WHEN DO STALE UPDATES HELP? FORECASTING, PARTICIPATION WEIGHTING, AND STEP SCALE IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Stale client updates can reduce the variance of federated aggregation, but accuracy gains do not by themselves identify the value of forecasting: participation weighting and aggregate scale can change simultaneously. We analyze these effects through a forecast–residual estimator and use OPTICAST, which learns client-specific linear forecasts, as a case study. Exact conditional moments expose the coupling between propensity error and forecast residuals. A complementary decomposition separates aggregate weight mass from relative client weighting, and an aggregation-aware damping criterion identifies the optimal scalar correction at a frozen history. Across the four main image-classification settings, OPTICAST achieves higher mean accuracy than FedAU, MIFA, U-FedVARP, and FedStale. The control experiments qualify this result. Across three CIFAR-10 participation regimes, last-update reuse yields paired gains of 0.14, 1.06, and 3.51 percentage points over zero forecasting as participation decreases, whereas a separate five-seed test does not establish an advantage of learned gains over simple reuse. On FEMNIST, a one-seed oracle scale intervention recovers 14.34 points of a 16.54-point zero-control-to-participating-mean gap. Learned and damped configurations also collapse on Shakespeare. These findings distinguish configured-method performance from evidence for predictor complexity, and motivate evaluating forecasting jointly with participation weights, aggregate scale, implementation consistency, and stability.

1 INTRODUCTION

Federated learning aggregates information from clients whose participation is incomplete and heterogeneous (McMahan et al., 2017). When participation probabilities differ, averaging only the updates that arrive need not follow the intended population objective. Two prominent responses address different aspects of this problem: participation-aware reweighting adjusts the influence of observed clients, while update caching substitutes information for absent clients. FedAU exemplifies the former; MIFA, U-FedVARP, and FedStale use different forms of cached information (Wang & Ji, 2024; Gu et al., 2021; Jhunjhunwala et al., 2022; Rodio & Neglia, 2024).

Complete-method comparisons nevertheless do not isolate the contribution of forecasting. A participation estimator changes both transient errors and clipped inverse weights. Adding a forecast then changes variance and, with imperfect weights, the conditional mean. Equal server learning rates do not equalize the expected total weight of observed updates. An apparent forecasting benefit can therefore reflect compensation for attenuation in the reference update, rather than more informative prediction.

We investigate *when stale-update forecasting can be distinguished from participation weighting and aggregate scale*. The common estimator adds a population forecast to inverse-weighted residuals from reporting clients. This construction is established: at known propensities, last-update and damped-update instances recover U-FedVARP and FedStale. Our analysis distinguishes the roles of historical prediction, client representation, and aggregate scale.

The analysis separates *forecast residuals*, *relative client weights*, and *aggregate weight mass*. Exact conditional moments connect residuals to participation weighting. A mass–weighting decomposition then shows why restoring aggregate scale need not restore the population objective. The experiments

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.

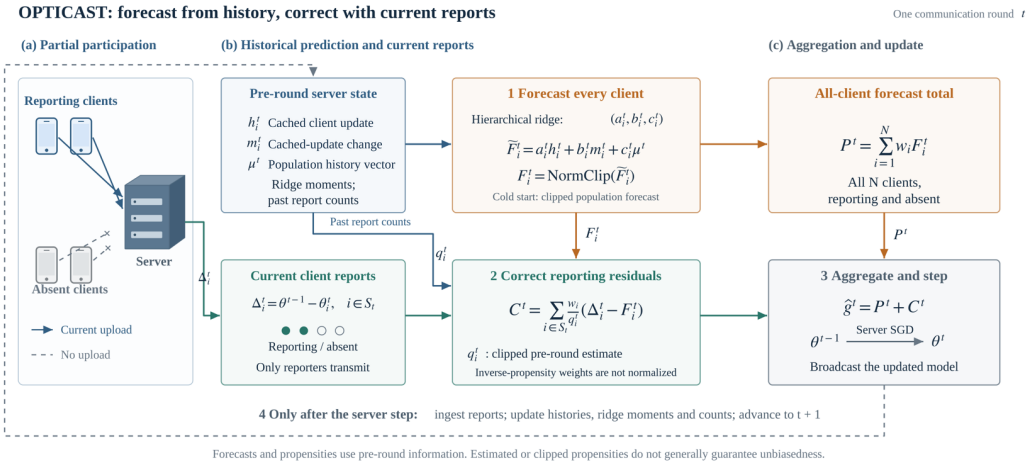


Figure 1: OPTICAST in one communication round. (a) Only reporting clients upload current updates. (b) The server forecasts every client from cached client updates h_i^t, cached-update changes m_i^t, and the population history vector μ^t; past report counts determine q_i^t. (c) The all-client forecast total and reporting-client residual corrections form the server update. Histories and ridge moments are updated only after the server step. Forecast inputs and the cold-start rule are defined in Section 2 and Appendix B. Estimated or clipped propensities do not generally preserve unbiasedness.

ask whether forecasting helps at fixed aggregation settings, whether learning gains improves on simple reuse, and whether conclusions survive normalization and scale controls. Sparse CIFAR-10 supports simple reuse; CIFAR-100 shows small forecast-only gains despite positive complete-method gains; FEMNIST and Shakespeare expose scale sensitivity and instability.

We make three contributions. First, we separate aggregate mass from relative weighting and derive an exact one-step criterion for damped stale reuse to reduce aggregation error. Second, we connect the moments and a forecast-error-dependent convergence bound to frozen-state checks, including interior damping optima. Third, paired empirical comparisons across four tasks characterize forecasting, learned complexity, normalization, and stability. The scope of single-factor attribution is assessed against the available implementation evidence.

2 A COMMON VIEW OF PARTICIPATION-AWARE AGGREGATION

Problem setting. Let $f(\theta) = \sum_{i=1}^{N} w_i f_i(\theta)$, where $w_i \geq 0$ and $\sum_i w_i = 1$. In round t, client i would return the local model difference $\Delta_i^t = \theta^{t-1} - \theta_i^t$. Its report indicator I_i^t is Bernoulli with probability $p_i > 0$. We assume independent indicators across clients and rounds, independent of local training randomness. The server history $\mathcal{H}_{t-1}$ contains information available before the round. The full-participation update is $\sum_i w_i \Delta_i^t$; it is a model difference, not generally an exact gradient after multiple local steps.

A forecast F_i^t and a propensity estimate $q_i^t > 0$ are computed from $\mathcal{H}_{t-1}$. We study

$$\hat{g}_F^t = \sum_i w_i F_i^t + \sum_i I_i^t \frac{w_i}{q_i^t} (\Delta_i^t - F_i^t), \qquad \theta^t = \theta^{t-1} - \eta_s \hat{g}_F^t. \tag{1}$$

This is the augmented Horvitz–Thompson form (Horvitz & Thompson, 1952; Robins et al., 1994). Setting $F_i^t = 0$ gives inverse-propensity aggregation; setting $F_i^t = h_i^t$, the last received update, gives a cached-update correction. History measurability is essential: fitting the forecast to the current report before evaluating its residual changes the estimator being analyzed.

Closest methods. Table 1 places the five relevant methods in a common residual-gain form, obtained by replacing $1/q_i$ in Eq. (1) by ρ_i. MIFA uses memory without inverse-propensity correction (Gu et al., 2021). U-FedVARP extends the variance-reduction construction of FedVARP to unequal

Table 1: Closest methods in $\sum_i w_i F_i + \sum_{i \in S} w_i \rho_i (\Delta_i - F_i)$. The taxonomy describes aggregation forms, not an empirical ranking. Known-propensity forms are distinguished from estimated-propensity adaptations. h_i is the last received update.

Method	Forecast F_i	Residual gain ρ_i
FedAU	0	Online interval weight ω_i
MIFA	h_i	1
U-FedVARP	h_i	$1/p_i$
FedStale	αh_i	$1/p_i$
FedSteer	Cached-subspace reconstruction	$1/p_i$

participation, and FedStale interpolates through a damping parameter (Jhunjhunwala et al., 2022; Rodio & Neglia, 2024). These relationships concern aggregation algebra rather than implementations with different storage, clipping, initialization, or optimization.

Recent subspace-based correction. FedSteer (Zhang et al., 2026) reconstructs stale client updates using cached projection coordinates and an evolving gradient subspace. OPTICAST instead fits client-specific gains on a cached client update, its change, and a population history vector. Our analysis focuses on the interaction of forecast residuals with estimated participation weights and aggregate scale. Applying the moment identities to a particular implementation requires pre-round measurability of its forecast.

Controlled configurations. The count-based configurations estimate participation through

$$q_i^t = \min\left\{1, \max\left\{\underline{q}, \frac{n_i^t + 1/2}{t}\right\}\right\}, \qquad n_i^t = \sum_{s<t} I_i^s, \qquad \underline{q} = 0.05. \tag{2}$$

The zero-forecast control, last-update forecast, moving average, and OPTICAST share the count-based estimator. Figure 1 shows the information flow of OPTICAST, whose linear forecast combines three internal inputs: the cached client update h_i, the cached-update change m_i (the difference of the two latest received updates), and the population history vector μ (an exponential moving average of the participating weighted-mean update). The resulting forecast is $F_i = a_i h_i + b_i m_i + c_i \mu$ before norm clipping. Per-client gains are fitted by hierarchical ridge regression on historical inner products. This learned linear predictor provides a concrete instance of the estimator family. Appendix B specifies fitting, clipping, and the different cold-start rules.

Two further controls answer questions that the named competitors alone cannot. The *participating mean*, $\sum_{i \in S} w_i \Delta_i / \sum_{i \in S} w_i$, tests whether forecasting is needed at all under the evaluated protocol. A residual-normalized, or Hájek-type, estimator (Hájek, 1971) divides the residual sum in Eq. (1) by $\sum_{i \in S} w_i / q_i$. It probes normalization at a fixed forecast, rather than adding another algorithmic competitor.

3 WHAT FORECASTING CHANGES

We suppress the round index and condition on the pre-round history. Let $d_i = \mathbb{E}[\Delta_i \mid \mathcal{H}]$, $\sigma_i^2 = \mathbb{E}[\|\Delta_i - d_i\|^2 \mid \mathcal{H}]$, and $r_i = d_i - F_i$. Assume conditional independence of the client updates and square integrability.

