CUMULATIVE UTILITY PARITY: FAIR LEARNING UNDER INTERMITTENT CLIENT PARTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermittent and heterogeneous client availability creates unequal opportunities to contribute to federated learning, a disparity that snapshot metrics for loss and accuracy do not capture and completely miss this unfairness. We introduce cumulative utility parity (CUP), which evaluates each client's accumulated training utility relative to its observed availability. CUP prioritizes clients with larger availability-normalized utility deficits and uses cached model updates as surrogates for unavailable clients. We characterize availability-normalized utility disparity, derive a max–min allocation target, and bound the discrepancy between cached surrogates and fresh updates. Across non-IID CIFAR-10 experiments with two and five labels per client, CUP without caching achieves lower mean temporal utility dispersion than all evaluated baselines in the two-label setting and than FedProx, q-FFL, AFL, and FairFedCS in the five-label setting. The cached CUP variants achieve the lowest mean CV and highest mean Jain index in both settings, together with the highest mean final accuracy in the five-label setting. These findings highlight the potential of availability-aware utility allocation to improve cumulative fairness alongside predictive performance. Anonymous code

1 INTRODUCTION

Federated learning (FL) trains a shared model by aggregating locally computed updates while leaving training data on clients (McMahan et al., 2017). Its deployment involves heterogeneous data, computational resources, and connectivity, making both statistical and system constraints central to algorithm design (Bonawitz et al., 2019; Li et al., 2020a). These differences also motivate fairness objectives that go beyond aggregate predictive performance (Li et al., 2020b; Mohri et al., 2019).

Prior work addresses several distinct fairness dimensions. Agnostic Federated Learning optimizes against mixtures of client distributions to account for uncertainty about the target population (Mohri et al., 2019), while q-FFL encourages a more uniform accuracy distribution across devices (Li et al., 2020b). CoreFed uses a cooperative-game notion of core stability, with guarantees under its stated assumptions (Chaudhury et al., 2022). Federated Fairness Analytics provides a framework for measuring fairness across different perspectives (Dilley et al., 2024). At the algorithmic level, FairFedCS uses Lyapunov optimization to adjust selection probabilities based on reputation, participation history, and contribution (Shi et al., 2023); FedFV mitigates conflicts among client gradients before averaging by modifying their directions and magnitudes (Wang et al., 2021).

Participation dynamics introduce a related but distinct challenge. Production FL systems must accommodate intermittently available devices (Bonawitz et al., 2019), and resource-aware selection methods such as FedCS explicitly account for heterogeneous computation and communication constraints (Nishio & Yonetani, 2019). Power-of-Choice analyzes the tradeoff between convergence speed and solution bias under biased client selection (Cho et al., 2020). More generally, Wang & Ji (2022) analyze convergence under arbitrary participation patterns. These studies motivate treating the participation process as part of the learning problem rather than assuming that all clients have comparable opportunities to contribute. Our focus is the cumulative benefit obtained per available scheduling opportunity. This differs from optimizing a performance distribution, mitigating gradient conflicts, or characterizing convergence under a given participation process. Historical scheduling objectives are already present in methods such as FairFedCS (Shi et al., 2023); CUP instead makes availability-normalized utility rates the explicit fairness target. Participation-aware performance

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.

fairness is also studied by PHP-FL, which addresses inconsistent participation probabilities in model-heterogeneous FL through dual-end aligned ensemble learning and importance-driven selective parameter updates (Wu et al., 2025). This is complementary to the performance-distribution objective of q-FFL (Li et al., 2020b) and to our accumulated utility relative to observed availability.

To address this limitation, we introduce CUP, a framework that combines availability-aware client selection (prioritizes clients with larger availability-normalized utility deficits) with cached model updates as surrogates for unavailable clients, see Figure 1. CUP tracks each client's accumulated training utility relative to its observed availability and increases scheduling priority when that utility falls behind or available opportunities are repeatedly missed. When clients are unavailable, their cached model updates serve as surrogates in aggregation, with weights that decrease as the updates become stale. Cache reuse preserves historical information without counting it as fresh participation or newly earned utility. Our analysis characterizes long-run availability-normalized utility rates, and derives a max–min common utility-rate target under an expected realized-participation budget. Separately, under surrogate-accuracy assumptions, we bound the error induced by cached surrogates in the aggregate update direction and its effect on a single optimization step (Section 2).

The implementation adapts online through telemetry, availability estimates, utility deficits, and opportunity debt, making CUP applicable when availability and client data are heterogeneous and evolve over time; the long-run assumptions used for selected analytical limits do not restrict the online mechanism to stationary deployments. We evaluate CUP against q-FFL, PHP-FL, AFL, FairFedCS, FedProx selection (Section 3), together with a cached-update ablation. The study distinguishes equality of selection rates conditional on availability from parity of availability-normalized realized utility, and evaluates predictive performance, cumulative utility dispersion, and conditional participation allocation.

2 Model Design

In FL, fairness traditionally focuses on balancing loss across clients while assuming uniform or randomly distributed participation. However, in real-world deployments, client availability is temporally skewed and often correlated with data heterogeneity. Some clients are consistently underrepresented due to intermittent connectivity, energy constraints, or operational cycles. This creates structural biases that are not captured by loss alone but manifest through underutilization and misrepresentation over time. To address this, we design a fairness-aware FL framework that combines:

1. Temporal Utility Tracking: We maintain a history of each client's cumulative utility (e.g., total loss reduction) throughout training. This allows the system to detect long-term disparities in benefit allocation and adjust future scheduling priorities to compensate for underrepresented clients.

2. Adaptive Sampling via Availability Modeling: We estimate each client's availability from telemetry and combine this estimate with its historical marginal utility and participation debt. This allows the scheduler to target equal availability-normalized utility rates while respecting the realized-participation budget.

3. Representation-Aware Surrogates: To mitigate representation loss from unavailable clients, we introduce surrogate update directions derived from previously cached prototypes or model gradients. Surrogate contributions are incorporated into the global update using exponentially decayed confidence weights that discount older cached information.

- Let $F_k(w)$ denote the loss of client k under model w.
- Let $A_k(t) \in \{0, 1\}$ be the availability signal obtained only from client telemetry.
- Let $\mathcal{C}_t = \{k \in [N] : A_k(t) = 1\}$ be the eligible set observed before scheduling, and let $S_k(t) \in \{0, 1\}$ satisfy $S_k(t) \leq A_k(t)$.
- Let π_k be the long-term availability of client k.
- Let $u_k(t)$ be the cumulative utility received by client k until round t.
- Let $\hat{F}_k(w) = F_k(w)/\pi_k$ be the availability-adjusted loss.
- In the surrogate analysis, $\tilde{F}_k(w)$ denotes a scalar surrogate loss; vector client signals use distinct notation $g_k(w)$ and $\tilde{g}_k(w)$.
- N is the population size and m the per-round capacity; the scheduler selects $\min(m, |\mathcal{C}_t|)$ eligible clients.

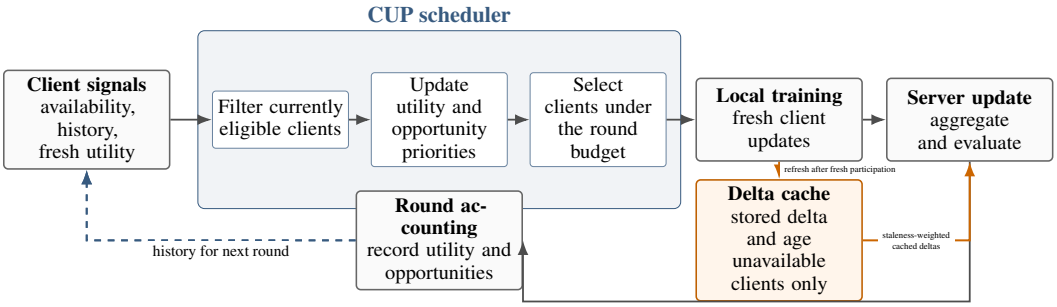


Figure 1. CUP architecture. The scheduler uses current availability and historical utility debt to prioritize eligible clients. Selected clients return fresh updates; the optional cache supplies staleness-weighted stored updates. After aggregation, realized utility and missed opportunities update the history used in the next round.

Lemma 1 (Conditional realized-utility rate). For each $k \in [N]$, assume the process $(A_k(t), S_k(t), Y_k(t))_{t\geq 1}$ is jointly stationary and ergodic, with $S_k(t) \leq A_k(t)$, and $Y_k(t) = S_k(t)\Delta u_k(t)$. Assume $\mathbb{E}|Y_k(t)| < \infty$, $\pi_k = \mathbb{E}[A_k(t)] > 0$, and $\phi_k = \mathbb{E}[S_k(t)] > 0$. Define

$$r_k = \Pr(S_k(t) = 1 \mid A_k(t) = 1), \qquad \mu_k = \mathbb{E}[\Delta u_k(t) \mid S_k(t) = 1]. \tag{1}$$

For $u_k(T) = \sum_{t=1}^{T} Y_k(t)$, almost surely,

$$\frac{u_k(T)}{T} \longrightarrow \pi_k r_k \mu_k, \qquad \frac{u_k(T)}{T\pi_k} \longrightarrow r_k \mu_k. \tag{2}$$

Moreover, stationarity gives $\mathbb{E}[u_k(T)]/T = \pi_k r_k \mu_k$ for every T. Normalizing by availability removes participation-frequency effects but not differences in scheduling or update quality. By Lemma 1, client k's long-run normalized utility rate is $r_k\mu_k$, where r_k is its conditional selection rate and μ_k is its mean utility when participating. The proof is in Appendix A. It requires joint stationarity and ergodicity, but not independence across clients. These are sufficient assumptions, not established properties of adaptive training.

Theorem 1 (Systematic normalized-utility-rate disparity). Under Lemma 1, let $\tilde{u}_k(T) = u_k(T)/\pi_k$, $\bar{u}(T) = N^{-1}\sum_k \tilde{u}_k(T)$, and

$$V_T = \frac{1}{N}\sum_{k=1}^{N}(\tilde{u}_k(T) - \bar{u}(T))^2, \qquad D = \frac{1}{N}\sum_{k=1}^{N}(r_k\mu_k - \overline{r\mu})^2, \tag{3}$$

where $\overline{r\mu} = N^{-1}\sum_k r_k\mu_k$. Then

$$\frac{V_T}{T^2} \longrightarrow D \quad \text{almost surely,} \qquad \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\mathbb{E}[\tilde{u}_k(T)]}{T} - \frac{\mathbb{E}[\bar{u}(T)]}{T}\right)^2 = D. \tag{4}$$

In particular, $r_k\mu_k = c$ for every client implies $D = 0$. If two policies each satisfy Lemma 1, and one attains equal products while the other's products are not all equal, only their systematic rate disparities are ordered: $D_{\text{equal}} = 0 < D_{\text{other}}$. The theorem identifies persistent inequality with differences among the client rates $r_k\mu_k$: unequal rates produce disparity that grows quadratically with the horizon before normalization. Equal rates eliminate this asymptotic systematic component. Appendix B proves this result.

