k(\theta^{(r-1)}, \xi_k^{(r-1)})\|} - 1, \tag{69}$$

This rule models the self-interested behavior of clients who preferentially engage when the server's global model aligns closely with their local data. The contribution $c_i^{(r)}$ is evaluated against the mean gradient norm of the active subset from the previous round, rendering it scale-invariant to the absolute number of participants. By shifting the ratio by -1, we establish a zero-centered cost threshold. Thus, if a client's gradient norm is higher than the active mean gradient norm, $c_i^{(r)} > 0$ and we will get a participation probability below 0.5, meaning his data diverge from the server global model.

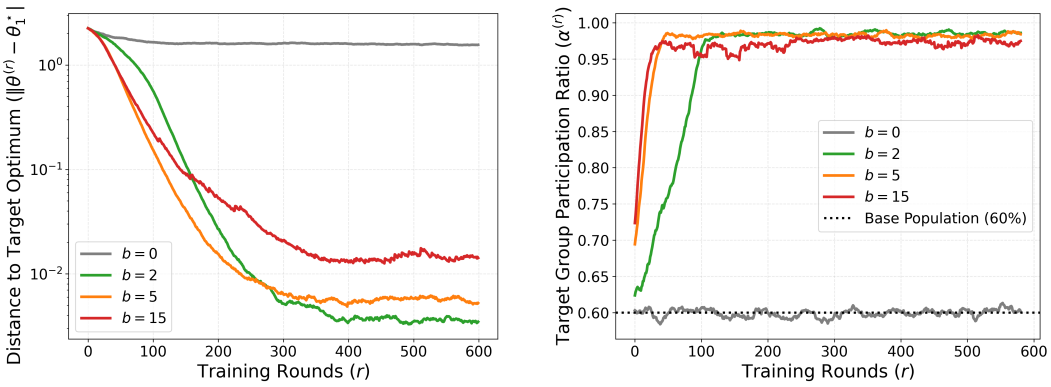
In Figure 2, we evaluate the learning dynamics across different autonomy parameters (b) when the target group G_1 maintains a 60/40 population majority over the biased group G_2. As Figure 2b illustrates, higher values of b accelerate the self-debiasing feedback loop, causing the target group's participation ratio ($\alpha^{(r)}$) to increase much faster.

However, as shown in Figure 2a, forcing α higher and faster does not automatically translate to a lower convergence error. When b is too high, the logistic selection threshold becomes hypersensitive, causing even highly aligned clients from the target group to drop out due to minor stochastic variations in their local gradients. According to Theorem 5, this starvation of the active subset size inflates the gradient variance weight $V_1^{(r)}$, increasing the convergence error through the term $\eta V_1 \sigma_1^2$. When $b = 0$, the system cannot filter out divergent updates, leaving the participation ratio near the base population rate ($\alpha \approx 0.6$). This traps the global model at a high error floor due to the heterogeneity bias penalty, which is proportional to $(1-\alpha)^2 \zeta_{1,2}$.

In Figure 8, we investigate how the server's global model initialization (θ_0) and the demographic group distribution ($\frac{N_1}{N_1+N_2}$) affect convergence.

When the model is initialized near θ_2^*, the local gradient norms for group G_2 are naturally small, yielding a high participation probability for G_2 and a low participation probability for G_1. As established by the convergence bound in Theorem 5, a low α maximizes the heterogeneity bias penalty $(1-\alpha)^2 \zeta_{1,2}$.

To escape this initialization, the expected number of active target clients ($m_1 = N_1 p_1$) must exceed the active biased clients ($m_2 = N_2 p_2$) to flip the gradient direction back toward ∇F_1. Because p_1 is very low, this condition is only met if the target population N_1 is overwhelmingly large, explaining the $> 90\%$ majority requirement. Conversely, under favorable initialization ($\theta_0 \approx \theta_1^*$), the proba-



(a) Distance to Target Optimum

(b) Target Group Participation Ratio

Figure 7: Impact of the autonomy parameter (b) on convergence with a $60/40$ population majority. (a) Static participation ($b = 0$) suffers from a high bias-induced error floor. Autonomous dropout allows the model to minimize the target error, though hyper-sensitive autonomy ($b = 15$) injects stochastic noise that raises the error floor compared to moderate autonomy ($b = 2$). (b) The underlying feedback loop demonstrates how higher autonomy rapidly scales the majority participation ratio ($\alpha \to 1.0$).