Proposition 1 (Conditional moments)**.** *For history-measurable F_i and $q_i > 0$, independent Bernoulli participation gives*

$$\mathbb{E}[\hat{g}_F \mid \mathcal{H}] - \sum_i w_i d_i = \sum_i w_i \left(\frac{p_i}{q_i} - 1\right) r_i, \tag{3}$$

$$\operatorname{tr} \operatorname{Var}(\hat{g}_F \mid \mathcal{H}) = \sum_i w_i^2 \frac{p_i}{q_i^2} \left[(1 - p_i)\|r_i\|^2 + \sigma_i^2\right]. \tag{4}$$

The proof is a conditional expectation and variance calculation, given in Appendix A. At $q_i = p_i$, every forecast has the same conditional mean, and forecasting acts through variance. With estimated

propensities, the mean also depends on the forecast residual. In particular,

$$\mathbb{E}[\hat{g}_F - \hat{g}_0 \mid \mathcal{H}] = \sum_i w_i \left(1 - \frac{p_i}{q_i}\right) F_i. \tag{5}$$

Thus a forecast-only intervention in code need not be a variance-only intervention statistically. The two error factors in Eq. (3) need not be independent. Although an exact conditional-mean forecast makes the bias zero, the identities do not prove consistency of the learned forecast or statistical double robustness under arbitrary availability.

Clipping has a persistent effect. Under the Bernoulli model, Eq. (2) converges almost surely to $\max\{p_i, \underline{q}\}$. For $p_i < \underline{q}$, the residual coefficient p_i/q_i tends to $p_i/\underline{q} < 1$. The limiting contribution is therefore pulled toward the forecast, or toward zero when no forecast is supplied. This is not a vanishing estimation error. Conversely, a client's true $p_i \le \underline{q}$ does not imply that its finite-sample estimate equals the floor in every round.

Proposition 2 (Aggregate mass and relative client weighting)**.** *Define $\varpi_i = p_i/q_i$, $c = \sum_i w_i \varpi_i > 0$, and $\widetilde{w}_i = w_i \varpi_i / c$. Then*

$$\mathbb{E}[\hat{g}_0 \mid \mathcal{H}] = c \sum_i \widetilde{w}_i d_i, \tag{6}$$

$$\mathbb{E}[\hat{g}_F \mid \mathcal{H}] = c \sum_i \widetilde{w}_i d_i + \sum_i w_i (1 - \varpi_i) F_i. \tag{7}$$

Multiplying $\hat{g}_0$ by $1/c$ restores its expected aggregate weight mass. It restores the full-participation mean for every *collection of client updates if and only if ϖ_i is constant on clients with positive w_i.*

This decomposition separates a scalar attenuation from altered client representation. It does not identify c with the realized update norm, and rescaling does not equate stochastic training trajectories. When all $q_i = \underline{q}$, $c = \sum_i w_i p_i / \underline{q}$ and $\widetilde{w}_i = w_i p_i / \sum_j w_j p_j$. The mass-corrected zero-forecast estimator therefore still overrepresents frequent participants relative to the original w_i. A normalized participating mean also generally differs from the population target; an expectation of a random ratio is not a ratio of expectations.

An analytic example. Take two equally weighted clients with $p_1 = 0.01$, $p_2 = 0.05$, and $q_1 = q_2 = 0.05$. For losses $f_1(\theta) = \frac{1}{2}(\theta - 1)^2$ and $f_2(\theta) = \frac{1}{2}(\theta + 1)^2$, the intended minimizer is 0. Here $c = 0.6$ and $\widetilde{w} = (1/6, 5/6)$. With exact-gradient updates, the expected zero-forecast direction, even after mass correction, vanishes at $-2/3$. A server learning-rate change corrects speed but not this displaced stationary point. This is an algebraic illustration, not an additional experiment or a claim that every empirical gap is caused by objective displacement.

Forecast fitting should reflect aggregation risk. The relevant prediction loss is not generally an unweighted per-client regression error. For a common damping factor $F_i = \alpha h_i$, define $u_i = p_i/q_i - 1$, $v = \sum_i w_i u_i h_i$, $b_0 = \sum_i w_i u_i d_i$, and $\nu_i = w_i^2 p_i (1 - p_i)/q_i^2$. Write $A = \langle b_0, v \rangle + \sum_i \nu_i \langle d_i, h_i \rangle$ and $D = \|v\|^2 + \sum_i \nu_i \|h_i\|^2$. Expanding the moments gives:

Corollary 1 (Aggregation-aware damping)**.** *Let $\mathcal{R}(\alpha) = \mathbb{E}[\|\hat{g}_{\alpha h} - \sum_i w_i d_i\|^2 \mid \mathcal{H}]$. Then*

$$\mathcal{R}(\alpha) - \mathcal{R}(0) = D\alpha^2 - 2A\alpha, \qquad \alpha_\star = \mathrm{clip}_{[0,1]}(A/D) \quad (D > 0). \tag{8}$$

For $D > 0$, positive damping improves on zero forecasting exactly when $A > 0$ and $0 < \alpha < 2A/D$. Full stale reuse improves exactly when $2A > D$.

At known propensities the bias term disappears, but the weights remain $w_i^2(1-p_i)/p_i$. With estimated propensities, cross-client bias also matters. If $D = 0$, $A = 0$ and the risk is constant in α. This criterion requires unknown current mean updates: it is an oracle diagnostic, not a deployable tuning rule or a guarantee about training accuracy. Section 5.4 evaluates it on the archived checkpoint; Appendix A.3 proves it.

Table 2: Evaluation settings. The endpoint is the mean of the last five evaluations within each run; the seed, not the evaluation, is the replication unit. CIFAR tasks evaluate held-out clients; the other tasks evaluate local test data of training clients.

Task	Training clients	Rounds	Participation interval	Seeds
CIFAR-10, dense	170	200	$[0.50, 0.90]$	3
CIFAR-10, moderate	170	200	$[0.10, 0.50]$	5
CIFAR-10, sparse	170	200	$[0.02, 0.15]$	5
CIFAR-100	170	200	$[0.10, 0.50]$	5
FEMNIST	3597	300	$[0.01, 0.05]$	3
Shakespeare	858	300	$[0.01, 0.05]$	3

Connection to optimization. A convergence bound makes the same distinction. For K local SGD steps, define $\widetilde{\eta} = \eta_s \eta_c K$, $R_i^t = \mathbb{E}\|\Delta_i^t - F_i^t\|^2$, and

$$V^t = \frac{\sum_i w_i^2 (1 - p_i) R_i^t / p_i}{(\eta_c K)^2}, \qquad B^t = \sum_i w_i \mathbb{E}\left[\left(\frac{p_i}{q_i^t} - 1\right)^2 \|\Delta_i^t - F_i^t\|^2\right]. \tag{9}$$

For smooth losses, bounded gradient noise and heterogeneity, and sufficiently small step sizes, the average squared gradient norm is bounded by an initial-gap term, local-noise and drift terms, a term proportional to $\widetilde{\eta} T^{-1} \sum_t V^t$, and a term proportional to $(\eta_c K)^{-2} T^{-1} \sum_t B^t$. Appendix A.4 gives a complete statement and proof. At known propensities the bias term vanishes. With clipping it need not vanish, even as the step size decreases. The bound is conditional on forecast errors; it does not establish stability of OPTICAST or convergence under the epoch-based, weight-decayed implementation used in the reported experiments.

4 EXPERIMENTAL DESIGN

We evaluate four tasks, summarized in Table 2. CIFAR-10 and CIFAR-100 use a label-Dirichlet partition with concentration 0.3, 170 training clients, and 30 held-out clients. FEMNIST and Shakespeare use writer and speaking-role clients, respectively. Data provenance follows CIFAR, LEAF, and federated visual-classification protocols (Krizhevsky, 2009; Caldas et al., 2018; Hsu et al., 2019). Appendix C gives local training and evaluation details.

Probabilities are spaced evenly by client index. At each seed, methods share the participation matrix. This is a controlled model of unequal availability, not a trace-based experiment or a deliberate correlation between availability and labels. Lowering the CIFAR-10 intervals changes both average participation and the distribution of inter-report gaps. It is distinct from holding one group always active while reducing another group's probability, as in experiments on participation heterogeneity (Rodio & Neglia, 2024).

The named comparison includes FedAU, MIFA, U-FedVARP, and FedStale. FedAU uses its interval estimator; the reported U-FedVARP and FedStale runs also use that estimator rather than oracle propensities. FedStale's primary setting is $\alpha = 0.5$. The record additionally includes FedAU with a 50-round interval cap and limited damping sweeps. FedSteer is discussed as recent related work but is not evaluated under the present protocol.

Client and server learning rates are fixed per task, not tuned per method. The reference and shared-code paths differ in storage precision and forecast clipping. Several exploratory arms also span source revisions, including the forecast–propensity factorial. These are comparisons of configured implementations rather than a tuned ranking or fully validated single-revision interventions. Controlled-input production tests pass, but do not identify the source of the historical training difference (Appendix F.1). We therefore distinguish paired accuracy differences from causal attribution to one component.

Statistical conventions. Each run contributes its sample-weighted mean accuracy over the final five scheduled evaluations; the seed is the replication unit. Exploratory seed sets are drawn from 12–16 for CIFAR and 12–14 for FEMNIST and Shakespeare. Two prospectively specified CIFAR-10 tests use disjoint seeds 17–21 and 22–26 (Appendix D.2). All other analyses are exploratory. Confidence

Table 3: Test accuracy (%; mean ± standard deviation of the last-five-evaluation endpoint). Each column uses a common seed set: five for CIFAR-10/100 and three for FEMNIST. CIFAR tasks evaluate held-out clients; FEMNIST evaluates local test data of training clients. The configurations use fixed task-level learning rates, not per-method tuning. Bold marks the largest displayed mean, not a significance test.

	CIFAR-10 Moderate	CIFAR-10 Sparse	CIFAR-100 Moderate	FEMNIST Sparse
FedAU	49.23 ± 1.10	42.16 ± 1.80	22.60 ± 0.54	46.76 ± 2.04
MIFA	49.14 ± 1.08	44.75 ± 0.74	22.93 ± 0.65	49.00 ± 2.58
U-FedVARP	49.94 ± 0.44	45.23 ± 1.39	23.05 ± 0.64	49.84 ± 2.51
FedStale	50.09 ± 0.81	45.30 ± 1.18	22.92 ± 0.58	48.51 ± 2.35
OPTICAST	$\mathbf{50.69 \pm 1.10}$	$\mathbf{48.72 \pm 0.75}$	$\mathbf{23.44 \pm 0.68}$	$\mathbf{52.57 \pm 2.55}$

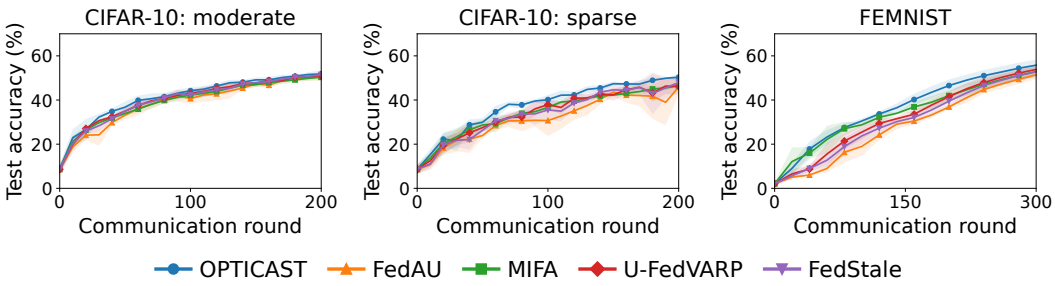


Figure 2: Test accuracy over the recorded training horizon. Curves show the across-seed mean and shaded regions one sample standard deviation on common recorded evaluation rounds (five seeds per CIFAR-10 setting; three for FEMNIST). Lines connect observations without smoothing or imputation. CIFAR-10 evaluates held-out clients; FEMNIST evaluates local test data of training clients. Tables summarize the final five evaluations, not the final plotted point.

intervals use paired run endpoints, not pooled evaluations. Reanalysis of archived records introduces no new training replications.