Lemma 2. Let N be the population size and let $S_k(t) \leq A_k(t)$ denote availability-aware selection with completed training for every selected client. For the rate allocation problem, take fixed $\pi_k > 0$ and $\mu_k > 0$, where μ_k is the conditional mean utility given $S_k(t) = 1$ as in Lemma 1. Write

$$r_k := \Pr(S_k(t) = 1 \mid A_k(t) = 1) \tag{5}$$

for the long-run conditional scheduling rate. Suppose the system can support an expected realized-participation budget $b \in (0, \sum_k \pi_k]$, so that

$$\sum_{k=1}^{N}\pi_k r_k \leq b, \qquad 0 \leq r_k \leq 1. \tag{6}$$

Then the max–min cumulative-utility-parity scheduler

$$\max_{\tau,\, r_1,\ldots,r_N} \tau \quad \text{subject to} \quad r_k\mu_k \geq \tau \;\; \forall k, \quad equation\ 6, \tag{7}$$

has optimal common utility-rate target

$$\tau^* = \min\left\{ \min_k \mu_k,\, \frac{b}{\sum_{j=1}^N \pi_j/\mu_j} \right\}, \qquad r_k^* = \frac{\tau^*}{\mu_k}. \tag{8}$$

This is the canonical optimal solution that uses the least budget for the target τ^*. If

$$b \leq \left(\min_k \mu_k \right) \sum_{j=1}^N \frac{\pi_j}{\mu_j}, \tag{9}$$

then the budget is active and

$$\tau^* = \frac{b}{\sum_{j=1}^N \pi_j/\mu_j}, \qquad r_k^* = \frac{b}{\mu_k \sum_{j=1}^N \pi_j/\mu_j}, \qquad r_k^*\mu_k = \tau^* \;\forall k. \tag{10}$$

If a scheduler realizes these target conditional rates and satisfies Lemma 1, then:

$$\frac{u_k(T)}{T\pi_k} \xrightarrow{a.s.} \tau^*, \qquad \forall k, \tag{11}$$

and the systematic rate disparity in Theorem 1 vanishes. This rate-level benchmark guides the online scheduler, while feasibility under a hard per-round capacity depends on the realized joint availability process. The proof is in Appendix C. Inverse availability alone is not the CUP objective. The max–min target schedules clients with lower mean utility more often so that the products $r_k\mu_k$ are equalized, subject to rate caps and the expected participation budget. This is an allocation benchmark: an online scheduler still needs to realize these conditional rates under the actual joint availability process. Appendix D defines sliding-window diagnostics and a fixed-size eligible-client sampler.

Theorem 2 (Conditional opportunity debt). Suppose $(A_k(t), S_k(t))$ is jointly stationary and ergodic for each $k \in [N]$, with $S_k(t) \leq A_k(t)$. Let $\pi_k = \mathbb{E}[A_k(t)] > 0$, $\phi_k = \mathbb{E}[S_k(t)] = \pi_k r_k$, and

$$d_k(t) = \sum_{s<t} (A_k(s) - S_k(s)), \qquad d_k(t+1) = d_k(t) + A_k(t) - S_k(t), \qquad d_k(1) = 0. \tag{12}$$

Then $\mathbb{E}[d_k(t)] = (t-1)(\pi_k - \phi_k)$ and, almost surely,

$$\frac{d_k(t)}{t-1} \longrightarrow \pi_k - \phi_k = \pi_k(1 - r_k), \qquad \frac{d_k(t)}{\sum_{s<t} A_k(s)} \longrightarrow 1 - r_k. \tag{13}$$

For comparison, define the separate reactive weighting rule $p_k(t) = \alpha_k(1 + \lambda d_k(t))/(\pi_k + \epsilon)$ and $\widehat{p}_k(t) = p_k(t)/\sum_{j=1}^N p_j(t)$, with $\alpha_k > 0$, $\epsilon > 0$, and $\lambda > 0$. If $Z = \sum_j \alpha_j(\pi_j - \phi_j)/(\pi_j + \epsilon) > 0$, then

$$\widehat{p}_k(t) \longrightarrow \frac{\alpha_k(\pi_k - \phi_k)/(\pi_k + \epsilon)}{Z} \quad \text{almost surely.} \tag{14}$$

The limit may be zero or one. No ratio limit is asserted from this argument if $Z = 0$. If $\lambda = 0$, the weights are constant and proportional to $\alpha_k/(\pi_k + \epsilon)$, with the regularizer retained. Opportunity debt grows with missed eligible opportunities, so offline rounds create no debt. The limiting weights describe the response to accumulated missed opportunities; utility parity remains governed by both selection and realized utility.

The proof is in Appendix E. Conditional debt counts missed eligible opportunities; offline rounds incur no debt. Appendix D provides sliding-window identities and a fixed-size sampler, not a nonstationary tracking guarantee. Theorem 2 concerns a separate reactive weighting rule; it does not establish convergence of Algorithm 1's deficit-based scheduler.

Proposition 1 (CUP Conditional Optimization). The preceding results concern utility allocation. Here we analyze the fresh-only, sample-weighted aggregation in Algorithm 1 with the local training in Algorithm 2, without importance correction. Fix

$$F(w) = \sum_{k=1}^N a_k F_k(w), \qquad a_k \geq 0, \quad \sum_k a_k = 1. \tag{15}$$

For the population sample-weighted objective take $a_k = n_k / \sum_j n_j$, where $n_k > 0$ is client k's fixed sample count. Let $\mathcal{F}_t$ contain the history and w_t before current telemetry, and let $\mathcal{G}_t$ additionally contain availability and the selected set $\mathcal{S}_t$, before local training. All selected clients complete, as in the evaluated runs. Define $\omega_{k,t} = n_k \mathbf{1}\{k \in \mathcal{S}_t\} / \sum_{j \in \mathcal{S}_t} n_j$ on nonempty rounds, and $\omega_{k,t} = 0$ otherwise. For a fixed scale $\gamma > 0$, write

$$d_{k,t} = \frac{w_t - w_{k,t}^{\mathrm{loc}}}{\gamma}, \qquad G_t = \sum_k \omega_{k,t} d_{k,t}, \qquad w_{t+1} = w_t - \gamma G_t. \tag{16}$$

Here $d_{k,t} = 0$ for unselected clients; it is a scaled local model displacement, not an assumed unbiased gradient. Equation 16 matches Algorithm 1 with $c = 0$, including $G_t = 0$ on empty rounds. Define

$$m_t = \mathbb{E}[G_t \mid \mathcal{F}_t], \qquad b_t = m_t - \nabla F(w_t), \qquad \mathcal{B}_T^2 = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|b_t\|^2. \tag{17}$$

The term $\mathcal{B}_T^2$ measures systematic error relative to the fixed population objective, whereas σ^2 below measures fluctuations around the conditional mean m_t. Unequal selection, random FedAvg normalization, and local training can make b_t nonzero even with exact local gradients. Assume w_0 is deterministic, F is bounded below by F_* and has an L-Lipschitz gradient, and the relevant second moments are finite. Suppose, almost surely, $\mathbb{E}[\|G_t - m_t\|^2 \mid \mathcal{F}_t] \le \sigma^2 < \infty$ uniformly in t. If $0 < \gamma \le 1/L$, then

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w_t)\|^2 \le \frac{2(F(w_0) - F_*)}{\gamma T} + L\gamma\sigma^2 + \mathcal{B}_T^2. \tag{18}$$

For a family of runs with $\gamma = \gamma_0/\sqrt{T}$, $0 < \gamma_0 \le 1/L$, and a variance bound uniform also in T,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w_t)\|^2 \le \frac{2(F(w_0) - F_*)/\gamma_0 + L\gamma_0\sigma^2}{\sqrt{T}} + \mathcal{B}_T^2. \tag{19}$$

Thus an $O(T^{-1/2})$ stationarity rate additionally requires $\mathcal{B}_T^2 = O(T^{-1/2})$; CUP scheduling and FedAvg alone do not establish this. With fixed γ and $\mathcal{B}_T^2 \le B^2$, the bound retains $L\gamma\sigma^2 + B^2$. The scale γ must be interpreted jointly with the local displacement: for H plain-SGD steps of size η, a natural choice is $\gamma = H\eta$. Merely rescaling a fixed implementation does not establish a diminishing-step rate, since the moments and bias of $d_{k,t}$ change with that scale.

Appendix F proves the bound and decomposes b_t into effective aggregation-weight mismatch and local-training error. No true inclusion probabilities or unbiased local model updates are required. The bound is conditional on its stated moment and smoothness assumptions; it is not a certificate that the experimental hyperparameters satisfy them. The utility-rate conclusions of Lemmas 1–2 and Theorems 1–2 retain their own assumptions and do not imply $b_t = 0$.

Cached representation surrogates. Let $\mathcal{S}_t = \{k : S_k(t) = 1\}$ be the freshly selected clients and $\mathcal{M}_t \subseteq [N] \setminus \mathcal{C}_t$ the unavailable clients represented by a stored cache. Each cache records its last refresh round $\tau_k < t$, giving staleness $\delta_k = t - \tau_k$; clients without a cache supply no surrogate. Write $g_k(w) \in \mathbb{R}^d$ for a client update direction, such as $\nabla F_k(w)$. We either reuse a cached gradient, $\widetilde{g}_k(w_t) = \widehat{g}_k(\tau_k)$, or construct a direction from a cached prototype representation $\mathcal{P}_k(\tau_k)$ via $\widetilde{g}_k(w_t) = H_k(w_t; \mathcal{P}_k(\tau_k))$. The map H_k must produce a direction with the same dimension and sign convention as g_k; prototype vectors are not directly added to model parameters. For nonnegative base aggregation weights $a_k(t)$, set

$$\eta_k(t) = \eta_0 e^{-\lambda\delta_k}, \qquad \eta_0 > 0, \quad \lambda > 0, \qquad Z_t = \sum_{k \in \mathcal{S}_t} a_k(t) + \sum_{k \in \mathcal{M}_t} a_k(t)\eta_k(t) \tag{20}$$

When $Z_t > 0$, the surrogate-enabled aggregate and global update are

$$\widetilde{G}_t(w_t) = \frac{\sum_{k \in \mathcal{S}_t} a_k(t) g_k(w_t) + \sum_{k \in \mathcal{M}_t} a_k(t)\eta_k(t)\widetilde{g}_k(w_t)}{Z_t}, \qquad w_{t+1} = w_t - \gamma \widetilde{G}_t(w_t) \tag{21}$$

If $Z_t = 0$, the model is unchanged. Cache reuse neither marks an unavailable client as selected nor credits it with newly realized training utility: $S_k(t) = 0$ on $\mathcal{M}_t$, and opportunity debt remains based on $A_k(t) - S_k(t)$. Lemma 3 and Theorem 3, their approximation assumptions (including those required for H_k), and their proofs appear in Appendices G and H.