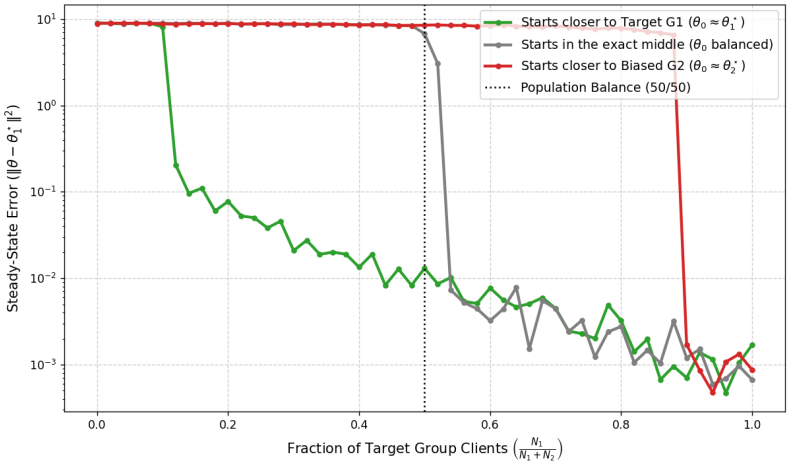


Figure 8: Steady-state convergence error as a function of the target group population fraction, $\frac{N_1}{N_1+N_2}$, evaluated under three distinct initialization states.

bilities are reversed ($p_1 \to 1$ and $p_2 \to 0$). The target group dominates the active subset, keeping $(1-\alpha)^2\zeta_{1,2} \approx 0$ and ensuring stable convergence even if their base population is a small minority.

F Multi-Server Framework

In the main part of the paper, we consider the case of a single server and multiple clients. As we show in this appendix, this setting can be easily extended into a more general framework that describes the dynamics between *multiple* servers competing for the participation of *multiple* clients.

We propose a game with N_c clients and N_s servers. Each client $i \in [N_c] := \{0, 1, \ldots, N_c - 1\}$ has a dataset drawn from a distribution D_i over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^n$ is the feature domain; and $\mathcal{Y} \subset \mathbb{R}$ is the label domain. The clients are interested in utilizing the data of other clients in order to improve their own predictions.

To achieve their goal, they choose a server (or a subset of servers) in each round of the game. Initially, PSs broadcast their models to all clients. The clients then select the servers of their choice

based on their selection rule $\phi_c(k^r, h_i^r, b_i)$. As detailed Section 2, this decision is a function of the common public knowledge k^r, the client's private history h_i^r, and a characteristic parameter b_i that dictates their willingness to explore different servers or to share data.

Each client sends their update to the servers of their choice. An example of an update using stochastic gradient descent (SGD) training proceeds as follows: Each client runs T local iterations in each round. They sample a batch of data for each iteration and perform SGD training on each sampled batch. Assume that the selected server of client i is $\phi_c(k^r, h_i^r, b_i) = j$ in round r. The client i will initiate their local model as $\theta_i^{(r,0)} = \theta_j^{(r)}$ and run T iterations, as:

$$\theta_i^{(r,k+1)} = \theta_i^{(r,k)} - \eta_i g_i(\theta_i^{(r,k)}). \tag{70}$$

Then, they will send the calculated local update to server j:

$$\Delta_i^{(r+1)} = \theta_i^{(r,T)} - \theta_j^{(r)}. \tag{71}$$

In return, each server will send its updated global model after aggregating the updates from all clients that chose it. Server j receives the updates $\{\Delta_i^{(r+1)}\}_{i\in\mathcal{S}_j}$ where $\mathcal{S}_j = \{i \in \{1,\dots,N_c\} : \phi_c(k^r, h_i^r, b_i) = j\}$ and uses an aggregation rule for its model update:

$$\theta_j^{(r+1)} = \theta_j^{(r)} + \mathrm{Agg}(\{\Delta_i^{(r+1)}\}_{i\in\mathcal{S}_j}). \tag{72}$$

In general, the aggregation rule can be an arbitrary function of the update information from the clients in $\mathcal{S}_j$. However, we impose a specific two-step structure on this rule, which is both reasonable and fairly rich. Specifically, we separate the aggregation rule into two steps from the server's point of view. First, each server uses a filtering rule $\phi_s(k^r, h'^r_j, b'_j)$, acting as a counterpart to the clients' selection rule, where h'^r_j is server j's history at round r and b'_j is server j's characteristic. This rule filters the client updates to use only relevant contributions ($\phi_s(k^r, h'^r_j, b'_j)_i$ returns 1 if client i's update is chosen, and 0 otherwise).

