BEYOND STALENESS DECAY: HISTORY-ENHANCED RESIDUAL ACCELERATION IN ASYNCHRONOUS FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Asynchronous federated learning produces a stream of client updates computed from stale model snapshots. We introduce History-Enhanced Residual Acceleration (HERA), a server-side method that reuses this stream as a short residual history for Anderson mixing. Lagged RMS stabilization balances the residual geometry, while a base-anchored update controls the full displacement of each history proposal. Our analysis separates base-update drift, delay, history amplitude, and client-selection bias. With bounded delay, uniform sampling, and a stepsize-scaled anchor, local-SGD residuals attain the $O(K^{-1/2})$ nonconvex stationarity rate; nonuniform arrivals and fixed anchors contribute explicit residual terms. A local error bound relates proposal quality to residual fitting, map conditioning, and evaluation mismatch. A four-way MNIST ablation examines the interaction between RMS stabilization and history correction, with their combination improving mean final accuracy over RMS alone in all five settings. Across ten CIFAR scenarios, HERA attains the highest mean final validation accuracy in eight, including all six non-IID settings.

1 INTRODUCTION

Federated learning enables multiple clients to collaboratively train a shared model while keeping their raw data local (McMahan et al., 2017). In standard synchronous federated learning, however, the server must wait for the selected clients to finish local computation before aggregating their updates. This synchronization barrier makes training vulnerable to stragglers caused by heterogeneous computation, communication, and availability. Asynchronous federated learning removes this barrier by allowing the server to update whenever a client response arrives (Xie et al., 2019). The resulting event-driven protocol improves resource utilization, but it introduces *staleness*: an arriving client update is typically computed from an outdated server model. Handling staleness is therefore a central challenge in asynchronous federated learning.

Existing methods handle staleness through complementary mechanisms: Taylor-based gradient correction (Zheng et al., 2017), delay-dependent mixing in FEDASYNC (Xie et al., 2019), buffered aggregation in FEDBUFF (Nguyen et al., 2022), heterogeneous-update calibration in CA2FL (Wang et al., 2024a), direction-aware weighting and gradient correction in FEDAC (Zang et al., 2024), and adaptive server updates in FADAS (Wang et al., 2024b). The same arrivals also encode optimization directions from earlier points on the server trajectory. As they accumulate, they form a short history that can support acceleration when its geometry remains informative at the current model. Reusing that history requires accounting jointly for delay, client heterogeneity, and stochastic local training. This motivates the question:

How can the history of stale arrivals be reused reliably to accelerate asynchronous federated learning?

We introduce History-Enhanced Residual Acceleration (HERA), which reuses stale client updates as a short residual history. Anderson mixing extracts affine coefficients from this history (Anderson, 1965; Walker & Ni, 2011; Toth & Kelley, 2015). Because asynchronous arrivals differ in their starting

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.

snapshots, client objectives, and stochastic local trajectories, HERA first stabilizes their residual geometry and then anchors the resulting history proposal to the current base update.

Related work. Asynchronous optimization has long been studied as a way to reduce synchronization overhead and improve throughput in large-scale learning. Early work established delay-tolerant and lock-free update rules in shared-memory or parameter-server settings, including HOGWILD!, asynchronous parallel SGD for nonconvex optimization, and the AROCK fixed-point framework (Niu et al., 2011; Lian et al., 2015; Peng et al., 2016). Federated methods extend these ideas through buffering, scheduling, adaptive updates, and server-side correction (Nguyen et al., 2022; Li et al., 2024; Wang et al., 2024b;a). Most of these methods treat delay primarily as a perturbation to be compensated, buffered, reweighted, scheduled, or corrected. HERA instead asks whether stale arrivals can also provide a useful short-window history after stabilization and safeguards.