The cache is an optional model-update mechanism, not a substitute for an actual participation event. Exponential decay reduces the influence of stale information, while normalization by Z_t keeps fresh and cached contributions on a common aggregation scale. Lemma 3 bounds scalar loss replacement error; under separate vector accuracy assumptions, Theorem 3 bounds aggregate-direction error.

3 Empirical Evaluation

3.1 Infrastructure

We evaluate the availability-aware framework in Algorithms 1 and 2 on CIFAR-10 under non-IID partitions with two and five labels per client (Li et al., 2020b; McMahan et al., 2017). The two-label and five-label experiments use identical settings; only the number of CIFAR-10 classes assigned to each logical client changes. We use a Torchvision-pretrained ResNet, freeze the backbone, and train only the classifier. Selected clients train for one local epoch with mini-batches of 32 using SGD with learning rate 0.01 and zero momentum. We run $T = 50$ rounds and select at most $m = 10$ clients.

Availability. Our implementation is based on the real-world availability traces of mobile devices in an FL system published by Yang et al. (2021). The distribution of device availability percentages indicates that most devices are available less than 40% of the time (Behfar & Mortier (2026)), where availability percentage is defined as the fraction of time (between the first and last observation of a device) during which it is in a state suitable for FL participation. The trace records timestamped events such as *WiFi on/off* and *battery charging on/off*. Following Google's FL readiness definition Bonawitz et al. (2019), we treat a device as available when it is simultaneously charging and connected to WiFi.

Process isolation. Each client and the coordinator run in separate Docker containers. Each container has an independent Python runtime containing the model weights and inference scripts. CIFAR-10 is mounted read-only from the host, avoiding duplicated storage while providing every worker with a consistent dataset view. Real FL deployments may contain thousands or millions of heterogeneous mobile and edge devices with non-IID data, intermittent availability, and correlated failures. Reproducing that population with one physical process per device is impractical because of its compute, memory, and orchestration cost. We therefore adopt *logical clients* (Lai et al., 2021): virtual participants whose data and availability are modeled explicitly, while a bounded pool of physical workers performs their local training. At each round, the scheduler selects logical clients and maps them to the physical workers in execution waves. If m logical clients are selected and P workers are available, local training requires $\lceil m/P \rceil$ waves, with workers processing multiple logical clients sequentially when necessary. Execution failures are outside the present evaluation.

3.2 CUP Scheduling and Evaluation

CUP credits utility only after completed participation: $u_k(t) = u_k(t-1) + S_k(t)\Delta u_k(t)$. Here, $\Delta u_k(t)$ is the observed loss reduction clipped to $[0, 1]$, and availability is estimated from telemetry as $\hat{\pi}_k(t) = t^{-1} \sum_{\tau=1}^{t} A_k(\tau)$. The scheduler combines availability-normalized utility deficits with missed-opportunity debt, so an offline client incurs no scheduling penalty. Online denominators use a positive floor for numerical stability; normalized utility, $u_k(T)/\hat{\pi}_k(T)$, is defined only for clients with at least one observed opportunity, while observed clients with zero utility remain in the evaluation. The performance figures show accuracy, utility CV, and Jain index at every logged round, see Algorithm 1 and 2 for the implementation steps.

3.3 Baseline evaluation Metrics

(i) Predictive performance. We distinguish among recorded predictive performance, dispersion in cumulative utility, and allocation of available scheduling opportunities. Accuracy, CV, and Jain index are evaluated at every logged round and summarized over the full trajectory. Figures 2 and 3 plot this field, while Tables 1 and 2 report final and peak global accuracy.

Algorithm 1. Availability-aware CUP with optional cache

Require: $T, m, b_0, w^{(0)}; b, c \in \{0,1\}; \epsilon, \lambda, \eta_0, \gamma > 0$
$w \leftarrow w^{(0)}; u_k, d_k, O_k, B_k \leftarrow 0; \widehat{\mu}_k \leftarrow 1; \mathcal{H} \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  Observe $A_k(t); \mathcal{C}_t \leftarrow \{k : A_k(t) = 1\}; h_t \leftarrow \min\{m, |\mathcal{C}_t|\}$
  $\pi_k^- \leftarrow O_k/(t-1)$ if $t > 1$, else 1; $O_k \leftarrow O_k + A_k(t)$
  $\widehat{\tau}_t \leftarrow 0$ if $\pi^- = \mathbf{0}$; else use (8) with $(\pi^-, \widehat{\mu}, b_0)$
  $e_k \leftarrow [t\widehat{\tau}_t - u_k / \max\{\pi_k^-, \epsilon\}]_+; q_k \leftarrow (e_k + \epsilon)(1 + \lambda d_k)/(\pi_k^- + \epsilon)$
  $q_k \leftarrow 1$ if $b = 0$
  Draw $\mathcal{S}_t \subseteq \mathcal{C}_t, |\mathcal{S}_t| = h_t$, sequentially without replacement with weights q_k
  $S_k(t) = P_k(t) \leftarrow \mathbf{1}\{k \in \mathcal{S}_t\}$
  Obtain $(w_k, n_k, \Delta u_k)$ for $k \in \mathcal{S}_t$ from Algorithm 2
  For $k \in \mathcal{S}_t$: $g_k \leftarrow (w - w_k)/\gamma; \mathcal{H}_k \leftarrow (g_k, t, n_k)$
  $\mathcal{M}_t \leftarrow \mathrm{dom}(\mathcal{H}) \setminus \mathcal{C}_t$ if $c = 1$, else $\varnothing$; $a_k \leftarrow n_k$
  Apply (20)–(21) with $(\widetilde{g}_k, \tau_k, n_k) = \mathcal{H}_k$; retain w if $Z_t = 0$
  $k \in \mathcal{S}_t$: $u_k \leftarrow u_k + \Delta u_k$; $\widehat{\mu}_k \leftarrow \max\{(B_k\widehat{\mu}_k + \Delta u_k)/(B_k + 1), 10^{-3}\}$
  $B_k \leftarrow B_k + S_k(t); \widehat{\pi}_k(t) \leftarrow O_k/t; d_k \leftarrow d_k + A_k(t) - S_k(t)$
end for

Algorithm 2. Client local training

Require: Data $\mathcal{D}_k$, epochs E, optimizer OPT
while request (t, w) is received **do**
  $w_k \leftarrow w; \ell_k \leftarrow F_k(w)$
  Train w_k on $\mathcal{D}_k$ for E epochs using OPT
  $\Delta u_k \leftarrow \min\{1, [\ell_k - F_k(w_k)]_+\}$; return $(w_k, |\mathcal{D}_k|, \Delta u_k)$
end while

(ii) Cumulative participation metrics.

Availability-normalized cumulative utility. Utility is measured as the reduction in client k's local loss. $z_k(t) = F_k(w_t) - F_k(w_{k,t}^{\mathrm{loc}})$, where both losses are evaluated on the same client data under the same evaluation procedure. The audit verifies that each logged increment equals $S_k(t)\min\{1, \max\{0, z_k(t)\}\}$. Thus, for this evaluation, $\Delta u_k(t) = \min\{1, \max\{0, z_k(t)\}\}$, and $u_k(T) = \sum_{t=1}^{T} S_k(t)\Delta u_k(t)$. This is clipped loss reduction, not an accuracy increment or a guarantee of net global loss improvement. Normalize by telemetry-derived $\hat{\pi}_k(T) = O_k(T)/T$ to obtain $\tilde{u}_k(T) = u_k(T)/\hat{\pi}_k(T)$ only for $\mathcal{I}_T = \{k : O_k(T) > 0\}$. Let $n_T = |\mathcal{I}_T|$. All normalized-utility summaries use e.g. n_T= available clients, including observed clients with zero utility/selections.

- Utility Coefficient of Variation (Utility CV): $\mathrm{CV}(\tilde{u}) = \sigma(\tilde{u})/\bar{u}$, where $\bar{u} = n_T^{-1}\sum_{k\in\mathcal{I}_T}\tilde{u}_k$ and $\sigma^2 = n_T^{-1}\sum_{k\in\mathcal{I}_T}(\tilde{u}_k - \bar{u})^2$. The denominator is positive for every completed run; no numerical floor is added to the CV. It is undefined when the mean is zero or $n_T = 0$.
- Jain's Fairness Index (Utility): $J(\tilde{u}) = (\sum_{k\in\mathcal{I}_T}\tilde{u}_k)^2/(n_T\sum_{k\in\mathcal{I}_T}\tilde{u}_k^2)$, undefined if its denominator is zero. With these conventions $J = 1/(1 + \mathrm{CV}^2)$, so CV and Jain are complementary presentations of the same dispersion statistic, not independent evidence.

Lower Utility CV and higher Jain index indicate stronger cumulative fairness.

Conditional opportunity allocation. The relevant denominator is the number of available scheduling opportunities, not the total number of rounds. Define

$$O_k(T) = \sum_{t=1}^{T} A_k(t), \qquad B_k(T) = \sum_{t=1}^{T} S_k(t), \qquad \widehat{r}_k(T) = \frac{B_k(T)}{O_k(T)} \quad \text{when } O_k(T) > 0. \tag{22}$$

Since $S_k(t) \le A_k(t)$, each defined rate lies in $[0,1]$. Let $\mathcal{I}_T = \{k \in [N] : O_k(T) > 0\}$. Then,

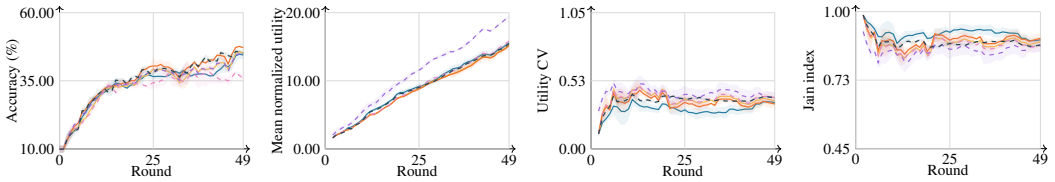
$$\mathrm{CondSelGap}(T) = \max_{k\in\mathcal{I}_T}\widehat{r}_k(T) - \min_{k\in\mathcal{I}_T}\widehat{r}_k(T). \tag{23}$$

This gap lies in $[0,1]$ and is undefined if $\mathcal{I}_T$ is empty. Report $O_k(T)$, $B_k(T)$, and coverage $|\mathcal{I}_T|/N$ with the gap; clients with no opportunities have undefined rates, whereas available but never-selected clients have rate zero. With one observed client the gap is zero but says nothing about population-wide parity. The scheduler selects $\min(m, |\mathcal{C}_t|)$ clients per round, so a uniform count target based on full eligibility and full capacity is inappropriate. Conditional selection-rate equality is also distinct from CUP: CUP targets equality of $r_k\mu_k$, not necessarily of r_k when conditional utilities differ.