Then, the server applies a *fixed* aggregation function on the filtered updates $\{\Delta_i^{(r+1)}\}_{i\in\mathcal{S}'_j}$ where $\mathcal{S}'_j = \{i \in \{1,\dots,N_c\} : \phi_c(k^r, h_i^r, b_i) = j \text{ and } \phi_s(k^r, h'^r_j, b'_j)_i = 1\}$. For example, using FedAvg (McMahan et al., 2017), where n_i is client i's number of data points and $n = \sum_{i=1}^{|\mathcal{S}'_j|} n_i$, we get:

$$\theta_j^{(r+1)} = \theta_j^{(r)} + \sum_{i=1}^{|\mathcal{S}'_j|} \frac{n_i}{n} \Delta_i^{(r+1)}. \tag{73}$$

For visual intuition, Figure 9 illustrates this general framework specifically in the case of $N_s = 2$ servers.

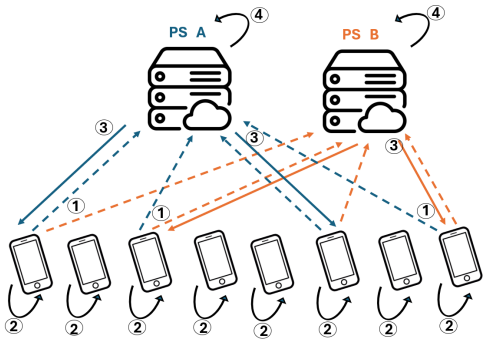


Figure 9: The proposed framework in the case of $N_s = 2$ servers and N_c clients, illustrating the clients' selection rule $\phi_c(k^r, h_i^r, b_i)$ ①, local update calculation ②, servers' filtering rule $\phi_s(k^r, h'^r_j, b'_j)$ ③, and final server aggregation ④.

The generalized process for clients and servers is formalized in Algorithms 2 and 3.

Algorithm 2: Client Update and Server Selection

Input: Round index r, client behavior parameter b_i, number of local steps T, step size η_i

Output: Local update $\Delta_i^{(r+1)}$, selected server $\phi_c(k^r, h_i^r, b_i)$

Receive from each server $j \in [N_s]$: its model $\theta_j^{(r)}$ and the gradients $\|\hat{g}_{k,j}^{(r)}\|$ of its active clients $k \in \mathcal{S}_j$;

Choose a server $j \in [N_s]$ according to: $\phi_c(k^r, h_i^r, b_i) = j,$ with probability $p_{i,j}^r$ where $\sum_{j=1}^{N_s} p_{i,j}^r = 1$;

Set $\theta_i^{(r,0)} = \theta_j^{(r)}$ where $j = \phi_c(k^r, h_i^r, b_i)$;

Run T local iterations of SGD: $\theta_i^{(r,k+1)} \leftarrow \theta_i^{(r,k)} - \eta_i g_i(\theta_i^{(r,k)})$;

Set $\Delta_i^{(r+1)} = \theta_i^{(r,T)} - \theta_j^{(r)}$;

return $\Delta_i^{(r+1)}$ *to selected server* j

Algorithm 3: Server Aggregation

Input: Number of rounds R, a set of clients $[N_c]$

Output: Global model $\theta_j^{(R)}$

Init $\theta_j^{(0)}$;

for *round* $r = 0, 1, \ldots, R-1$ **do**

  Broadcast $\theta_j^{(r)}$ to all clients;

  Receive a set of updates $\{\Delta_i^{(r+1)}\}_{i \in \mathcal{S}_j}$ where $\mathcal{S}_j = \{i \in \{1, \ldots, N_c\} : \phi_c(k^r, h_i^r, b_i) = j\}$;

  Filter the updates according to $\phi_s(k^r, h'^r_j, b'_j)$ to form $\mathcal{S}'_j$;

  Update model: $\theta_j^{(r+1)} \leftarrow \theta_j^{(r)} + \mathrm{Agg}(\{\Delta_i^{(r+1)}\}_{i \in \mathcal{S}'_j})$

return $\theta_j^{(R)}$

While this multi-server framework provides a realistic model of competitive distributed networks, extending our Markovian convergence analysis to this setting introduces some theoretical challenges. In the single-server environment analyzed in Section 3, the system state is described by a single active subset $\mathcal{S}^{(r)}$, which we modeled using a time-invariant transition matrix to bound the mixing time τ_{mix}. However, in the multi-server setting, the state space expands to the joint distribution of client allocations across all N_s servers. Because the clients' selection rule ϕ_c depends on the competing models of *all* servers, the transition probabilities become coupled and non-stationary. A client's decision to participate in server j's learning process depends on the relative divergence of server j's model compared to server k's model. Consequently, establishing the existence of a unique stationary distribution π, bounding the spectral gap, and proving convergence requires analyzing multi-agent dynamical systems with non-convex, interacting drift functions. We leave the derivation of convergence bounds for this coupled, competitive Markov chain as a compelling direction for future work.