5 RESULTS

5.1 COMPARISON WITH PARTICIPATION-AWARE BASELINES

Table 3 reports test accuracy for OPTICAST and the four evaluated participation-aware baselines. OPTICAST has the highest mean among these configurations on the four image-classification settings; this ordering is descriptive, not a claim that every pairwise difference is statistically significant. Under moderate CIFAR-10 participation, OPTICAST exceeds FedAU by 1.46 percentage points; under sparse participation, the difference is 6.56 points. However, the zero-forecast control itself improves over FedAU by 3.24 points in the sparse setting. The full difference from FedAU cannot therefore be assigned to prediction alone.

Figure 2 shows the corresponding recorded training trajectories. On CIFAR-100, positive complete-method gains coexist with small forecast increments. The held-out accuracies are $22.60 \pm 0.54\%$ for FedAU, $23.21 \pm 0.51\%$ for zero forecasting, $23.32 \pm 0.53\%$ for last-update forecasting, and $23.44 \pm 0.68\%$ for OPTICAST. Against the same zero-forecast control, the paired gains are only 0.11 ± 0.33 and 0.22 ± 0.37 points for last-update and learned forecasts. The learned-minus-last mean is approximately 0.1 points, positive at three of five seeds. These exploratory results do not extend the registered CIFAR-10 test to another dataset. Appendix D.4 provides the per-seed values and absolute endpoints.

Table 4: Forecasting and aggregation ablations. Endpoint definition and matched seed sets are the same as in Table 3. The shared-code configurations use count-based propensities with floor 0.05; the participating mean uses no inverse propensities.

	CIFAR-10 Moderate	CIFAR-10 Sparse	CIFAR-100 Moderate	FEMNIST Sparse
Zero forecast	49.80 ± 1.29	45.39 ± 1.92	23.21 ± 0.51	40.10 ± 2.10
Last-update forecast	50.86 ± 1.04	48.91 ± 1.07	23.32 ± 0.53	50.82 ± 2.31
OPTICAST	50.69 ± 1.10	48.72 ± 0.75	23.44 ± 0.68	52.57 ± 2.55
Participating mean	49.21 ± 1.20	46.25 ± 1.85	23.01 ± 0.66	55.68 ± 2.05

Table 5: Prospectively specified moderate-participation CIFAR-10 comparisons. Differences are percentage points. The first test is one-sided; the others are two-sided. The registered test directions are retained; statistics are independently recomputed from the exported unrounded endpoints.

Comparison	Seeds	Mean $\pm$ SD	95% interval	p
OPTICAST $-$ FedAU	17–21	1.76 ± 1.24	$[0.57, \infty)$	0.017
OPTICAST $-$ last update	22–26	-0.13 ± 0.44	$[-0.67, 0.42]$	0.56
OPTICAST $-$ moving average	22–26	-0.10 ± 0.54	$[-0.77, 0.56]$	0.69

5.2 FORECASTING AND AGGREGATION ABLATIONS

Table 4 shows that OPTICAST improves on zero forecasting across the four image-classification settings. Simple cached-update reuse attains slightly higher mean accuracy on CIFAR-10, whereas OPTICAST attains higher means on CIFAR-100 and FEMNIST. The stronger participating mean on FEMNIST motivates the subsequent normalization and scale analysis. These orderings distinguish complete-configuration performance from evidence that learned gains are necessary.

The count-based configurations share the nominal estimator, floor, and task-level settings; their paired differences quantify the observed forecast contrast, subject to the source-revision limitation in Section 4. Last-update reuse improves over zero forecasting by 0.14 ± 0.25, 1.06 ± 0.41, and 3.51 ± 1.18 points in dense, moderate, and sparse CIFAR-10, respectively. The corresponding OPTICAST gains are 0.34 ± 0.25, 0.89 ± 0.63, and 3.33 ± 1.29. The observed forecasting advantage increases as reporting becomes less frequent within this task (Figure 3), but these comparisons do not establish that learned gains improve on storing the last update.

The two prospectively specified tests make this distinction explicit (Table 5). The first compares complete configurations: OPTICAST has higher endpoint accuracy than FedAU at all five seeds. The second holds the count-based estimator fixed and changes the forecast. Both confidence intervals include zero. These comparisons bound the plausible mean advantage within that setting but do not prove equality, non-inferiority, or a universal absence of benefit.

A four-cell comparison also crossed zero versus last-update forecasting with count versus *clipped interval* estimation in the shared implementation. Its reported interaction is 0.07 points, with a 95% interval of $[-0.44, 0.58]$. This interval is not contained in the predeclared ± 0.5-point margin. More importantly, the unresolved implementation audit prevents treating the interval branch as a verified FedAU control. We retain the factorial as provisional supplementary evidence, not as an attribution or a claim of negligible interaction (Appendix F.1).

5.3 A BINDING FLOOR CHANGES THE INTERPRETATION OF LARGE GAINS

On FEMNIST, where all true probabilities are at or below 0.05, last-update forecasting improves over its zero-forecast control by 10.72 ± 0.70 points and OPTICAST by 12.47 ± 1.33. These are configured forecast contrasts, but Eq. (5) shows that they include a change in expected update direction and magnitude. They cannot be read exclusively as benefits of lower prediction error.

The scale intervention addresses a narrower question. At FEMNIST seed 12, only the zero-forecast aggregate is multiplied by $s = q / \sum_i w_i p_i$; server weight decay and update order are unchanged. The factor uses simulator probabilities and is therefore an *oracle diagnostic*. With estimated propensities,

rescaling raises accuracy from 38.27% to 52.61%, against 54.81% for the participating mean. It recovers 14.34 of the 16.54-point gap, leaving 2.20 points. With true probabilities still clipped at 0.05, recovery is 100.27%. Both interventions use one seed; Table 14 gives the exact paired calculations.

The intervention measures the zero-control-to-mean gap, not the fraction of a forecasting method's gain due to scale. Scalar correction also leaves relative weights distorted (Proposition 2). Finite-sample estimates are not all at the floor: the recorded diagnostics give 80% of reporting clients' raw estimates at or below it in the second half of FEMNIST training. The average applied weight mass is 0.56 for the count-based control and 0.60 for the true-but-clipped control, compared with one by construction for the participating mean. Thus the constant multiplier exactly corrects expected mass in the latter arm, not every realized step or the former arm's transient weights.

A separate intervention changes the floor on sparse CIFAR-10 from 0.05 to 0.15, holding the task and participation probabilities fixed. The zero-forecast configuration moves from $+3.24$ to -1.25 points relative to FedAU, while the last-update contrast against zero forecasting increases from 3.51 to 6.44 points. Yet the participating mean still trails the last-update configuration by 1.10 points on average. Clipping reproduces part, but not all, of the FEMNIST ordering. Dataset differences remain relevant, and the higher floor is a mechanism probe rather than a recommended hyperparameter.

Normalization changes the outcome as well. On FEMNIST, residual normalization improves over unnormalized HT by 4.61 points with the last-update forecast and by 2.92 points with learned gains, with the improvement present at every matched seed. The resulting gains over FedAU, 8.67 and 8.72 points, are close to but below the participating mean's 8.92. On moderate CIFAR-10, the same normalization changes mean accuracy by only about 0.05 points. This contrast motivates evaluating normalization in the clipping regime rather than treating it as an interchangeable implementation detail.

5.4 Frozen-state validation

At a frozen CIFAR-10 checkpoint, predicted-to-measured variance ratios range from 0.992 to 1.01 over 10,000 sampled masks. At known propensities, measured normalized trace variance falls from 1.14 without a forecast to 0.212 with last-update reuse and 0.186 with learned gains. The archived Gram matrix also permits direct evaluation of Corollary 1: the optimal scalar gains are 0.883909, 0.875657, and 0.899160 under known, estimated, and synthetic below-floor propensities, respectively. Each agrees with a fine-grid minimizer to within 3.0×10^{-7} (Appendix E). The risk-minimizing gain is interior in all three regimes at this checkpoint; this is not an accuracy advantage established by retraining.

The three-estimator comparison further limits what this diagnostic can establish. Under synthetic below-floor participation, the normalized one-step root mean squared errors with learned gains are 0.77 for HT, 1.27 for Hájek, and 1.79 for the mixed mean. This frozen CIFAR-10 ordering differs from the FEMNIST training ordering, but the tasks and states differ too: it is not a controlled demonstration of a ranking reversal along one trajectory. Frozen-state checks establish estimator algebra, not a general ranking of training accuracy. Current local-update noise is zero during mask resampling, and neither cold-start forecasting nor the last-update norm clip is exercised at this checkpoint.

5.5 Stability and computational cost

Shakespeare limits any broad recommendation for the learned configuration. OPTICAST collapses at two of three seeds, as does the reported FedStale configuration, whereas the count-based zero-forecast and last-update configurations collapse at none of those seeds. The respective gaps for OPTICAST, FedStale, and last-update forecasting are -25.94 ± 24.13, -21.22 ± 19.63, and $+1.28 \pm 1.65$ points relative to FedAU. A trust-weight variant avoids the observed collapse but has median late-training weight 0.0014, effectively suppressing forecasting. It is not evidence that the original learned predictor has been made reliably useful. The cause of the failures is unresolved.

On FEMNIST, reported end-to-end times are 108.7 minutes for FedAU, 112.9 for U-FedVARP, 162.4 for the shared last-update configuration, and 176.1 for OPTICAST, each averaged over three runs. OPTICAST is therefore about 62% slower than FedAU in this implementation. Its 15.1 GiB aggregation-state budget is comparable in bytes to U-FedVARP's 15.3 GiB, but uses different

quantization and history lengths. These are implementation measurements on one machine, not evidence of a negligible systems cost.

6 Discussion and conclusion

Forecasting, participation weighting, and aggregate scale are distinct design choices with interacting effects. At known propensities, reducing participation-weighted forecast error reduces variance without changing the conditional mean. With estimated or clipped propensities, the same forecast also changes the mean. Restoring aggregate weight mass removes one scalar distortion but can leave relative client weighting incorrect. This distinction explains why a large gain over a weak zero-forecast control does not establish a correspondingly large gain from prediction quality.