Distribution-wide allocation inequality and lower-tail utility. For final-horizon allocation context, we also compute the Gini coefficient of conditional selection rates,

$$G_r(T) = \frac{\sum_{i,j\in\mathcal{I}_T}|\widehat{r}_i(T) - \widehat{r}_j(T)|}{2n_T\sum_{k\in\mathcal{I}_T}\widehat{r}_k(T)} \tag{24}$$

It is undefined when its denominator is zero. Unlike the max–min gap, it depends on all pairwise rate differences. Define utility per observed opportunity as $v_k(T) = u_k(T)/O_k(T) = \tilde{u}_k(T)/T$ for $k \in \mathcal{I}_T$. Let $v_{(1)} \le \cdots \le v_{(n_T)}$ be their ordered values. For $n_T > 0$, define



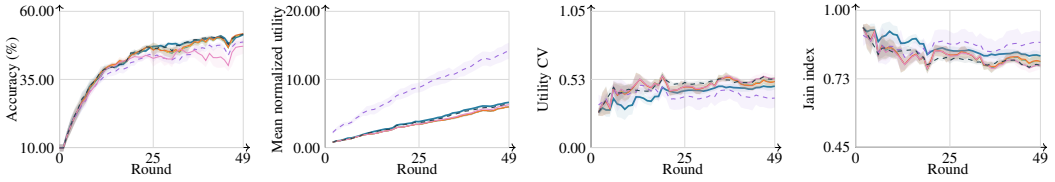
(a) Accuracy **(b)** Mean normalized utility **(c)** Utility CV ↓ **(d)** Jain index ↑

Figure 2. Round-wise two-label means over three complete seeds per method. Colors identify (CUP without caching: blue; FedProx: yellow; q-FFL: orange; PHP-FL: purple; AFL: pink; FairFedCS: navy)

.

Table 1. Two-label results, including CUP with and without cached model deltas.

Method	n	Final acc. (%)	Peak acc. (%)	Mean CV ↓	Mean Jain ↑	Gap ↓	G_r ↓	L_{10} ↑	Z_u ↓
CUP (no cache)	3	44.47 ± 1.48	44.83 ± 1.89	0.303 ± 0.034	0.913 ± 0.017	0.619 ± 0.029	0.134 ± 0.009	0.1532 ± 0.0062	1.0 ± 0.0
CUP + cache	3	46.66 ± 1.44	46.66 ± 1.44	0.297 ± 0.031	0.917 ± 0.015	0.548 ± 0.041	0.128 ± 0.005	0.1847 ± 0.0102	0.0 ± 0.0
FedProx	3	45.18 ± 0.83	45.73 ± 0.46	0.380 ± 0.024	0.872 ± 0.014	0.791 ± 0.149	0.150 ± 0.017	0.1048 ± 0.0154	0.7 ± 0.5
q-FFL	3	47.15 ± 0.53	47.66 ± 0.41	0.356 ± 0.029	0.885 ± 0.017	0.791 ± 0.149	0.150 ± 0.017	0.1202 ± 0.0174	0.3 ± 0.5
PHP-FL	3	44.96 ± 0.41	45.20 ± 0.31	0.427 ± 0.044	0.844 ± 0.027	0.791 ± 0.149	0.150 ± 0.017	0.1198 ± 0.0306	0.7 ± 0.5
AFL	3	36.02 ± 1.19	40.13 ± 2.91	0.387 ± 0.026	0.867 ± 0.016	0.791 ± 0.149	0.150 ± 0.017	0.0972 ± 0.0151	0.7 ± 0.5
FairFedCS	3	45.13 ± 0.39	45.79 ± 0.65	0.371 ± 0.009	0.878 ± 0.005	0.622 ± 0.031	0.163 ± 0.001	0.1289 ± 0.0057	0.0 ± 0.0



(a) Accuracy **(b)** Mean normalized utility **(c)** Utility CV ↓ **(d)** Jain index ↑

Figure 3. Round-wise five-label means over three complete seeds per method. Colors identify CUP without caching (blue), uniform available (gray), FedAvg (green), FedProx (yellow), q-FFL (orange), PHP-FL (purple), AFL (pink), and FairFedCS (navy). Bands show population standard deviations.

Table 2. Five-label results, including CUP with and without cached model deltas.

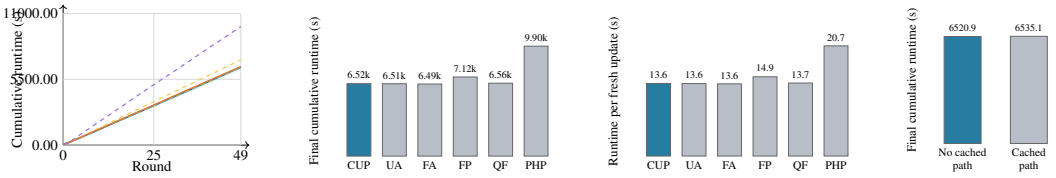
Method	n	Final acc. (%)	Peak acc. (%)	Mean CV ↓	Mean Jain ↑	Gap ↓	G_r ↓	L_{10} ↑	Z_u ↓
CUP (no cache)	3	51.46 ± 0.40	51.46 ± 0.40	0.413 ± 0.030	0.852 ± 0.017	0.529 ± 0.010	0.134 ± 0.001	0.0342 ± 0.0062	0.3 ± 0.5
CUP + cache	3	53.14 ± 0.43	53.70 ± 0.19	0.397 ± 0.030	0.864 ± 0.015	0.483 ± 0.022	0.130 ± 0.001	0.0372 ± 0.0023	0.0 ± 0.0
FedProx	3	51.65 ± 0.21	51.65 ± 0.21	0.471 ± 0.024	0.817 ± 0.015	0.791 ± 0.149	0.150 ± 0.017	0.0209 ± 0.0084	1.7 ± 0.9
q-FFL	3	51.81 ± 0.18	51.81 ± 0.18	0.466 ± 0.025	0.820 ± 0.016	0.791 ± 0.149	0.150 ± 0.017	0.0206 ± 0.0080	1.3 ± 0.5
PHP-FL	3	48.76 ± 0.63	48.76 ± 0.63	0.405 ± 0.072	0.857 ± 0.044	0.791 ± 0.149	0.150 ± 0.017	0.1229 ± 0.0230	0.3 ± 0.5
AFL	3	47.19 ± 2.77	48.16 ± 1.79	0.477 ± 0.019	0.813 ± 0.012	0.791 ± 0.149	0.150 ± 0.017	0.0198 ± 0.0064	1.7 ± 0.5
FairFedCS	3	51.65 ± 0.42	51.65 ± 0.42	0.478 ± 0.017	0.813 ± 0.011	0.542 ± 0.019	0.129 ± 0.004	0.0182 ± 0.0018	0.3 ± 0.5

$$L_{10}(T) = \frac{1}{q_T} \sum_{i=1}^{q_T} v_{(i)}(T), \qquad q_T = \lceil 0.1 n_T \rceil, \qquad Z_u(T) = \sum_{k \in \mathcal{I}_T} \mathbf{1}\{u_k(T) = 0\} \tag{25}$$

Thus L_{10} is a lower-tail mean, not the tenth-percentile threshold; with $n_T = 47$ (available clients), it averages the five lowest values, including zeros. It is undefined when $n_T = 0$. The zero-utility count Z_u includes both never-selected clients and selected clients whose clipped gains remain zero. Served coverage is $|\{k \in \mathcal{I}_T : B_k(T) > 0\}|/n_T$, undefined when $n_T = 0$. These metrics retain observed clients with zero utility and exclude clients without opportunities. Lower-tail utility is interpreted within each label setting, since the scale of clipped loss reduction can change with the partition.

3.4 Round-wise results and surrogate ablation

Figure 2 shows four aligned trajectories, averaged over three complete seeds per method. CUP without caching reaches $44.83 \pm 1.89\%$ peak and $44.47 \pm 1.48\%$ final global accuracy, compared with $47.15 \pm 0.53\%$ final accuracy for q-FFL. The shaded bands represent population standard



(a) Cumulative runtime **(b)** Final runtime **(c)** Runtime/fresh-update **(d)** CUP ablation

Figure 4. The graphs show directly measured systems cost. They abbreviate uniform available as UA, FedAvg as FA, FedProx as FP, q-FFL as QF, and PHP-FL as PHP.

deviations. The cumulative metrics reflect accumulated clipped loss reductions normalized by availability; their late-round behavior should be distinguished from instantaneous training gains. CUP without caching has lower temporal mean utility dispersion than the two-label baselines, with CV 0.303 ± 0.034 and Jain 0.913 ± 0.017. Figure 3 and Table 2 summarize the five-label setting. Mean final accuracy is $51.46 \pm 0.40\%$ for CUP without caching, $51.65 \pm 0.21\%$ for FedProx, $51.81 \pm 0.18\%$ for q-FFL, and $48.76 \pm 0.63\%$ for PHP-FL. CUP's temporal mean CV (0.413 ± 0.030) and Jain index (0.852 ± 0.017) beat almost all baselines except that PHP-FL has slightly better means (0.405 ± 0.072 and 0.857 ± 0.044, respectively). All methods have three complete seeds; reported deviations describe between-run variation, not statistical significance.

Table 1 and Table 2 compare replicated CUP variants with and without cached model deltas. In the two-label setting, caching increases mean final accuracy from $44.47 \pm 1.48\%$ to $46.66 \pm 1.44\%$, reduces temporal mean CV from 0.303 to 0.297, and increases Jain from 0.913 to 0.917. In the five-label setting, mean final accuracy increases from $51.46 \pm 0.40\%$ to $53.14 \pm 0.43\%$, while CV decreases from 0.413 to 0.397 and Jain increases from 0.852 to 0.864. Thus, CUP variants with caching have higher accuracy, lower utility CV, and higher Jain index than their counterparts.