Anderson acceleration is a classical residual-based acceleration mechanism. Multisecant formulations and regularized nonlinear acceleration relate Anderson-type extrapolation to approximate secant and quasi-Newton ideas (Fang & Saad, 2009; Scieur et al., 2016). The approach has been extended to proximal-gradient and nonsmooth settings, with globalized type-I schemes and analyses for nonsmooth fixed-point problems (Mai & Johansson, 2020; Zhang et al., 2020; Bian et al., 2021). Recent work studies stochastic Anderson mixing, filtering, restarting, explicit global convergence rates, and numerically stable orthogonalization procedures (Wei et al., 2021; Pollock & Rebholz, 2023; Ouyang et al., 2024; Scieur, 2024; Tang et al., 2024). Anderson-type ideas have also appeared in federated optimization (Feng et al., 2025).

Most of this acceleration literature assumes that the history is generated by one coherent underlying fixed-point or optimization map. In asynchronous federated learning, the server history instead consists of stale, client-dependent, and stochastic local updates, so its residual geometry is less consistent. HERA connects these directions by stabilizing asynchronous residuals and anchoring their history correction to the current base update. Appendix D gives the extended comparison.

Contributions. We make three contributions.

- We develop HERA, a server-side residual-history method for asynchronous federated learning that combines lagged RMS stabilization, regularized Anderson fitting, and base-anchored acceptance using updates already available at the server.
- We derive a residual-level bound separating base-update drift, delay, history amplitude, and client-selection bias. Uniform local SGD with stepsize-scaled anchoring attains $O(K^{-1/2})$ stationarity; nonuniform arrivals and fixed anchors retain explicit residual terms. Local bounds connect residual fitting to proposal quality.
- We provide complementary mechanism and performance evaluations. A four-way MNIST ablation examines RMS stabilization and history correction across five asynchronous settings. A ten-scenario CIFAR comparison evaluates the complete HERA configurations against six asynchronous baselines.

Notation. Unless otherwise stated, vectors lie in $\mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. We use $\| \cdot \|_2$ for the Euclidean vector norm and the induced matrix spectral norm, writing also $\| \cdot \|$ for the former; for a positive semidefinite matrix M, $\|u\|_M = (u^\top M u)^{1/2}$. Let $\mathbf{1}$ be the all-ones vector, $[N] = \{1, \dots, N\}$, and $\mathrm{Diag}(q)$ the diagonal matrix with diagonal q. The symbols $\odot, \oslash$ denote coordinatewise multiplication and division, and $\mathrm{clip}(z, q_{\min}, q_{\max})$ coordinatewise clipping to $[q_{\min}, q_{\max}]$. For symmetric matrices, $\preceq$ is the Löwner order and $\mathrm{spec}(M)$ the spectrum. Finally, $[r_j]_{j \in \mathcal{H}_k}$ has columns r_j in retained-window order.

2 FROM ASYNCHRONOUS LOCAL UPDATES TO RESIDUAL ANDERSON MIXING

We formulate asynchronous federated learning through per-arrival server updates, reinterpret returned client displacements as residual surrogates, and identify the geometric mismatch that arises when Anderson mixing is applied to stale, client-dependent updates.

2.1 ASYNCHRONOUS FEDERATED LEARNING

We consider the standard federated objective

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x),$$

where f_i is the local objective of client i. We index committed server updates by k, writing $x_{k+1} = x_k + u_k$. In single-arrival execution, update k processes a client displacement Δ_k computed from an earlier snapshot x_{s_k}, where $s_k \leq k$. Its staleness, $\tau_k = k - s_k$, counts the intervening server updates. A returned displacement may combine multiple local training steps. The symbols Δ_k and u_k distinguish the local change from the committed server update. A standard asynchronous baseline commits a delay- or reliability-weighted version of the returned update,

$$u_k = r_k^{\text{base}} = w(\tau_k)\Delta_k,$$

so that $x_{k+1} = x_k + r_k^{\text{base}}$. Here $w(\tau_k)$ is a prescribed staleness weight, possibly together with other scalar reliability factors. HERA uses the base displacement as the reference for its server-side history correction.

2.2 RESIDUAL VIEW AND ANDERSON MIXING

For a deterministic synchronous base rule, the fixed-point form is $x_{k+1} = G(x_k)$, with residual $r(x) = G(x) - x$. An optimization step can therefore be viewed as a residual-like displacement. In the asynchronous setting, r_k^{base} serves as a *residual surrogate*: a displacement computed from an earlier snapshot and applied at the current server model.