The strongest empirical conclusion is correspondingly conditional. Simple historical forecasts help under the evaluated sparse CIFAR-10 protocol, while additional learned complexity is not established by the separate forecast-only comparison. In the FEMNIST regime with all true probabilities at or below the floor, normalization and scale are central to interpreting performance. Instability on Shakespeare prevents an unconditional recommendation of the learned or damped configurations.

The study has important limits. Source-revision differences and controlled-input tests do not establish equivalence of historical training trajectories; the exploratory ablations are therefore not fully validated single-factor interventions. Shared learning rates do not establish a tuned ranking. Finally, participation is independent and index-assigned; the single-seed scale diagnostic and the fixed-step theory do not establish generality under realistic availability traces or the implemented epoch-based dynamics. The learned forecast also changes the cold-start rule, so its FEMNIST gain over last-update reuse cannot be assigned to fitted gains alone.

For evaluating stale-update methods, the resulting standard is concrete: compare against the same estimator without a forecast, retain a simple last-update forecast, report normalization and effective weight mass, and account for failures and server cost.

AI use statement

Generative AI tools assisted with formulation of mathematical claims, figure design, numerical presentation, and manuscript drafting. The authors are responsible for the final scientific content and its verification.

Ethics statement

The evaluated methods retain client-specific model updates on the server and therefore require access beyond an interface that reveals only securely aggregated updates. No privacy guarantee is claimed. Independent, index-assigned availability does not support conclusions about demographic fairness or selection correlated with sensitive attributes. The experiments use established image-classification and language-modeling datasets; the limitations of client weighting, evidence selection, and unstable outcomes are reported alongside the results.

Reproducibility statement

Appendices A–C provide estimator definitions, proofs, and training configurations. Appendix D reports paired numerical results, while Appendix E describes the frozen-checkpoint moment and damping evaluations. The numerical analyses draw on an archive of 390 run endpoints and their recorded evaluation histories; the frozen-state diagnostics use the archived Gram matrix. Appendix F.1 documents the controlled-input production checks and their limitations. These checks assess implementation and numerical consistency, not reproduction of the historical training runs. No separate code or data supplement accompanies this submission. Executable source code, original event files, datasets, and model checkpoints are not included.

REFERENCES

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jaroslav Hájek. Comment on "an essay on the logical foundations of survey sampling, part one". In V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, pp. 236. Holt, Rinehart and Winston, Toronto, 1971.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 906–916, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Angelo Rodio and Giovanni Neglia. FedStale: Leveraging stale updates in federated learning. In *ECAI 2024*, 2024. doi: 10.3233/FAIA240849.

Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. In *International Conference on Learning Representations*, 2024.

Haoran Zhang, Cainã Figueiredo Pereira, Marie Siew, Xutong Liu, Carlee Joe-Wong, and Rachid El-Azouzi. FedSteer: Taming extreme gradient staleness in federated learning with corrective projections and caching. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 8090–8114, 2026.

A PROOFS AND SCOPE OF THE THEORETICAL RESULTS

A.1 CONDITIONAL MOMENTS

Proof of Proposition 1. Condition on $\mathcal{H}$ and put $Y_i = \Delta_i - F_i$. The forecast sum is fixed, $\mathbb{E}[Y_i \mid \mathcal{H}] = r_i$, and $\operatorname{tr}\operatorname{Var}(Y_i \mid \mathcal{H}) = \sigma_i^2$. Independence of I_i from current local randomness gives $\mathbb{E}[I_i Y_i \mid \mathcal{H}] = p_i r_i$. Therefore

$$\mathbb{E}[\hat{g}_F \mid \mathcal{H}] = \sum_i w_i F_i + \sum_i w_i \frac{p_i}{q_i} r_i = \sum_i w_i d_i + \sum_i w_i \left(\frac{p_i}{q_i} - 1\right) r_i.$$

For each client, $I_i^2 = I_i$, so

$$\operatorname{tr}\operatorname{Var}(I_i Y_i \mid \mathcal{H}) = p_i(\|r_i\|^2 + \sigma_i^2) - p_i^2\|r_i\|^2 = p_i(1-p_i)\|r_i\|^2 + p_i\sigma_i^2.$$

Conditional independence across clients removes cross-covariances. Multiplication by w_i^2/q_i^2 and summation yields Eq. (4). □

Equivalently, for $R_{i|\mathcal{H}} = \mathbb{E}[\|\Delta_i - F_i\|^2 \mid \mathcal{H}]$,

$$\operatorname{tr}\operatorname{Var}(\hat{g}_F \mid \mathcal{H}) = \sum_i w_i^2 \left(\frac{p_i}{q_i}\right)^2 \left[\frac{1-p_i}{p_i} R_{i|\mathcal{H}} + \sigma_i^2\right]. \tag{10}$$

At $q_i = p_i$, the second term is the local-update noise that would remain under full participation. The participation penalty weights squared forecast errors by $(1-p_i)/p_i$. A history-based forecast cannot remove current-round local noise. The conditional-mean forecast $F_i = d_i$ achieves zero bias and the minimum variance, but this is an oracle statement rather than a property established for a fitted predictor.

A.2 MASS, RELATIVE WEIGHTS, AND THE LIMITING OBJECTIVE

Proof of Proposition 2. By Eq. (3), $\mathbb{E}[\hat{g}_0 \mid \mathcal{H}] = \sum_i w_i \varpi_i d_i$. Since $w_i \varpi_i = c\widetilde{w}_i$, Eq. (6) follows. Adding the forecast contribution of Eq. (5) proves Eq. (7). After division by c, the conditional mean equals $\sum_i \widetilde{w}_i d_i$. This equals $\sum_i w_i d_i$ for every collection of d_i if and only if each positive-weight coefficient satisfies $\widetilde{w}_i = w_i$. To see necessity, choose all $d_j = 0$ except one arbitrary nonzero d_i. The condition is then $\varpi_i = c$ for all positive-weight clients. Constancy of ϖ_i also suffices because the weights sum to one. □

The residual error after oracle mass correction is exactly

$$\mathbb{E}[\hat{g}_0/c \mid \mathcal{H}] - \sum_i w_i d_i = \sum_i (\widetilde{w}_i - w_i) d_i. \tag{11}$$

Thus scalar and relative-weight errors are separately identifiable at a frozen history with known propensities. They are not generally identifiable from terminal accuracy alone.

If updates are exact gradients and the positive ratios ϖ_i are constant over time, define $\widetilde{f} = \sum_i \widetilde{w}_i f_i$. Then $\mathbb{E}[\hat{g}_0 \mid \mathcal{H}] = c\nabla\widetilde{f}$. The corresponding expected direction is the gradient of a reweighted objective, scaled by c. This statement does not by itself prove convergence of a stochastic process to its minimizer; it identifies the mean field. The two-client example in the main text has $\nabla\widetilde{f} = \theta + 2/3$. Multiplication by any positive scalar preserves its zero at $-2/3$.

For the count estimator, $n_i^t/t \to p_i$ almost surely by the strong law. Continuity of clipping implies $q_i^t \to \max\{p_i, \underline{q}\}$. The limiting zero-forecast mean therefore uses ratios $\varpi_i \to \min\{1, p_i/\underline{q}\}$. When the floor is strictly below every p_i, these ratios tend to one. When some $p_i < \underline{q}$, the limiting change in relative weights generally persists.

A common propensity also simplifies the residual-normalized estimator. When $q_i = q$ on reporting clients and $F_i = 0$, its common $1/q$ cancels between numerator and denominator, leaving the participating mean. This identity holds for a nonempty realized reporting set; it does not make that random ratio unbiased for $\sum_i w_i d_i$.

A.3 AGGREGATION-AWARE DAMPING

Proof of Corollary 1. At a frozen history and with $F_i = \alpha h_i$, the conditional bias is $b_0 - \alpha v$. The conditional mean squared error against the conditional full-participation mean is bias squared plus trace variance:

$$\mathcal{R}(\alpha) = \|b_0 - \alpha v\|^2 + \sum_i \nu_i \|d_i - \alpha h_i\|^2 + \sum_i w_i^2 \frac{p_i}{q_i^2} \sigma_i^2.$$

The local-noise term is constant in α. Expanding the other terms gives $\mathcal{R}(\alpha) = C - 2\alpha A + \alpha^2 D$, where

$$A = \langle b_0, v \rangle + \sum_i \nu_i \langle d_i, h_i \rangle, \qquad D = \|v\|^2 + \sum_i \nu_i \|h_i\|^2.$$

For $D > 0$ and $\alpha > 0$, $\mathcal{R}(\alpha) < \mathcal{R}(0)$ is equivalent to $D\alpha < 2A$. It requires $A > 0$, and at $\alpha = 1$ it is equivalent to $2A > D$. When $D > 0$, the unconstrained minimizer is A/D; projecting it onto $[0, 1]$ minimizes the convex quadratic on that interval. If $D = 0$, then $v = 0$ and every term $\nu_i \|h_i\|^2 = 0$, implying $A = 0$ and a constant risk. At $q_i = p_i$, $u_i = 0$ and $\nu_i = w_i^2(1 - p_i)/p_i$, giving the stated known-propensity reduction. □

This risk uses the conditional population mean as its target, not a realized noisy full-participation update. The distinction matters when local-update noise is nonzero. Appendix E evaluates its minimizer directly on the archived Gram matrix. The accompanying deterministic test separately checks the quadratic identity and the improvement criterion by enumerating all reporting masks on small synthetic vectors. That test does not execute the original aggregation classes, so it does not close the cross-implementation issue of Appendix F.1.

A.4 FULL CONVERGENCE STATEMENT

The following result bounds the average squared gradient norm in terms of forecast residuals and propensity error. It formalizes how these quantities enter the convergence analysis. The local-drift argument is of the type used in federated optimization analyses such as Karimireddy et al. (2020).

Assume every f_i is L-smooth, $f \geq f_\star > -\infty$, and

$$\sum_i w_i \|\nabla f_i(\theta) - \nabla f(\theta)\|^2 \leq \delta^2 \quad \text{for every } \theta.$$

Each client starts at θ^{t-1} and performs exactly K steps

$$\theta_i^{t,k+1} = \theta_i^{t,k} - \eta_c \big(\nabla f_i(\theta_i^{t,k}) + \xi_i^{t,k} \big).$$

Conditioned on the history before its local step, $\xi_i^{t,k}$ has zero mean and second moment at most σ^2. Local-noise families are conditionally independent across clients. Participation is as in Section 2, forecasts and propensities are pre-round measurable, and all relevant second moments are finite. Let $\chi \geq 1$ satisfy $(p_i/q_i^t)^2 \leq \chi$ almost surely.