3.5 RUNTIME OVERHEAD

Figure 4 retains the earlier runtime measurements; these are not recomputed from the additional baseline runs used to update Figure 2 and Table 1. It reports the directly measured systems cost for that earlier run set. Panel (a) shows how elapsed time accumulates: the near-linear trajectories indicate broadly stable per-round cost, while their slopes expose method-level differences. CUP finishes in 6520.9 seconds; the baseline means range from 6487.1 seconds for FedAvg to 9900.5 seconds for PHP-FL. FedProx is also slower at 7119.0 seconds, whereas uniform available and q-FFL remain close to CUP at 6513.8 and 6559.4 seconds. Panels (b) and (c) separate total experiment duration from an event-normalized view. Dividing final runtime by the number of fresh selected-client updates gives 13.6 seconds per update for CUP, compared with 13.6 for uniform available, 13.6 for FedAvg, 14.9 for FedProx, 13.7 for q-FFL, and 20.7 for PHP-FL. This normalization is useful for diagnosing whether a longer run merely processed more selected updates; evaluation, data loading, scheduling, and orchestration time remain included in the numerator. Panel (d) isolates the archived CUP cache variants. Enabling the cached path changes final runtime from 6520.9 to 6535.1 seconds, an increase of 14.2 seconds or 0.22%. This small difference shows that no substantial wall-clock penalty is visible; it does not establish negligible memory cost.

4 CONCLUSION

This paper introduces cumulative utility parity as a resource-allocation view of fairness under intermittent client availability. The availability-aware scheduler combines telemetry, historical utility deficits, and conditional opportunity debt without treating offline rounds as missed allocations. Under stated assumptions, the analysis identifies the normalized utility rate $r_k \mu_k$, characterizes systematic disparity, and derives a closed-form max–min allocation benchmark under an expected participation budget. CUP without caching achieves lower mean temporal utility dispersion than all evaluated baselines in the two-label setting and than FedProx, q- FFL, AFL, and FairFedCS in the five-label setting. Across both settings, the cached CUP variants achieve the lowest mean CV and highest mean Jain index. Also, cached CUP achieves almost the highest mean final accuracy. These results highlight CUP's potential to improve cumulative fairness alongside predictive performance.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, consistency checks, revisions to algorithm pseudocode, clarification of theoretical claims. The authors take responsibility for the final content of this work.

REFERENCES

Stefan Behfar and Richard Mortier. Data-availability co-correlation: A missing dimension in fair federated learning. *EdgeSys @MobiSys'26*, 2026.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečny, Stefano Mazzocchi, and Brendan McMahan et al. Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1: 374–388, 2019.

Bhaskar Ray Chaudhury, Linyi Li, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/pdf/2211.02091.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies, 2020. URL https://arxiv.org/abs/2010.01243.

Oscar Dilley, Juan Marcelo Parra-Ullauri, Rasheed Hussain, and Dimitra Simeonidou. Federated fairness analytics: Quantifying fairness in federated learning, 2024. URL https://arxiv.org/abs/2408.08214.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*, pp. 19–35, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020a. URL https://arxiv.org/abs/1812.06127.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *ICLR*, 2020b. URL https://arxiv.org/pdf/1905.10497.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2017. URL https://arxiv.org/abs/1602.05629.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning (ICML)*, pp. 4615–4625, 2019. URL https://arxiv.org/abs/1902.00146.

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. doi: 10.1109/ICC.2019.8761315.

Yuxin Shi, Zelei Liu, Zhuan Shi, and Han Yu. Fairness-aware client selection for federated learning. *IEEE International Conference on Multimedia and Expo (ICME)*, 2023. URL https://arxiv.org/abs/2307.10738.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.13648.

Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1615–1623, 2021. doi: 10.24963/ijcai.2021/223. URL https://www.ijcai.org/proceedings/2021/223.

Siyuan Wu, Yongzhe Jia, Haolong Xiang, Xiaolong Xu, Xuyun Zhang, Lianyong Qi, and Wanchun Dou. A fair federated learning method for handling client participation probability inconsistencies in heterogeneous environments. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chengxu Yang, Qipeng Wang, Mengwei Xu, Zhenpeng Chen, Kaigui Bian, Yunxin Liu, and Xuanzhe Liu. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *Proceedings of the Web Conference 2021*, pp. 935–946. ACM, 2021. doi: 10.1145/3442381.3449851. URL https://dl.acm.org/doi/10.1145/3442381.3449851.

A Proof of Lemma 1

Because $S_k(t) \le A_k(t)$,

$$\phi_k = \Pr(S_k(t) = 1) = \Pr(A_k(t) = 1) \Pr(S_k(t) = 1 \mid A_k(t) = 1) = \pi_k r_k. \tag{26}$$

Conditioning on realized participation, not on a hypothetical unconditional utility, yields

$$\mathbb{E}[Y_k(t)] = \phi_k \mathbb{E}[\Delta u_k(t) \mid S_k(t) = 1] = \pi_k r_k \mu_k. \tag{27}$$

Applying the ergodic theorem to the integrable process $Y_k(t)$ yields

$$\frac{1}{T} \sum_{t=1}^{T} Y_k(t) \longrightarrow \mathbb{E}[Y_k(t)] \qquad \text{almost surely.} \tag{28}$$

Division by $\pi_k > 0$ proves equation 2. Linearity of expectation and stationarity give the finite-T expectation identity. The argument permits cross-client dependence and does not require availability and selection to be independent. Stationarity of availability alone would not suffice for applying the ergodic theorem to realized utility.

B Proof of Theorem 1: systematic rate disparity

By Lemma 1, $\tilde{u}_k(T)/T \to r_k \mu_k$ almost surely. Since N is finite, these limits hold simultaneously for all clients. Their average converges to $\overline{r\mu}$, and continuity of the finite sum of squared centered coordinates gives

$$\frac{V_T}{T^2} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{\tilde{u}_k(T)}{T} - \frac{\bar{u}(T)}{T} \right)^2 \longrightarrow \frac{1}{N} \sum_{k=1}^{N} (r_k \mu_k - \overline{r\mu})^2 = D \tag{29}$$

almost surely. The expectation identity in Lemma 1 gives $\mathbb{E}[\tilde{u}_k(T)]/T = r_k \mu_k$ directly, so its cross-client variance is D for every horizon. All centered products vanish exactly when all products are equal. This proves the systematic comparison in Theorem 1 without comparing upper bounds on random variances.

To distinguish the systematic term from fluctuations, suppose additionally that each $\tilde{u}_k(T)$ has a finite second moment at a fixed horizon. Define $\xi_k(T) = \tilde{u}_k(T) - \mathbb{E}[\tilde{u}_k(T)]$ and $\bar{\xi}(T) = N^{-1} \sum_k \xi_k(T)$. Expanding squares and using $\mathbb{E}[\xi_k(T)] = 0$ yields the exact identity

$$\mathbb{E}[V_T] = T^2 D + \mathbb{E}\left[\frac{1}{N} \sum_{k=1}^{N} (\xi_k(T) - \bar{\xi}(T))^2 \right]. \tag{30}$$

The final term depends on the policy and includes cross-client covariance. Even when one policy has $D = 0$, this term need not be smaller than the other policy's. Thus the proof establishes neither a strict nor a weak ordering of actual finite-horizon expected variances. No bound on this fluctuation term is assumed in the theorem.

C Lemma 2

Lemma 2. Let N be the population size and let $S_k(t) \le A_k(t)$ denote availability-aware selection with completed training for every selected client. For the rate allocation problem, take fixed $\pi_k > 0$ and $\mu_k > 0$, where μ_k is the conditional mean utility given $S_k(t) = 1$ as in Lemma 1. Write

$$r_k := \Pr(S_k(t) = 1 \mid A_k(t) = 1) \tag{31}$$

for the long-run conditional scheduling rate. Suppose the system can support an expected realized-participation budget $b \in (0, \sum_k \pi_k]$, so that

$$\sum_{k=1}^{N} \pi_k r_k \le b, \qquad 0 \le r_k \le 1. \tag{32}$$

Then the max–min cumulative-utility-parity scheduler

$$\max_{\tau, r_1,\dots,r_N} \tau \quad \text{subject to} \quad r_k\mu_k \geq \tau \ \forall k, \quad \textit{equation } 32, \tag{33}$$

has optimal common utility-rate target

$$\tau^* = \min\left\{ \min_k \mu_k, \frac{b}{\sum_{j=1}^N \pi_j/\mu_j} \right\}, \qquad r_k^* = \frac{\tau^*}{\mu_k}. \tag{34}$$

This is the canonical optimal solution that uses the least budget for the target τ^*. If

$$b \leq \left(\min_k \mu_k\right) \sum_{j=1}^N \frac{\pi_j}{\mu_j}, \tag{35}$$

then the budget is active and

$$\tau^* = \frac{b}{\sum_{j=1}^N \pi_j/\mu_j}, \qquad r_k^* = \frac{b}{\mu_k \sum_{j=1}^N \pi_j/\mu_j}, \qquad r_k^*\mu_k = \tau^* \ \forall k. \tag{36}$$

If a scheduler realizes these target conditional rates and satisfies Lemma 1, then:

$$\frac{u_k(T)}{T\pi_k} \xrightarrow{a.s.} \tau^*, \qquad \forall k, \tag{37}$$

and the systematic rate disparity in Theorem 1 vanishes. This is a conditional consequence, not a proof that the online heuristic realizes the rates. The expected-budget constraints are a rate-level relaxation; realizability with a hard per-round capacity depends on the joint availability process. Inverse availability is not itself the CUP objective.

Interpretation. The max–min target assigns higher conditional scheduling rates to clients with lower mean utility so that the products $r_k\mu_k$ are equalized, subject to the expected budget and rate caps.

Proof. Lemma 1 shows that client k's availability-normalized utility converges at rate $r_k\mu_k$. Therefore, maximizing the smallest normalized utility rate gives program equation 7. For any proposed target τ, the smallest scheduling rate that satisfies $r_k\mu_k \geq \tau$ is $r_k = \tau/\mu_k$. This choice is feasible only if $\tau \leq \min_k \mu_k$, because $r_k \leq 1$, and its expected participation cost is

$$\sum_{k=1}^N \pi_k \frac{\tau}{\mu_k} = \tau \sum_{k=1}^N \frac{\pi_k}{\mu_k}. \tag{38}$$

Thus the largest feasible target in the rate relaxation is exactly equation 8. Under equation 9, the participation budget is active and gives equation 10. If the budget is larger, the max–min target saturates at $\min_k \mu_k$ and the canonical parity solution leaves the excess budget unused. If these rates are realized and Lemma 1 applies, Theorem 1 gives zero systematic rate disparity.

Appendix D gives descriptive window identities and a fixed-size eligible-client sampler, not a nonstationary or correlated-availability guarantee.

D SLIDING-WINDOW MONITORING AND FIXED-SIZE SAMPLING

Use N for population size and m for per-round capacity. For the window $I_t = \{\max(1, t - W + 1), \dots, t\}$ with $L_t = |I_t|$, define

$$\begin{aligned} O_{k,W}(t) &= \sum_{s\in I_t} A_k(s), & B_{k,W}(t) &= \sum_{s\in I_t} S_k(s), \\ \hat{\pi}_{k,W}(t) &= O_{k,W}(t)/L_t, & \hat{r}_{k,W}(t) &= B_{k,W}(t)/O_{k,W}(t), \\ U_{k,W}(t) &= \sum_{s\in I_t} S_k(s)\Delta u_k(s), & \hat{\mu}_{k,W}(t) &= U_{k,W}(t)/B_{k,W}(t) \end{aligned} \tag{39}$$

Ratios with zero denominators are undefined. If both counts are positive, the definitions give the exact descriptive identity

$$\frac{U_{k,W}(t)}{L_t \hat{\pi}_{k,W}(t)} = \hat{r}_{k,W}(t)\hat{\mu}_{k,W}(t) \tag{40}$$

No stationarity assumption is needed for this finite-window identity; it is not a guarantee of tracking a time-varying population rate.

Availability-first fixed-size sampling. After observing $A(t)$, form $\mathcal{C}_t = \{k \in [N] : A_k(t) = 1\}$ and $h_t = \min(m, |\mathcal{C}_t|)$. A valid dependent sampling procedure is to draw h_t clients sequentially without replacement. At draw ℓ, let $\mathcal{R}_{t,\ell}$ be the remaining eligible clients, and draw k with probability

$$\frac{q_k(t)}{\sum_{j \in \mathcal{R}_{t,\ell}} q_j(t)}, \qquad k \in \mathcal{R}_{t,\ell}, \qquad q_k(t) > 0. \tag{41}$$

The scores may be specified deficit-based priorities. The choice $q_k(t) = 1/\max\{\hat{\pi}_{k,W}(t), \epsilon\}$ is a separate inverse-availability heuristic, not the exact CUP optimizer. Draw probabilities sum to one *at each draw*; they are not the first-order inclusion probabilities of the final subset. If $h_t = 0$, select nobody. Otherwise the final indicators satisfy $S_k(t) \le A_k(t)$, and $\sum_k S_k(t) = h_t$.

Let $\mathcal{H}_{t-1}$ denote the information available before the current telemetry observation, and define the actual first-order inclusion probabilities

$$\rho_k(t) = \Pr(S_k(t) = 1 \mid \mathcal{H}_{t-1}, A(t)). \tag{42}$$

Then $0 \le \rho_k(t) \le A_k(t)$ and $\sum_{k=1}^N \rho_k(t) = h_t$, by taking conditional expectations of the fixed-size constraint. Uniform positive scores give $\rho_k(t) = h_t/|\mathcal{C}_t|$ for eligible clients; unequal scores generally do not give $\rho_k(t) = h_t q_k(t)/\sum_{j \in \mathcal{C}_t} q_j(t)$. Marginals must be computed for the actual sampler, not replaced by its scores.

Expectation and scope. By the tower property, $\mathbb{E}[S_k(t)] = \mathbb{E}[\rho_k(t)]$. Equivalently this is $\pi_k(t)\mathbb{E}[\rho_k(t) \mid A_k(t) = 1]$ when $\pi_k(t) = \Pr(A_k(t) = 1) > 0$. The inner expectation includes the other clients' availability and the history; it cannot be replaced by a deterministic ratio over $[m]$. The earlier drift bound with unspecified constants is removed. A quantitative guarantee would require an explicit joint availability model, estimator-error bounds, the actual sampling law, and control of changing utility and feasible targets.

E Proof of Theorem 2: conditional opportunity debt

The jointly stationary ergodic pair (A_k, S_k) induces the bounded stationary ergodic process $A_k - S_k$. Since $\phi_k = \mathbb{E}[S_k(t)]$, its mean is $\pi_k - \phi_k$. The ergodic theorem gives equation 13; dividing by the similarly convergent availability count gives the conditional missed-opportunity fraction $1 - r_k$. Stationarity and linearity of expectation give

$$\mathbb{E}[d_k(t)] = \sum_{s<t} \mathbb{E}[A_k(s) - S_k(s)] = (t-1)(\pi_k - \phi_k). \tag{43}$$

For the separate reactive-weight heuristic with $\lambda > 0$, divide numerator and denominator of $\widehat{p}_k(t)$ by $\lambda(t-1)$:

$$\widehat{p}_k(t) = \frac{\dfrac{\alpha_k}{\pi_k + \epsilon}\left(\dfrac{d_k(t)}{t-1} + \dfrac{1}{\lambda(t-1)}\right)}{\displaystyle\sum_{j=1}^N \frac{\alpha_j}{\pi_j + \epsilon}\left(\frac{d_j(t)}{t-1} + \frac{1}{\lambda(t-1)}\right)}. \tag{44}$$

The numerator converges to $\alpha_k(\pi_k - \phi_k)/(\pi_k + \epsilon)$ and the denominator to $Z > 0$, giving equation 14. These numerators are nonnegative but need not be positive. All but one may vanish, so the limiting weights need not lie strictly inside $(0, 1)$. If $Z = 0$, the leading-order argument gives a zero-over-zero ratio and establishes no limit. With $\lambda = 0$, direct substitution gives weights proportional to $\alpha_k/(\pi_k + \epsilon)$.

For comparison only, define the old *unconditional* debt $d_k^{\mathrm{uncond}}(t) = \sum_{s<t}(1 - S_k(s))$. Under stationary ergodic participation with mean ϕ_k, its rate is $1 - \phi_k$. Repeating the ratio argument for that separate heuristic yields numerators $\alpha_k(1 - \phi_k)/(\pi_k + \epsilon)$, provided their sum is positive. It is incorrect to replace ϕ_k by π_k unless every available client participates.

F Proof of Proposition 1: bias-aware optimization of CUP–FedAvg

Step 1: Uncached model-update representation. On a nonempty round, substituting the definition of $d_{k,t}$ into equation 16 gives $w_t - \gamma G_t = \sum_k \omega_{k,t} w_{k,t}^{\mathrm{loc}}$ because $\sum_k \omega_{k,t} = 1$. On an empty round both the algorithm and this representation retain w_t. This identity permits multiple local steps and does not require any gradient interpretation of the returned model.

Step 2: Conditional descent with bias. Put $g_t = \nabla F(w_t)$. Conditional variance decomposition gives

$$\mathbb{E}[\|G_t\|^2 \mid \mathcal{F}_t] = \|m_t\|^2 + \mathbb{E}[\|G_t - m_t\|^2 \mid \mathcal{F}_t] \le \|m_t\|^2 + \sigma^2. \tag{45}$$

Smoothness and the identity $-\langle g_t, m_t\rangle = -(\|g_t\|^2 + \|m_t\|^2 - \|m_t - g_t\|^2)/2$ yield

$$\begin{aligned}\mathbb{E}[F(w_{t+1}) \mid \mathcal{F}_t] &\le F(w_t) - \gamma\langle g_t, m_t\rangle + \frac{L\gamma^2}{2}(\|m_t\|^2 + \sigma^2)\\ &= F(w_t) - \frac{\gamma}{2}\|g_t\|^2 - \frac{\gamma}{2}(1 - L\gamma)\|m_t\|^2 + \frac{\gamma}{2}\|b_t\|^2 + \frac{L\gamma^2\sigma^2}{2}\\ &\le F(w_t) - \frac{\gamma}{2}\|g_t\|^2 + \frac{\gamma}{2}\|b_t\|^2 + \frac{L\gamma^2\sigma^2}{2}.\end{aligned} \tag{46}$$

The last inequality uses $\gamma \le 1/L$; it does not set the bias to zero.

Step 3: Telescoping. Taking expectations and summing gives

$$\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w_t)\|^2 \le F(w_0) - F_* + \frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\|b_t\|^2 + \frac{TL\gamma^2\sigma^2}{2}. \tag{47}$$

Division by $\gamma T/2$ proves equation 18; substitution gives equation 19. No independence across clients or rounds, stationary availability, or positive inclusion-probability lower bound is needed for this implication. Such generality is possible because selection and availability effects remain in the bias and variance terms. A history with no possible participation has $m_t = 0$ and $b_t = -\nabla F(w_t)$; no progress is asserted there.

Step 4: Selection and local-training components. On selected clients define $e_{k,t} = \mathbb{E}[d_{k,t} \mid \mathcal{G}_t] - \nabla F_k(w_t)$, and set $e_{k,t} = 0$ otherwise. Let $\bar{\omega}_{k,t} = \mathbb{E}[\omega_{k,t} \mid \mathcal{F}_t]$. Since weights are $\mathcal{G}_t$-measurable, the tower property gives the exact decomposition

$$\begin{aligned} b_t &= b_t^{\mathrm{weight}} + b_t^{\mathrm{local}},\\ b_t^{\mathrm{weight}} &= \sum_k (\bar{\omega}_{k,t} - a_k)\nabla F_k(w_t),\\ b_t^{\mathrm{local}} &= \mathbb{E}\left[\sum_k \omega_{k,t} e_{k,t} \,\middle|\, \mathcal{F}_t\right].\end{aligned} \tag{48}$$

In particular,

$$\|b_t\|^2 \le 2\left\|\sum_k (\bar{\omega}_{k,t} - a_k)\nabla F_k(w_t)\right\|^2 + 2\mathbb{E}\left[\sum_k \omega_{k,t}\|e_{k,t}\|^2 \,\middle|\, \mathcal{F}_t\right]. \tag{49}$$

This follows from conditional Jensen and weighted Cauchy–Schwarz, using $\sum_k \omega_{k,t} \le 1$. The first term includes availability, adaptive selection, the random FedAvg denominator, and empty rounds. Neither inverse availability nor equality of the utility products $r_k\mu_k$ implies $\bar{\omega}_{k,t} = a_k$.