Theorem 1 (A forecast-error-dependent bound)**.** *Let $\widetilde{\eta} = \eta_s \eta_c K$, and use V^t and B^t from Eq.* (9)*. If $\widetilde{\eta} L \leq 1/16$, $\chi \widetilde{\eta} L \leq 1/2$, and $\eta_c K L \leq 1/24$, then*

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\theta^{t-1})\|^2 \leq \frac{8(f(\theta^0) - f_\star)}{\widetilde{\eta} T} + \frac{8 \chi \widetilde{\eta} L \sigma^2}{K} \sum_i w_i^2 \\ + 70 \eta_c^2 (K-1) L^2 \sigma^2 + 420 \eta_c^2 K (K-1) L^2 \delta^2 \\ + \frac{4 \chi \widetilde{\eta} L}{T} \sum_{t=1}^{T} V^t + \frac{40}{(\eta_c K)^2 T} \sum_{t=1}^{T} B^t. \end{aligned} \tag{12}$$

The expectation defining B^t must remain joint: q_i^t and the forecast residual both depend on the past. Replacing it by a product of marginal expectations is not justified.

A.5 LOCAL-UPDATE BOUNDS

Fix a round, write $a = \nabla f(\theta^{t-1})$, and use $\mathbb{E}_{t-1}$ for conditional expectation. Define

$$\mathcal{D} = \sum_i w_i \sum_{k=0}^{K-1} \mathbb{E}_{t-1}\|\theta_i^{t,k} - \theta^{t-1}\|^2, \qquad \bar{d} = \sum_i w_i \mathbb{E}_{t-1}\Delta_i^t.$$

Lemma 1 (Mean and noise of the local update)**.** *Under the local-SGD assumptions,*

$$\bar{d} = \eta_c K a + e, \qquad \|e\|^2 \le \eta_c^2 K L^2 \mathcal{D}, \tag{13}$$

$$\sum_i w_i^2 \sigma_{i,t}^2 \le 2\eta_c^2 K \sigma^2 \sum_i w_i^2 + 2\eta_c^2 K L^2 \mathcal{D}. \tag{14}$$

Proof. The mean-zero noise condition gives $\bar{d} = \eta_c \sum_{i,k} w_i \mathbb{E}_{t-1} \nabla f_i(\theta_i^{t,k})$. Subtract $\eta_c K a$, use Jensen's inequality across clients and conditional expectation, Cauchy–Schwarz across steps, and smoothness to obtain Eq. (13).

For Eq. (14), write $\Delta_i = D_i + Z_i$ where $D_i = \eta_c \sum_k \nabla f_i(\theta_i^{t,k})$ and $Z_i = \eta_c \sum_k \xi_i^{t,k}$. Martingale differences imply $\mathbb{E}_{t-1}\|Z_i\|^2 \le \eta_c^2 K \sigma^2$. Since the conditional mean minimizes conditional squared distance,

$$\mathbb{E}_{t-1}\|D_i - \mathbb{E}_{t-1} D_i\|^2 \le \eta_c^2 K L^2 \sum_k \mathbb{E}_{t-1}\|\theta_i^{t,k} - \theta^{t-1}\|^2.$$

Apply $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$, multiply by w_i^2, sum, and use $w_i^2 \le w_i$ in the drift term. □

Lemma 2 (Local drift)**.** *If $\eta_c K L \le 1/4$, then*

$$\mathcal{D} \le 7K(K-1)\eta_c^2\sigma^2 + 42K^2(K-1)\eta_c^2(\delta^2 + \|a\|^2). \tag{15}$$

Proof. For one client, put $u_k = \mathbb{E}_{t-1}\|\theta_i^{t,k} - \theta^{t-1}\|^2$ and $\zeta_i^2 = \|\nabla f_i(\theta^{t-1}) - a\|^2$. The case $K = 1$ is immediate because $\mathcal{D} = 0$. For general K, applying Young's inequality with parameter $1/(2K-1)$ and splitting the local gradient into its drift, heterogeneity, and global-gradient components gives

$$u_{k+1} \le \left(1 + \frac{1}{2K-1} + 6K\eta_c^2 L^2\right) u_k + B_i, \quad B_i = \eta_c^2\sigma^2 + 6K\eta_c^2(\zeta_i^2 + \|a\|^2).$$

Here $u_0 = 0$. The step-size condition implies $6K\eta_c^2 L^2 \le 1/(2K-1)$. Unrolling the recursion yields

$$u_k \le B_i \frac{2K-1}{2}\left(e^{2K/(2K-1)} - 1\right) \le 7KB_i.$$

Sum over $k = 1, \dots, K-1$, weight by w_i, and use $\sum_i w_i \zeta_i^2 \le \delta^2$. □

A.6 DESCENT AND TELESCOPING

Proof of Theorem 1. Let $b = \mathbb{E}_{t-1}\hat{g} - \bar{d}$ and

$$\mathcal{V} = \sum_i w_i^2 \frac{1-p_i}{p_i} \mathbb{E}_{t-1}\|\Delta_i^t - F_i^t\|^2.$$

Smoothness of f gives

$$\mathbb{E}_{t-1} f(\theta^t) \le f(\theta^{t-1}) - \eta_s\langle a, \bar{d} + b\rangle + \frac{L\eta_s^2}{2}\left(\|\bar{d} + b\|^2 + \operatorname{tr}\operatorname{Var}_{t-1}(\hat{g})\right).$$

Using Eq. (13),

$$-\eta_s\langle a, \bar{d}\rangle \le -\frac{\widetilde{\eta}}{2}\|a\|^2 + \frac{\eta_s\eta_c L^2}{2}\mathcal{D}, \quad -\eta_s\langle a, b\rangle \le \frac{\widetilde{\eta}}{16}\|a\|^2 + \frac{4\eta_s^2}{\widetilde{\eta}}\|b\|^2.$$

Also $\|\bar{d}+b\|^2 \le 4(\eta_c K)^2\|a\|^2 + 4\eta_c^2 K L^2 \mathcal{D} + 2\|b\|^2$. Equation (10) bounds the variance by $\chi(\mathcal{V} + \sum_i w_i^2 \sigma_{i,t}^2)$; apply Eq. (14) to its second term. Under the two conditions on $\widetilde{\eta}$, collecting coefficients gives

$$\mathbb{E}_{t-1} f(\theta^t) \le f(\theta^{t-1}) - \frac{5\widetilde{\eta}}{16}\|a\|^2 + \frac{5\eta_s\eta_c L^2}{4}\mathcal{D} + \frac{5\eta_s^2}{\widetilde{\eta}}\|b\|^2 + \frac{\chi L \eta_s^2}{2}\mathcal{V} + \chi L \eta_s^2 \eta_c^2 K \sigma^2 \sum_i w_i^2.$$

Substitute Eq. (15). Its gradient component is at most $\frac{105}{2}\widetilde{\eta}(\eta_c K L)^2\|a\|^2$. Since $5/16 - 105/1152 \ge 1/8$, the remaining descent coefficient is at least $\widetilde{\eta}/8$. The other drift coefficients are $35\widetilde{\eta}\eta_c^2(K-1)L^2\sigma^2/4$ and $105\widetilde{\eta}\eta_c^2 K(K-1)L^2\delta^2/2$.

By conditional Jensen's inequality,

$$\mathbb{E}\|b\|^2 \le \sum_i w_i \mathbb{E}\left[\left(\frac{p_i}{q_i^t} - 1\right)^2 \|\Delta_i^t - F_i^t\|^2\right] = B^t.$$

Moreover, $\mathbb{E}\mathcal{V} = (\eta_c K)^2 V^t$. Take total expectations, sum over rounds, telescope with $f \ge f_\star$, and multiply by $8/(\widetilde{\eta}T)$. The resulting coefficients are exactly those in Eq. (12). □

With known propensities, $B^t = 0$ and $\chi = 1$. If η_s is fixed, η_c is proportional to $T^{-1/2}$, and $T^{-1}\sum_t V^t$ remains bounded, the right side is $O(T^{-1/2})$. The theorem does not establish that residual condition for the particular forecasting algorithm. It also does not model variable local step counts caused by fixed local epochs, server weight decay, time-correlated availability, or conditional dependence between reporting and current updates.

B Forecasting rules and implementation details

B.1 Fixed and learned forecasts

The fixed configurations use $F_i = 0$, $F_i = h_i$, $F_i = 0.5h_i$, or the mean of the stored updates for client i. A missing fixed forecast is zero. For learned forecasting let $x_i = (h_i, m_i, \mu)$, treating missing basis vectors as zero. On receipt of a new update, the pre-round basis and that update define moment increments

$$G_i \leftarrow \psi G_i + (1-\psi)(\langle x_{i,a}, x_{i,b}\rangle)_{a,b=1}^3, \qquad v_i \leftarrow \psi v_i + (1-\psi)(\langle x_{i,a}, \Delta_i\rangle)_{a=1}^3.$$

Pooled moments receive the same observation contributions. The implementation uses $\psi = 0.9$. The hierarchical ridge solves are

$$u_{\text{pool}} = (G_{\text{pool}} + \lambda_{\text{pool}} I)^{-1}(v_{\text{pool}} + \lambda_{\text{pool}} u_0), \quad u_i = (G_i + \lambda_i I)^{-1}(v_i + \lambda_i u_{\text{pool}}),$$

where $u_0 = (1,0,0)$, $\lambda_i = 0.1\,\mathrm{tr}(G_i)/3$, and the pooled ridge coefficient is defined analogously. A level with no usable moments or a failed solve returns its prior. Coefficients are clamped componentwise to $[-1.5, 2]$.

The forecast is $F_i = (u_i)_1 h_i + (u_i)_2 m_i + (u_i)_3 \mu$, followed by

$$F_i \leftarrow F_i \min\{1, 3\bar{\nu}/\|F_i\|\},$$

with a zero-norm forecast left at zero. Here $\bar{\nu}$ is an EMA of the median norm of reported updates. The population vector μ is an EMA of the sample-weighted participating mean. These EMAs must be evaluated from previous rounds when constructing the current forecast. A client without a stored update receives the clipped μ under learned forecasting, with unit gain, not its learned c_i. Before any report exists, it receives zero. This differs from the cold-start rule of the fixed forecasts and confounds an interpretation specific to gain learning.

B.2 Required order of operations

For Eq. (1), the logical round order is as follows. First, snapshot the pre-round server state and compute F_i^t and q_i^t. Second, collect the reporting clients' updates and form the estimator using that snapshot. Third, apply the server step. Only then ingest reports into the counts, update histories, regression moments, and running population statistics. Local-update generation and report collection may be implemented before forecast evaluation provided the forecast state has not changed. The decisive condition is no use of the current report to construct its own baseline.