For an explicit local-error bound, suppose selected client k performs $H_{k,t}$ plain-SGD steps with step sizes $\eta_{k,t,h} \ge 0$, fixed conditional on $\mathcal{G}_t$, from $w_{k,t,0} = w_t$. Assume each stochastic gradient

has conditional mean $\nabla F_k(w_{k,t,h})$ given the current local history (including $\mathcal{G}_t$), and ∇F_k is L_k-Lipschitz. Set $s_{k,t} = \sum_{h=0}^{H_{k,t}-1} \eta_{k,t,h}$. Then

$$\begin{aligned}\|e_{k,t}\| \le \left|\frac{s_{k,t}}{\gamma} - 1\right| \|\nabla F_k(w_t)\| \\ + \frac{L_k}{\gamma} \sum_{h=0}^{H_{k,t}-1} \eta_{k,t,h} \mathbb{E}[\|w_{k,t,h} - w_t\| \mid \mathcal{G}_t].\end{aligned} \tag{50}$$

To obtain this inequality, telescope the SGD updates, take conditional expectations, and add and subtract $(s_{k,t}/\gamma)\nabla F_k(w_t)$. The first term accounts for unequal cumulative local step sizes; the second is local drift. For common H, η and $\gamma = H\eta$ only the drift term remains. If additionally the conditional expected norm of every local stochastic gradient is at most M_k, then the triangle inequality gives $\mathbb{E}[\|w_{k,t,h} - w_t\| \mid \mathcal{G}_t] \le \eta h M_k$, and hence

$$\|e_{k,t}\| \le \tfrac{1}{2} L_k \eta M_k (H-1). \tag{51}$$

These are sufficient conditions, not assumptions automatically supplied by one local epoch, random reshuffling, momentum, or a generic optimizer. For those implementations the displacement-based proposition still applies under its own assumptions, but the local-error bound must be justified separately.

Step 5: A sufficient variance condition. If $\mathbb{E}[\|d_{k,t}\|^2 \mid \mathcal{G}_t] \le M^2$ on selected clients, uniformly in k, t, then

$$\mathbb{E}[\|G_t\|^2 \mid \mathcal{G}_t] \le \mathbb{E}\left[\sum_k \omega_{k,t} \|d_{k,t}\|^2 \,\middle|\, \mathcal{G}_t\right] \le M^2. \tag{52}$$

The tower property and conditional variance decomposition give

$$\mathbb{E}\big[\|G_t - m_t\|^2 \mid \mathcal{F}_t\big] \le \mathbb{E}\big[\|G_t\|^2 \mid \mathcal{F}_t\big] \le M^2. \tag{53}$$

Thus the conditional variance assumption of Proposition 1 holds with $\sigma^2 = M^2$.

Step 6: Connection to the optional cache and Theorem 3. For the surrogate-enabled update, apply the same descent argument to its actual direction $\widetilde{G}_t$, with its own conditional variance bound. Use Theorem 3's matched-weight comparator G_t^{match} in the same direction units and with exactly the same normalizer as $\widetilde{G}_t$. Write $\Delta_t = \widetilde{G}_t - G_t^{\mathrm{match}}$ and suppose that theorem's vector assumptions give $\|\Delta_t\| \le B_t$. Then

$$\begin{aligned} b_t^{\mathrm{sur}} &= \mathbb{E}[G_t^{\mathrm{match}} \mid \mathcal{F}_t] - \nabla F(w_t) + \mathbb{E}[\Delta_t \mid \mathcal{F}_t], \\ \|b_t^{\mathrm{sur}}\|^2 &\le 2\|\mathbb{E}[G_t^{\mathrm{match}} \mid \mathcal{F}_t] - \nabla F(w_t)\|^2 + 2\mathbb{E}[B_t^2 \mid \mathcal{F}_t]. \end{aligned} \tag{54}$$

Substitute this bound for the bias term in equation 18. The matched comparator generally differs from the fresh-only FedAvg aggregate because adding caches changes normalization and coverage; its bias must be accounted for separately. A bound for cached gradients cannot be silently transferred to multi-step model displacements. Lemma 3's scalar loss bound alone does not imply this vector bound. Neither confidence decay nor Theorem 3 proves that the total bias vanishes.

Relation to cumulative utility parity. Lemmas 1–2 and Theorems 1–2 describe conditional utility rates, feasible rate benchmarks, and opportunity debt under their stated assumptions. They neither impose objective-gradient alignment nor guarantee that the online scheduler attains the benchmark. Conversely, Proposition 1 does not establish equal utility rates. Cache reuse remains excluded from fresh participation and utility credits. These distinctions allow the optimization, allocation, and replacement-error results to apply to their respective parts of the same algorithm without identifying their different guarantees.

G Lemma 3: scalar surrogate replacement error

Sufficient accuracy condition for cached gradients. Fix a round and a client $k \in \mathcal{M}_t$. Suppose g_k is L_k-Lipschitz on a domain containing w_t and w_{τ_k}, and a cached gradient direction satisfies

$\|\widehat{g}_k(\tau_k) - g_k(w_{\tau_k})\| \le \varepsilon_k^{\mathrm{cache}}$. Then

$$\|\widetilde{g}_k(w_t) - g_k(w_t)\| \le \varepsilon_k^{\mathrm{cache}} + L_k\|w_t - w_{\tau_k}\| =: E_k^{\mathrm{grad}}(t). \tag{55}$$

For a prototype-derived direction, assume instead a construction-specific bound $\|H_k(w_t; \mathcal{P}_k(\tau_k)) - g_k(w_t)\| \le E_k^{\mathrm{proto}}(t)$. Let $E_k(t)$ be the corresponding bound for the chosen surrogate, so that

$$\|\widetilde{g}_k(w_t) - g_k(w_t)\| \le E_k(t). \tag{56}$$

The prototype bound is an explicit assumption, not a consequence of caching or exponential decay. The cached-gradient bound is deterministic conditional on its cache-error assumption; it does not follow merely from an unbiased stochastic gradient.

Lemma 3 (Scalar surrogate replacement error). Fix a round t. Let $\mathcal{A}_t = \mathcal{C}_t$ denote the available clients and $\mathcal{M}_t \subseteq [N] \setminus \mathcal{A}_t$ the unavailable clients represented by a surrogate. Let $F_k, \widetilde{F}_k : \mathcal{W} \to \mathbb{R}$ be scalar losses, with $|\widetilde{F}_k(w) - F_k(w)| \le \epsilon$ for every $k \in \mathcal{M}_t$ and $w \in \mathcal{W}$. Hold the nonnegative weights q_k and η_k fixed across the following two objectives:

$$\begin{aligned} F_t^{\mathrm{sur}}(w) &= \sum_{k\in\mathcal{A}_t} q_k F_k(w) + \sum_{k\in\mathcal{M}_t} \eta_k \widetilde{F}_k(w), \\ F_t^{\mathrm{true}}(w) &= \sum_{k\in\mathcal{A}_t} q_k F_k(w) + \sum_{k\in\mathcal{M}_t} \eta_k F_k(w). \end{aligned} \tag{57}$$

Set $q_k = 0$ for available but unselected clients. Here η_k denotes the effective surrogate aggregation weight: for the normalized mechanism above it is $a_k(t)\eta_k(t)/Z_t$, not the raw confidence factor alone. Then, for all $w \in \mathcal{W}$,

$$|F_t^{\mathrm{sur}}(w) - F_t^{\mathrm{true}}(w)| = \left|\sum_{k\in\mathcal{M}_t} \eta_k\big(\widetilde{F}_k(w) - F_k(w)\big)\right| \le \epsilon \sum_{k\in\mathcal{M}_t} \eta_k. \tag{58}$$

The comparator is the same-weight true-loss objective. It equals a specified full-data objective only when its client coverage and weights match that objective; otherwise, additional weighting or omission errors are outside this bound. Normalized objectives must use the same positive normalizer on both sides. Scalar loss accuracy alone does not imply accuracy of gradients or other vector updates, and an arbitrary cached update need not be the gradient of a surrogate loss. The proof follows below.

Interpretation. Replacing unavailable clients' scalar losses by uniformly accurate surrogates changes the matched-weight objective by at most the per-client error times the total surrogate weight. This bound isolates replacement error only; it does not cover missing clients, mismatched weights, or vector-update error.

Scalar replacement error. Subtracting the two objectives in equation 57 cancels their identical available-client terms. Thus $F_t^{\mathrm{sur}}(w) - F_t^{\mathrm{true}}(w) = \sum_{k\in\mathcal{M}_t} \eta_k(\widetilde{F}_k(w) - F_k(w))$. The scalar triangle inequality, nonnegativity of the weights, and the uniform scalar error assumption give

$$\left|\sum_{k\in\mathcal{M}_t} \eta_k(\widetilde{F}_k(w) - F_k(w))\right| \le \sum_{k\in\mathcal{M}_t} \eta_k|\widetilde{F}_k(w) - F_k(w)| \le \epsilon \sum_{k\in\mathcal{M}_t} \eta_k. \tag{59}$$

H THEOREM 3: VECTOR REPLACEMENT AND OBJECTIVE PERTURBATION

Theorem 3 (Confidence-weighted surrogate aggregation). Fix round t, cache times $\tau_k < t$, and staleness $\delta_k = t - \tau_k$ for $k \in \mathcal{M}_t$. Use distinct notation $g_k(w), \widetilde{g}_k(w) \in \mathbb{R}^d$ for true and surrogate vector signals. Assume $\|\widetilde{g}_k(w) - g_k(w)\| \le \epsilon$ for every $k \in \mathcal{M}_t$ and $w \in \mathcal{W}$, with $w_t \in \mathcal{W}$. More generally, the pointwise bounds $E_k(t)$ in equation 56 suffice at the current iterate. These vector accuracy assumptions are separate from Lemma 3.

Let $\beta_k(t) \ge 0$ be finite weights held fixed across both aggregates, and set them to zero for clients supplying neither a fresh signal nor a surrogate. Define

$$\widetilde{G}_t(w_t) = \sum_{k\in\mathcal{S}_t} \beta_k(t) g_k(w_t) + \sum_{k\in\mathcal{M}_t} \beta_k(t)\widetilde{g}_k(w_t), \tag{60}$$

and its same-weight true-signal comparator

$$G_t(w_t) = \sum_{k \in \mathcal{S}_t \cup \mathcal{M}_t} \beta_k(t) g_k(w_t). \tag{61}$$

The aggregate replacement error is

$$\Delta_t(w_t) = \widetilde{G}_t(w_t) - G_t(w_t) = \sum_{k \in \mathcal{M}_t} \beta_k(t)(\widetilde{g}_k(w_t) - g_k(w_t)), \tag{62}$$

and satisfies

$$\|\Delta_t(w_t)\| \le B_t := \sum_{k \in \mathcal{M}_t} \beta_k(t) E_k(t), \tag{63}$$

$$\begin{aligned} B_t &= \frac{\eta_0}{Z_t} \sum_{k \in \mathcal{M}_t} a_k(t) e^{-\lambda \delta_k} E_k(t) \quad \text{(normalized mechanism)}, \\ \|\Delta_t(w_t)\| &\le \epsilon \eta_0 \sum_{k \in \mathcal{M}_t} e^{-\lambda \delta_k} \qquad\qquad \text{if } \beta_k(t) = \eta_0 e^{-\lambda \delta_k}. \end{aligned} \tag{64}$$

The first specialization uses $Z_t > 0$, $\beta_k(t) = a_k(t)/Z_t$ for fresh clients and $\beta_k(t) = a_k(t)\eta_k(t)/Z_t$ for missing clients, recovering equation 21. The second uses the uniform-error assumption and unnormalized exponential weights; no $1/Z_t$ factor may be dropped from a normalized rule. For cached gradients, $E_k(t)$ may be taken as $E_k^{\text{grad}}(t)$ from equation 55. Under the uniform bound, take $E_k(t) = \epsilon$ to obtain $\|\Delta_t(w_t)\| \le \epsilon \sum_{k \in \mathcal{M}_t} \beta_k(t)$.

Let the scalar objective $f : \mathbb{R}^d \to \mathbb{R}$ have an L-Lipschitz gradient, with $L > 0$, and use Euclidean norms. Consider $w_{t+1} = w_t - \gamma \widetilde{G}_t(w_t)$ with $\gamma > 0$. Independently of any descent-direction assumption,

$$\begin{aligned} |f(w_t - \gamma \widetilde{G}_t(w_t)) - f(w_t - \gamma G_t(w_t))| &\le \gamma \|\nabla f(w_t)\| \, \|\Delta_t(w_t)\| \\ &\quad + L\gamma^2 \|G_t(w_t)\| \, \|\Delta_t(w_t)\| + \frac{L\gamma^2}{2} \|\Delta_t(w_t)\|^2. \end{aligned} \tag{65}$$

Substituting B_t for each error norm gives a cache- and staleness-dependent objective perturbation bound. If, additionally,

$$\langle \nabla f(w_t), \widetilde{G}_t(w_t) \rangle \ge c \|\nabla f(w_t)\| \, \|\widetilde{G}_t(w_t)\|, \qquad c \in (0, 1], \tag{66}$$

then

$$f(w_{t+1}) \le f(w_t) - \gamma c \|\nabla f(w_t)\| \, \|\widetilde{G}_t(w_t)\| + \frac{L\gamma^2}{2} \|\widetilde{G}_t(w_t)\|^2. \tag{67}$$

For nonzero gradient and aggregate, this guarantees strict descent when $0 < \gamma < 2c\|\nabla f(w_t)\|/(L\|\widetilde{G}_t(w_t)\|)$. The angle condition is not implied by exponential confidence decay. Nor does decay alone guarantee that B_t decreases: approximation error can grow with model drift, and Z_t also changes. In a surrogate-only round, common decay factors may cancel under normalization.

The comparator G_t need not equal $\nabla f(w_t)$ or a standard full-data aggregate. These results control the effect of cached representations under matched weights, not selection bias or CUP convergence. The proof follows below; empirical effects of this mechanism require surrogate-enabled evaluation.

Interpretation. The theorem propagates client-level vector approximation errors into a bound on aggregate-update error and then into a bound on the one-step objective perturbation. Exponential confidence weights can reduce a stale surrogate's direct contribution, but they do not by themselves guarantee descent, decreasing total error, or convergence to CUP.

Cached-gradient error. For $\widetilde{g}_k(w_t) = \widehat{g}_k(\tau_k)$, add and subtract the true direction at the cached model. The triangle inequality, the cache-error assumption, and Lipschitz continuity give

$$\begin{aligned} \|\widehat{g}_k(\tau_k) - g_k(w_t)\| &\le \|\widehat{g}_k(\tau_k) - g_k(w_{\tau_k})\| + \|g_k(w_{\tau_k}) - g_k(w_t)\| \\ &\le \varepsilon_k^{\text{cache}} + L_k \|w_t - w_{\tau_k}\|. \end{aligned} \tag{68}$$

This proves equation 55. Moreover, $\|w_t - w_{\tau_k}\| \le \sum_{s=\tau_k}^{t-1} \|w_{s+1} - w_s\|$, so the drift term depends on the intervening model updates, not merely the elapsed rounds. For prototype-derived directions, equation 56 is precisely the stated construction-specific assumption; it is not derived from scalar loss accuracy or from storing a prototype. The proof does not assert such an error certificate for an unspecified reconstruction map.

We prove the replacement bound, the conditional progress bound, and the objective perturbation bound separately.

Step 1: Deterministic bias bound. From the definition equation 62 and the triangle inequality,

$$\|\Delta_t(w_t)\| = \left\| \sum_{k' \in \mathcal{M}_t} \beta_{k'}(t)\big(\tilde{g}_{k'}(w_t) - g_{k'}(w_t)\big) \right\| \le \sum_{k' \in \mathcal{M}_t} \big\|\beta_{k'}(t)\big(\tilde{g}_{k'}(w_t) - g_{k'}(w_t)\big)\big\| \tag{69}$$

By positive homogeneity of norms and nonnegativity of $\beta_{k'}(t)$,

$$\big\|\beta_{k'}(t)\big(\tilde{g}_{k'}(w_t) - g_{k'}(w_t)\big)\big\| = \beta_{k'}(t)\,\|\tilde{g}_{k'}(w_t) - g_{k'}(w_t)\| \tag{70}$$

Apply the vector accuracy assumption of Theorem 3, using either $E_k(t) = \epsilon$ or the client-specific bound equation 56:

$$\|\Delta_t(w_t)\| \le \sum_{k \in \mathcal{M}_t} \beta_k(t) E_k(t) = B_t. \tag{71}$$

For the normalized mechanism, substituting $\beta_k(t) = a_k(t)\eta_0 e^{-\lambda\delta_k}/Z_t$ for $k \in \mathcal{M}_t$ proves the first specialization in equation 64. The same positive Z_t is used in both aggregates, so no additional normalization difference appears. Under uniform accuracy, take $E_k(t) = \epsilon$ to obtain $\|\Delta_t(w_t)\| \le \epsilon \sum_{k \in \mathcal{M}_t} \beta_k(t)$. Substituting the unnormalized weights $\beta_k(t) = \eta_0 e^{-\lambda\delta_k}$ gives the second specialization. If base weights $a_k(t)$ are also used in an unnormalized rule, they remain in the sum. A normalized rule must retain its $1/Z_t$ factor.

Step 2: Smoothness-based one-step progress. The L-Lipschitz-gradient assumption of Theorem 3 gives the upper half of the two-sided remainder bound equation 75, derived in Step 3 below. Apply it with $x = w_t$ and $y = w_{t+1} = w_t - \gamma\widetilde{G}_t(w_t)$:

$$f(w_{t+1}) \le f(w_t) + \Big\langle \nabla f(w_t),\, -\gamma\tilde{G}_t(w_t) \Big\rangle + \frac{L}{2}\|\gamma\tilde{G}_t(w_t)\|^2 \tag{72}$$

Rearranging gives

$$f(w_{t+1}) \le f(w_t) - \gamma\langle \nabla f(w_t),\, \tilde{G}_t(w_t)\rangle + \frac{L\gamma^2}{2}\|\tilde{G}_t(w_t)\|^2 \tag{73}$$

Using the angle condition equation 66, i.e., $\langle \nabla f(w_t), \tilde{G}_t(w_t)\rangle \ge c\|\nabla f(w_t)\|\,\|\tilde{G}_t(w_t)\|$, yields equation 67.

Step 3: Quantifying the impact of surrogate bias on the update. Define the two candidate post-update points

$$w_t^{\mathrm{sur}} := w_t - \gamma\tilde{G}_t(w_t), \qquad w_t^{\mathrm{true}} := w_t - \gamma G_t(w_t) \tag{74}$$

so that $w_t^{\mathrm{sur}} - w_t^{\mathrm{true}} = -\gamma(\tilde{G}_t(w_t) - G_t(w_t)) = -\gamma\Delta_t(w_t)$. For arbitrary $x, y \in \mathbb{R}^d$, set $h = y - x$. The fundamental theorem of calculus and the L-Lipschitz-gradient assumption give

$$\begin{aligned} \big|f(y) - f(x) - \langle \nabla f(x), h\rangle\big| &= \left| \int_0^1 \langle \nabla f(x + sh) - \nabla f(x), h\rangle\, ds \right| \\ &\le \int_0^1 Ls\|h\|^2\, ds = \frac{L}{2}\|h\|^2. \end{aligned} \tag{75}$$

This proves the two-sided remainder bound equation 75; it does not take absolute values of a one-sided inequality. By the triangle inequality and Cauchy–Schwarz, it follows that

$$|f(y) - f(x)| \le \|\nabla f(x)\|\,\|y - x\| + \frac{L}{2}\|y - x\|^2. \tag{76}$$

Apply this bound with $x = w_t^{\text{true}}$ and $y = w_t^{\text{sur}}$. Lipschitz continuity of the gradient also yields

$$\|\nabla f(w_t^{\text{true}})\| \le \|\nabla f(w_t)\| + L\|w_t^{\text{true}} - w_t\| = \|\nabla f(w_t)\| + L\gamma\|G_t(w_t)\| \tag{77}$$

Therefore,

$$\begin{aligned}\left|f(w_t^{\text{sur}}) - f(w_t^{\text{true}})\right| &\le \|\nabla f(w_t^{\text{true}})\|\,\|w_t^{\text{sur}} - w_t^{\text{true}}\| + \frac{L}{2}\|w_t^{\text{sur}} - w_t^{\text{true}}\|^2 \\ &\le \big(\|\nabla f(w_t)\| + L\gamma\|G_t(w_t)\|\big)\cdot\gamma\|\Delta_t(w_t)\| + \frac{L}{2}\gamma^2\|\Delta_t(w_t)\|^2 \\ &= \gamma\,\|\nabla f(w_t)\|\,\|\Delta_t(w_t)\| \;+\; L\gamma^2\,\|G_t(w_t)\|\,\|\Delta_t(w_t)\| \;+\; \frac{L\gamma^2}{2}\|\Delta_t(w_t)\|^2\end{aligned} \tag{78}$$

which proves equation 65. Since all coefficients on its right-hand side are nonnegative, replacing each $\|\Delta_t(w_t)\|$ by B_t gives the claimed objective perturbation bound. For cached gradients this is explicit in $\varepsilon_k^{\text{cache}}$, $L_k\|w_t - w_{\tau_k}\|$, the confidence weights, and Z_t. This comparison requires no angle assumption; that assumption is used only for the conditional progress result in Step 2.