For an empty reporting set, Eq. (1) gives $\hat{g}_F = \sum_i w_i F_i$. With zero forecasting, both tested production paths return without an aggregate and skip the optimizer step. In particular, the imputation path does not apply a zero-gradient decay step in this case. With a nonzero forecast sum, the imputation path can apply a step while centralized FedAU skips. A nonempty reporting set whose updates cancel is a separate case: applying its zero aggregate with weight decay is not equivalent to skipping. No completed reported training run contains an empty reporting set.

B.3 Normalization and optional variants

The residual-normalized estimator is

$$\hat{g}_H = \sum_i w_i F_i + \frac{\sum_{i\in S}(w_i/q_i)(\Delta_i - F_i)}{\sum_{i\in S} w_i/q_i}.$$

For an empty S, its specified value is the forecast sum. The auxiliary mixed-mean configuration in the experimental record is

$$\hat{g}_\beta = (1-\beta)\frac{\sum_{i\in S} w_i\Delta_i}{\sum_{i\in S} w_i} + \beta\frac{\sum_{i\in S} w_i\Delta_i + \sum_{i\in A\setminus S} w_i F_i}{\sum_{i\in S} w_i + \sum_{i\in A\setminus S} w_i}, \qquad \beta = 0.5,$$

where A is the set of clients holding a forecast. It skips empty reporting rounds. This is retained as a secondary estimator check, not as an additional principal competitor.

The trust variant replaces F_i by $\kappa_i F_i$, with

$$\kappa_i = \mathrm{clip}_{[0,1]}\left(1 - \frac{E_i^{\mathrm{res}}}{E_i^{\mathrm{upd}}}\right) 0.9^{\tau_i - 1}.$$

The two E_i terms are historical EMAs of $\|\Delta_i - F_i\|^2$ and $\|\Delta_i\|^2$; τ_i is absence length. Before residual observations, the stated rule retains only the absence discount. In the available source snapshot, missing residual statistics or a nonpositive update-energy estimate return the absence discount alone; otherwise the clipped ratio above is used.

C Reported training configuration

CIFAR uses ResNet-18 with two-group GroupNorm, a 3×3 stem, and no initial max-pooling. The CIFAR-10 model has 11,173,962 parameters; the CIFAR-100 output head differs. FEMNIST uses a two-convolution CNN and a hidden fully connected layer with 62 outputs, totaling 2,278,334 parameters. Shakespeare uses embeddings of dimension 8 for 100 printable characters, two 256-unit LSTM layers, length-80 sequences, and a next-character output at every position, totaling 825,220 parameters.

The CIFAR partition assigns class samples to 200 clients using a Dirichlet draw with concentration 0.3 and then holds out 30 clients. Original training and test membership determines the within-client training and test portions. FEMNIST uses a seeded 80/20 split within writers; Shakespeare uses the first 80% of each role's text for training. The data generators use seed 12345. The training run seed controls initialization, data order, participation, and the evaluated client subset where applicable. Local optimizers are SGD without momentum. The experimental record reports weight decay in both client and server updates.

Weights w_i are proportional to local training sample counts. The main endpoint is the average accuracy across the final five scheduled evaluations. These evaluations cover a late-training window

Table 6: Local training and server configuration. A shared setting is not evidence that it is optimal for every method.

	CIFAR-10	CIFAR-100	FEMNIST	Shakespeare
Training clients	170	170	3597	858
Held-out clients	30	30	0	0
Rounds	200	200	300	300
Evaluation interval	10	10	20	20
Local epochs	5	5	5	2
Batch size	128	128	128	64
Client learning rate	0.03	0.03	0.005	0.5
Server learning rate	1	1	1	1
Weight decay	0.0005	0.0005	0.0005	0.0005
Evaluated clients	All held-out	All held-out	600	300
History length	3	3	2	2
History storage	fp16	fp16	int8	int8
Forecast norm multiplier	3	3	3	3
Propensity floor	0.05	0.05	0.05	0.05

rather than independent trials. CIFAR held-out-client evaluation is distinct from local test evaluation on FEMNIST and Shakespeare; their absolute accuracies should not be pooled into one metric. The underlying experimental record also contains unweighted per-client tercile summaries; these are distinct from the main endpoint.

Reference-code U-FedVARP and FedStale store one fp32 vector per CIFAR client and one fp16 vector per FEMNIST client. The shared forecasting code stores multiple quantized vectors. Forecast clipping is applied in the shared code but not to baseline histories. All configurations share one client/server learning-rate pair per task, with no method-specific tuning. These implementation differences remain part of the evaluated configurations.

FedAU's unbounded interval implementation uses $\omega_i = s_i/n_i$ after at least one prior report and 1 otherwise, where s_i is the latest prior report round. A finite-cap version limits individual intervals to 50 rounds; it should be reproduced from the reference interval algorithm, not replaced by an assumed clipping of the final average. In the experimental record, the finite cap changes no applied weights for some moderate-participation seeds. The available finite-cap table includes only the seeds actually re-run.

D Supplementary numerical evidence

Training tables retain the experimental record's display precision. The exported unrounded endpoints reproduce the paired training comparisons, including all 266 printed per-seed contrasts, and their primary endpoint summaries. Recomputing exported arithmetic is not independent training replication. Quantities absent from the endpoint export, including CIFAR-100 seen-client summaries and memory measurements, retain their technical-report provenance. A missing entry is not a failed run or a zero improvement; standard deviations of differences do not determine standard deviations of absolute accuracies without paired covariance.

D.1 Absolute endpoints and simple-forecast contrasts

A full-participation CIFAR-10 FedAU run is reported at 52.11% for seed 12 only. It is a single-run reference, not a multi-seed upper bound.

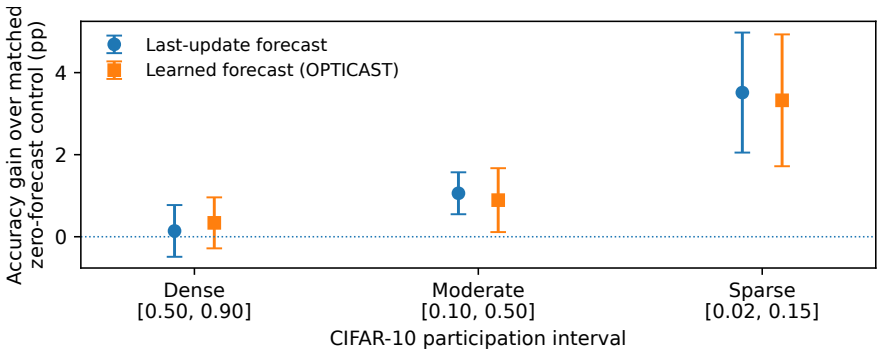


Figure 3: A forecast helps increasingly as participation falls within CIFAR-10. Points are mean differences from the matched zero-forecast count-based HT control; bars are two-sided 95% Student-t intervals recomputed from unrounded paired endpoints (3, 5, and 5 seeds). These are not intervals for the difference between learned and last-update forecasting. FEMNIST is not plotted on this within-task axis.

Table 7: Reported absolute endpoint accuracy (%). Entries are mean ± standard deviation across seeds. Dense has three seeds, moderate and sparse five, and FEMNIST three. CIFAR-100 endpoints are in Appendix D.4.

	Dense	Moderate	Sparse	FEMNIST
FedAU	50.73 ± 0.75	49.23 ± 1.10	42.16 ± 1.80	46.76 ± 2.04
Zero forecast	50.97 ± 1.20	49.80 ± 1.29	45.39 ± 1.92	40.10 ± 2.10
Last update	51.11 ± 1.08	50.86 ± 1.04	48.91 ± 1.07	50.82 ± 2.31
OPTICAST	51.30 ± 1.06	50.69 ± 1.10	48.72 ± 0.75	52.57 ± 2.55
Participating mean	50.65 ± 0.89	49.21 ± 1.20	46.25 ± 1.85	55.68 ± 2.05

Table 8: Forecast minus matched zero-forecast control under count-based HT (percentage points). Entries are paired mean ± standard deviation; seed counts match the preceding table.

Forecast	Dense	Moderate	Sparse	FEMNIST
Last update	0.14 ± 0.25	1.06 ± 0.41	3.51 ± 1.18	10.72 ± 0.70
Moving average	—	1.21 ± 0.44	—	9.74 ± 0.64
Damped last update	—	0.82 ± 0.27	—	6.29 ± 0.46
OPTICAST	0.34 ± 0.25	0.89 ± 0.63	3.33 ± 1.29	12.47 ± 1.33

D.2 PROSPECTIVELY SPECIFIED COMPARISONS AND THE FACTORIAL

Table 9: Per-seed differences for the two separate CIFAR-10 evaluation groups. Their reported summary tests are in Table 5.

Seed	OPTICAST − FedAU	Seed	OPTICAST − last	OPTICAST − average
17	3.54	22	−0.80	−0.96
18	2.35	23	−0.34	−0.29
19	1.27	24	0.14	0.24
20	0.26	25	0.25	0.37
21	1.36	26	0.13	0.13

The first comparison was specified as a one-sided paired t-test. The registration records 5 September 2026 as its commitment date, before any of its seeds 17–21 began. The second registration was committed on 11 September before any of its seeds 22–26 began, and specified two two-sided paired t-tests. Its decision rule reserved a gain claim for a significant positive contrast against last-update

reuse and required a nonsignificant outcome to be called not distinguishable, not equivalent. The first registration followed exploratory results: it does not make those earlier results confirmatory. The registration documents record this chronology. The original repository history is not part of the numerical supplement; the stated dates are documented provenance rather than independently authenticated timestamps.

Table 10: Four-cell moderate-participation factorial, expressed as differences from the same reference FedAU run. Both propensity branches in the shared code are clipped to $[0.05, 1]$; the interval cells are not reference FedAU. The interaction concerns these configured branches and remains provisional pending verification of the historical real-data trajectories.

Configuration	12	13	14	15	16
Zero, count	1.50	−0.11	−0.07	1.07	0.46
Last, count	2.27	0.97	0.65	2.82	1.45
Zero, clipped interval	0.61	−0.47	−1.24	−0.09	0.08
Last, clipped interval	1.16	0.34	0.06	2.10	0.86
Reported interaction	−0.22	−0.26	0.58	0.45	−0.21

The interaction is $(\text{last}_{\text{clipped interval}} - \text{zero}_{\text{clipped interval}}) - (\text{last}_{\text{count}} - \text{zero}_{\text{count}})$. Its reported mean is 0.07 points, standard deviation 0.41, two-sided 95% interval $[-0.44, 0.58]$, and $p = 0.74$. The interval is not contained in the prespecified ± 0.5-point margin, so the analysis establishes neither an interaction nor a practically negligible one. The exported endpoints reproduce the interaction as 0.0660759 points ($p = 0.737508$). Arithmetic on independently rounded cells can differ by 0.01 points from the directly computed contrast. These cells are retained as provisional empirical evidence; neither reference-FedAU equivalence nor a pure propensity-estimation attribution follows from them. Appendix F.1 separates the known clipping difference from the unexplained residual.

D.3 Main competitors: per-seed differences

Table 11: Moderate CIFAR-10 differences from FedAU. All rows use seeds 12–16.

Configuration	12	13	14	15	16
MIFA	0.66	−0.41	−0.92	0.92	−0.67
U-FedVARP	0.07	0.57	0.04	1.94	0.94
FedStale	0.78	0.84	0.60	1.51	0.57
OPTICAST	1.62	0.80	1.70	2.22	0.98
Participating mean	0.61	−0.74	−0.60	0.47	0.17

Table 12: Sparse CIFAR-10 differences from FedAU. The last three rows change only the count-estimator floor within each shared configuration.

Configuration	12	13	14	15	16
MIFA	2.70	1.65	4.79	3.75	0.06
U-FedVARP	3.87	2.49	5.68	3.98	−0.64
FedStale	2.58	2.86	5.13	3.37	1.75
Zero, count	6.19	0.65	2.67	4.94	1.72
Last, count	8.60	4.48	8.07	7.76	4.83
OPTICAST	7.73	4.33	7.65	8.71	4.39
Participating mean	7.22	0.65	4.83	5.51	2.25
Zero, floor 0.15	0.86	−3.72	0.34	−0.08	−3.67
Last, floor 0.15	6.42	2.99	7.34	6.06	3.14
OPTICAST, floor 0.15	3.54	1.72	4.63	2.90	0.28

On the true-but-clipped FEMNIST arm, $q_i = \underline{q}$ exactly. The unrounded endpoints give, at seed 12, a 13.7147-point gap to the participating mean, a 13.7522-point gain from scaling, and a remaining

Table 13: FEMNIST differences from FedAU. "Oracle-scaled" uses the simulator's true probabilities to set a fixed multiplier; the scaled observations are single-seed diagnostics. "True, clipped" is not unbiased oracle weighting.

Configuration	12	13	14
MIFA	2.30	1.52	2.89
U-FedVARP	3.39	2.20	3.63
FedStale	1.75	1.35	2.13
Zero, count	−7.02	−6.28	−6.71
Last, count	4.13	3.64	4.39
OPTICAST	6.38	4.67	6.36
Participating mean	9.52	8.31	8.94
Last, residual-normalized	9.17	8.09	8.73
Learned, residual-normalized	9.26	8.23	8.69
Zero, oracle-scaled	7.32	—	—
Zero, true, clipped	−4.20	−3.89	−3.47
Zero, true, clipped, oracle-scaled	9.56	—	—

gap of -0.0375 points. The recovery ratio is 100.2731%, not a clipped 100%. This near-match is an empirical outcome, not an identity between the estimators or proof of restored client weighting. On the estimated arm, the absolute endpoints are 45.2880% for FedAU, 54.8067% for the participating mean, 38.2706% for zero forecasting, and 52.6066% for its scaled version. Independently rounding endpoints and contrasts to four decimals can change the final digit of a subtraction; Table 14 reports contrasts calculated before rounding. The scale factor multiplies the aggregate only, leaving server weight decay and update order unchanged. Seeds 13 and 14 have not been run for this intervention.

Table 14: Oracle scale intervention on FEMNIST, seed 12. Gap is participating mean minus unscaled zero forecast; recovered is scaled minus unscaled; remaining is gap minus recovered. Values are calculated from unrounded run endpoints. Neither row partitions a forecasting method's gain.

Propensity rule	Gap (pp)	Recovered (pp)	Remaining (pp)	Recovery
Estimated count, clipped	16.5361	14.3360	2.2001	86.6951%
True p, clipped	13.7147	13.7522	−0.0375	100.2731%

D.4 CIFAR-100: ABSOLUTE AND PAIRED RESULTS

Tables 15–17 report CIFAR-100 endpoint levels and seed-wise comparisons. All use seeds 12–16, the moderate participation interval, and held-out-client accuracy; none is a new training experiment or a prospectively specified test on this dataset.

Table 15: Absolute CIFAR-100 endpoint accuracy (%). Mean and standard deviation over the same five seeds. Seen accuracy is measured on local test data of training clients; held-out accuracy is the main endpoint.

Configuration	Seen clients	Held-out clients
FedAU	21.77 ± 0.43	22.60 ± 0.54
Participating mean	22.20 ± 0.29	23.01 ± 0.66
MIFA	22.14 ± 0.23	22.93 ± 0.65
U-FedVARP	22.31 ± 0.18	23.05 ± 0.64
FedStale	22.15 ± 0.24	22.92 ± 0.58
Zero, count	22.52 ± 0.37	23.21 ± 0.51
Last, count	22.65 ± 0.40	23.32 ± 0.53
OPTICAST	22.74 ± 0.33	23.44 ± 0.68

Table 16: CIFAR-100 paired differences from FedAU (percentage points). The same five seeds underlie every row.

Configuration	12	13	14	15	16
Participating mean	0.03	0.70	0.48	0.66	0.17
MIFA	0.91	0.48	0.30	0.34	-0.36
U-FedVARP	1.18	0.86	0.35	-0.04	-0.09
FedStale	0.66	0.52	0.36	0.04	-0.01
Zero, count	0.68	0.66	0.59	0.48	0.64
Last, count	1.30	0.90	0.59	0.38	0.43
OPTICAST	1.28	1.10	1.01	0.24	0.55

Table 17: Direct CIFAR-100 forecast-minus-zero contrasts (percentage points). These are independently rounded direct contrasts, not subtractions of the rounded cells in Table 16. The learned configuration includes its different cold-start rule.

Forecast	12	13	14	15	16	Mean $\pm$ SD
Last update	0.62	0.24	0.00	-0.11	-0.21	0.11 ± 0.33
OPTICAST	0.59	0.44	0.42	-0.24	-0.09	0.22 ± 0.37

The unrounded endpoint comparisons give positive-seed counts of $2/5$ for last-update forecasting and $3/5$ for OPTICAST. A printed 0.00 is not used to infer its sign. The mean learned-minus-last contrast is about 0.1 points and is positive at three seeds. These observations limit the claim about learned complexity on this task; they do not establish equivalence or replicate a registered decision rule on a new population. Residual-normalized and finite-cap FedAU configurations were not evaluated on this block.

D.5 FINITE INTERVAL CAPS AND DAMPING SENSITIVITY

Table 18: Reported changes from FedAU's unbounded interval cap to cap 50 (percentage points). These are baseline sensitivity results, not selected best-of-two baselines.

Setting	Mean $\pm$ SD	Seeds	Per-seed differences
Moderate CIFAR-10	0.04 ± 0.53	12,13,15	$0.51, -0.54, 0.15$
Sparse CIFAR-10	1.66 ± 1.74	12–16	$3.37, -0.14, 0.79, 3.68, 0.61$
FEMNIST	-2.57 ± 0.19	12–14	$-2.74, -2.37, -2.62$
Shakespeare	0.90 ± 0.77	12–14	$1.38, 1.31, 0.01$

Against the finite-cap FedAU baseline, mean sparse CIFAR-10 differences are $+1.57$ for zero forecasting, $+5.09$ for last-update forecasting, and $+4.90$ for OPTICAST, over five seeds. The corresponding FEMNIST differences are -4.09, $+6.63$, and $+8.38$, over three seeds. The sign pattern is not an artifact of considering only the unbounded interval cap.

The available FedStale sweep selected $\alpha = 0.7$ on moderate CIFAR-10 and $\alpha = 0.9$ on FEMNIST using seed 12. Excluding that selection seed, gains over FedAU are 0.96 points on seeds 13–16 and 2.70 points on seeds 13–14, respectively. These are limited parameter checks, not complete tuning. On FEMNIST the selection-seed result improves up to the edge of the damping grid, and the undamped U-FedVARP endpoint is higher still. Quoting only the default damped baseline would conceal this sensitivity.

D.6 STABILITY AND COMPUTATIONAL COST

Collapse is defined as an endpoint below half the best observed accuracy. Its onset is the first logged round after which every evaluation remains below half the best value observed before that onset. FedStale collapses at rounds 140 and 200 on seeds 12 and 14; OPTICAST at rounds 200 and 80

Table 19: Shakespeare per-seed differences from FedAU (percentage points). Completed but collapsed runs are retained.

Configuration	12	13	14
MIFA	1.85	1.97	1.71
U-FedVARP	0.48	−0.06	0.53
FedStale	−31.71	1.42	−33.38
Zero, count	1.45	1.38	0.86
Last, count	2.22	2.25	−0.63
OPTICAST	−36.94	−42.61	1.74
OPTICAST, trust weight	1.97	1.72	0.55
Participating mean	1.39	1.37	−0.67

on seeds 12 and 13. Applying this onset rule to the newly exported complete accuracy histories reproduces all four recorded onset rounds; no new training is involved.

Possible explanations include pooled regression sensitivity, the cold-start population forecast, and large local LSTM updates. They are hypotheses, not established causes. Median trust weights in the latter half of training are about 0.67 on moderate CIFAR-10, 0.022 on FEMNIST, and 0.0014 on Shakespeare. Suppression sufficient to avoid a collapse can also remove the benefit of a forecast: the trust-weighted OPTICAST configuration is 4.90 points below FedAU on FEMNIST.

Table 20: Reported FEMNIST server cost on one hardware platform. Times are means ± standard deviations across three exclusive-GPU runs. Memory figures refer to the reported stored state, not a common precision budget.

Configuration	Aggregation state (GiB)	Logging state (GiB)	Wall time (min)
FedAU	Approximately 0	—	108.7 ± 1.0
U-FedVARP	15.3[a]	—	112.9 ± 1.1
FedStale	15.3[a]	—	112.6 ± 0.8
Last, count	15.1	1.9	162.4 ± 1.6
OPTICAST	15.1	1.9	176.1 ± 2.0

[a]The reference implementations do not report a separate aggregation/logging split; 15.3 GiB is their reported tracker state.

The shared implementation also retains update histories for the zero-forecast control. That is a property of the experimental code path, not an inherent memory requirement of inverse-propensity aggregation.

E FROZEN-CHECKPOINT MOMENT CHECK

The frozen-checkpoint experiment trains OPTICAST for 50 moderate-participation CIFAR-10 rounds, freezing the model and server state, and replaying each client's local training under three fixed batch orders. The mean replayed vector is then treated as a fixed update. For each propensity regime, 10,000 Bernoulli reporting masks are sampled. Fifty masks are also passed through the aggregation implementation to check the closed-form calculation. All local-update vectors remain fixed during mask resampling, so $\sigma_i^2 = 0$ in the tested moment identities.

The known-propensity regime uses probabilities above the floor and divides by those probabilities. The estimated regime reuses the same masks with historical count estimates. The synthetic below-floor regime changes the probabilities and masks, with about 4.9 expected reporting clients per draw. It is not a frozen FEMNIST checkpoint.

At known propensities, finite-sample measured bias norms are nonzero even when the exact bias is zero. Ratios of measured bias norm to the vector-mean error scale are consistent with this Monte Carlo noise. The reported predicted-to-measured variance ratios are 0.992–1.01; recomputing ratios from the rounded cells can change the last displayed digit.

Table 21: Reported Monte Carlo and analytic moments. Bias norms are divided by the full-update norm; variances by its square. MC SE is the archived root-mean-square Monte Carlo error scale for the vector mean, not an independently estimated scalar-bias confidence interval.

Propensities	Forecast	Bias	MC SE	Predicted bias	Variance	Predicted var.
Known	None	0.00831	0.0107	0	1.14	1.13
Known	Last	0.00479	0.00460	0	0.212	0.210
Known	Learned	0.00432	0.00431	0	0.186	0.184
Estimated	None	0.178	0.0120	0.177	1.45	1.44
Estimated	Last	0.0798	0.00540	0.0808	0.292	0.292
Estimated	Learned	0.0861	0.00511	0.0857	0.261	0.261
All below floor	None	0.428	0.0196	0.422	3.85	3.89
All below floor	Last	0.264	0.00796	0.261	0.634	0.639
All below floor	Learned	0.246	0.00729	0.243	0.531	0.538

E.1 Comparing estimator error on the same frozen state

Table 22: Normalized one-step root mean squared error on the archived CIFAR-10 checkpoint. Values are divided by the full-participation-update norm. The same fixed update vectors and corresponding masks are used across estimator choices within a column.

Estimator	Forecast	Known p	Estimated p	All p below floor
HT	None	1.07	1.22	2.01
HT	Last	0.46	0.55	0.84
HT	Learned	0.43	0.52	0.77
Hájek	None	1.05	1.13	3.51
Hájek	Last	0.45	0.50	1.41
Hájek	Learned	0.42	0.47	1.27
Mixed mean	None	0.91	0.91	3.51
Mixed mean	Last	0.51	0.51	1.78
Mixed mean	Learned	0.53	0.53	1.79

The mixed mean does not use propensities, so its known- and estimated-propensity columns coincide when masks coincide. Below the floor, all q_i share one value: without forecasts, both Hájek and the mixed mean reduce to the participating mean and have equal errors. These coincidences are algebraic checks. HT has the lowest one-step error below the floor for each fixed forecast in this table. The training comparisons on FEMNIST concern a different dataset and an evolving trajectory; they do not constitute a same-task experiment in which smaller one-step error causes worse learning.

E.2 Aggregation-aware damping check

The archived Gram matrix contains three blocks: current reference updates, stored last updates, and learned forecasts for 170 clients. The matrix is stored as float64; the analysis also accumulates its contractions in float64. Writing its blocks as G_{dd}, G_{dh}, and G_{hh}, and $z_i = w_i(p_i/q_i - 1)$, the coefficients are

$$A = z^\top G_{dh} z + \sum_i \nu_i (G_{dh})_{ii}, \qquad D = z^\top G_{hh} z + \sum_i \nu_i (G_{hh})_{ii}.$$

The analysis package includes these arrays and recomputes the analytic optimum and a grid search with spacing 10^{-6} on $[0, 1]$. This requires no new local training or checkpoint. It also recomputes the analytic moments and compares them with the archived Monte Carlo statistics; it does not resample the original masks.

The analytic and grid minima agree to numerical precision in each regime. The optimum is interior at this checkpoint, rather than at zero forecasting or full stale reuse. This is an oracle one-step statement conditional on the frozen history, not an evaluated adaptive damping method. No trained model in the main comparison used these retrospectively calculated gains. The checkpoint has no client without history, and the last-update norm clip does not bind, so this calculation checks neither cold-start

Table 23: Oracle scalar damping of the stored last-update direction, independently recomputed from the archived arrays. The analytic optimum lies in $[0, 1]$ in every regime. The last column is its absolute difference from the fine-grid minimizer, not an uncertainty estimate across checkpoints.

Propensity regime	$\alpha_\star = A/D$	Grid discrepancy
Known p	0.883909	3.0×10^{-7}
Estimated p	0.875657	1.3×10^{-7}
Synthetic below floor	0.899160	7.8×10^{-8}

behavior nor clip activation. The current-round reference updates are fixed during mask resampling; finite replay noise in those vectors is shared by the analytic and Monte Carlo calculations.

F Implementation verification and evidence scope

F.1 Controlled-input production checks

The supplied verification snapshot contains the production centralized and imputation aggregators. Four CPU programs were independently rerun: the aggregation fixtures, complete-state random-generator checks, participation-estimator checks, and synthetic multiround trace. Each completed successfully. The separate production-scale reduction test retains its archived result; its reported memory requirement exceeded the available verification memory. These checks exercise the verification snapshot, not each historical training revision. The accompanying analysis-only supplement retains the test descriptions and outcomes, but not upstream-derived production source files whose redistribution permissions remain unresolved.

The aggregation suite contains 100 matched-input cases: 76 cases compute aggregates within the declared tolerance and 24 cases skip the optimizer step on both paths. Two additional cases compare raw and clipped deployed weights with their respective FP64 oracles. Fixtures cover nonuniform weights, several tensor shapes, empty and nonempty reporting sets, cancellation, mixed magnitudes, client permutations, zero-weight entries, and two optimizer settings. The relative tolerance is 10^{-6} against the accumulation scale $\sum_{i \in S} |w_i/q_i| \|\Delta_i\|$, with absolute tolerance 10^{-9}. Prescribed-update client fixtures do not exercise the complete data-loading or historical training pipeline.

The random-generator test compares full Torch state tensors, complete legacy NumPy tuples, Python random states, and explicitly instantiated generator states. Reporter identities and step counts are checked separately; a guard rejects unequal byte arrays with equal sums. The tested paths agree under these checks. No CUDA state is initialized by this CPU test. An earlier byte-sum comparison was superseded and is not used as evidence of state identity.

Tests of the production participation estimator establish prefix invariance when future reports change, pre-round timing when the reporting sequence is truncated, and the effect of a finite cap within the interval algorithm. The finite-cap output differs from clipping the final interval average. These checks validate the tested provider and distinguish a finite interval cap from the imputer's propensity floor.

For zero forecasting and an empty reporting set, both paths skip the optimizer step. Nonzero forecasts can change empty-round behavior, but no empty round occurs in the reported training sequences. Archived reduction-route tests show small FP32 differences between matrix reduction and sequential accumulation. The archived synthetic 200-round trace records final relative parameter difference 2.164×10^{-7}; its CPU re-execution gives 2.052×10^{-7} in the verification environment. Neither trace nor a parameter-distance bound alone identifies the effect on historical classification accuracy.

F.2 Historical revisions and the scope of attribution

At moderate CIFAR-10 seed 13, reference FedAU has endpoint accuracy 48.8849556% and the shared zero-forecast clipped-interval configuration has 48.4129798%, a difference of -0.4719758 percentage points. No propensity estimate falls below the shared floor at that seed. The difference therefore cannot be assigned to that floor. Across seeds 12–16, floor activations are respectively 0.047%, 0, 0.018%, 0.012%, and 0.068% of all client-round estimates, not of reporting clients alone.

The configurations were executed at different source revisions. At seed 13 the recorded revisions are `rev04` for reference FedAU and `rev09` for the shared interval control. The historical source comparison reports changes in aggregation, imputation, training orchestration, and utilities. Other factorial arms also span revisions. Equal nominal hyperparameters and seeds do not establish identical implementations. The current controlled-input checks therefore validate their own source snapshot without retrospectively certifying every archived intervention.

The archived diagnostics report the same round-zero accuracy for several historical configurations. This equality is a scalar diagnostic, not proof of identical parameters, buffers, partitions, or evaluation state. It also reports differing accuracies by the first subsequent logged evaluation. A coupled original-task execution would be needed to locate the first upstream or numerical divergence; it has not been performed. We retain paired outcomes and the count-versus-clipped-interval factorial without assigning their complete difference to a single component.

F.3 Numerical records and integrity

The received anonymous archive contains 49 checksum-listed payload files and its checksum manifest. Its external checksum and all 49 payload checksums were verified. Its numerical-analysis command completed successfully from the extracted archive, reproducing its 107 archive-wide summary cells. Those summaries pool the recorded seeds within configurations and are not substituted for the manuscript's explicitly matched seed sets; the accompanying manuscript analysis regenerates the tables and tests with their stated seeds. The two experiment manifests contain 427 entries: 416 successful completions, including four unfavorable Shakespeare outcomes, and 11 technical failures. The compact endpoint export contains 390 unique configuration–seed rows. Its 26 omitted successful entries belong to the report-only London (12) and MNIST (14) blocks. Counts describe records, not independent replications of a single hypothesis.

Each exported endpoint equals the mean of its five listed scheduled evaluations. The 266 per-seed contrasts checked against the export agree at their reported precision. Recomputed prospective-test differences are 1.7557521 points ($p = 0.0170453$, one-sided), -0.1250738 points ($p = 0.5611445$), and -0.1038342 points ($p = 0.6876929$). The frozen-array moment and damping calculations require no additional client training. Integrity checks establish consistency of the experimental records, not independent authentication of the original event files.

The history export contains 30,668 observations across four logged metrics, including 7,667 primary-accuracy observations from 390 runs. The final five accuracy values of every history exactly match the previously exported tail, so no endpoint or seed was changed. The main test-accuracy figure uses common recorded evaluation rounds throughout training, with no smoothing, missing-value imputation, or horizon extension. One FEMNIST FedAU run has seven additional early evaluations between the common 20-round checkpoints; these remain in the export but are not used to change the seed count in the figure. All scheduled checkpoints are present. The four recorded Shakespeare collapse-onset labels are independently confirmed under the stated onset rule. A finite-cap inventory-label defect was corrected without changing endpoints. A historical duplicate-writer incident is documented separately: four tainted manifest entries were removed and archived before the affected configurations were rerun. The manifests are not literally unmodified.

F.4 Evidence selection and remaining limitations

The four-task evaluation retains the completed unfavorable results, the stronger FEMNIST participating mean, and the prospective forecast-only null. London was excluded from the experimental programme and has an incomplete grid. The legacy single-seed MNIST blocks lack the controls required here; their alternative dropout process also forbids consecutive absences. They are not used as additional evidence for the present conclusions.

The FEMNIST scale intervention remains single-seed. The source-revision comparison remains unresolved, and no new full training run was performed. Any demonstrated implementation correction requires identifying affected results and preserving the original record. Numerical reconstruction and controlled-input verification do not by themselves resolve historical training equivalence.