For example, $f(x) = x^2/2$ with stepsize 0.1 gives $G(x) = 0.9x$. A client starting at $x_{s_k} = 2$ returns -0.2; applying it at $x_k = 1$ gives 0.8, rather than $G(x_k) = 0.9$. History reuse must account for both the stale displacement and the intervening server movement.

Traditional delay weighting controls how strongly a stale arrival influences the server. Anderson mixing instead fits *affine coefficients*—weights that sum to one—from a short residual history. Its effectiveness therefore depends on the comparability of the candidate–residual pairs, which asynchronous delay and client-specific local training disrupt.

A direct extension is to maintain a short history of recent residual surrogates r_j^{base} and candidates $x_j + r_j^{\text{base}}$, and to compute affine coefficients by a regularized residual fit,

$$\alpha_k = \arg\min_{\mathbf{1}^\top \alpha = 1} \|R_k \alpha\|^2 + \lambda \|\alpha\|^2, \qquad x_{k+1} = \sum_{j \in \mathcal{H}_k} \alpha_{k,j} (x_j + r_j^{\text{base}}), \tag{1}$$

where the columns of $R_k = [\, r_j^{\text{base}} \,]_{j \in \mathcal{H}_k}$ collect recent residuals and $\mathcal{H}_k$ denotes the admissible window. The coefficients adapt to the geometry of the recent residuals, complementing the age information used by delay weighting.

2.3 GEOMETRIC MISMATCH IN ASYNCHRONOUS ANDERSON MIXING

Classical Anderson mixing pairs each point with a residual $r_j = G(x_j) - x_j$ of a single map G. In asynchronous federated learning, each r_j^{base} instead depends on a stale snapshot, a client objective, and a stochastic local trajectory. The resulting mismatch affects both the residual fit and the displacement of the mixed candidate. To make this explicit, we rewrite the direct Anderson candidate (1) as

$$x_{k+1} - x_k = \underbrace{\sum_{j \in \mathcal{H}_k} \alpha_{k,j} (x_j - x_k)}_{\text{history drift}} + \underbrace{\sum_{j \in \mathcal{H}_k} \alpha_{k,j} r_j^{\text{base}}}_{\text{fitted residual}}.$$

The residual least-squares solve controls the second term, while the first is the displacement induced by mixing past server states. In classical Anderson, a common map G links the two. For asynchronous local updates, the fitted residual and the movement of the stored models must instead be controlled together through the full step from x_k.

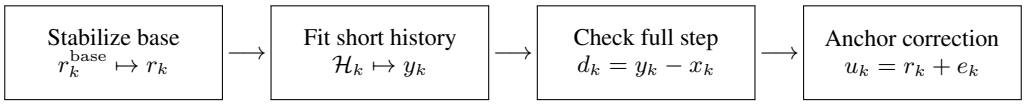

Figure 1: HERA's server update. The stabilized residual defines the base candidate and the reference for the history correction. A rejected or unavailable proposal gives $e_k = 0$.

Client heterogeneity and stochastic noise further affect the relative scales and directions of the arriving updates. HERA addresses these effects by stabilizing the residual surrogates, fitting a short history, and checking the resulting displacement against the current base step.

3 HERA: History-Enhanced Residual Acceleration

HERA combines a stabilized asynchronous base step with a bounded correction from residual history (Figure 1). Let r_k be the stabilized base residual, $b_k = x_k + r_k$ its candidate, and y_k a history proposal. The committed update has the form

$$x_{k+1} = b_k + e_k, \qquad e_k = \mathbf{1}_{\text{acc},k}(1 - \rho_k)(y_k - b_k), \qquad \|e_k\| \le \Gamma_k \|r_k\|. \tag{2}$$

The indicator $\mathbf{1}_{\text{acc},k}$ records acceptance and is zero when no proposal is available. The anchor budget Γ_k bounds the history correction relative to the stabilized base residual.

3.1 Residual stabilization

Let r_k^{base} denote the displacement supplied by the asynchronous base updater, such as the delay-weighted update $w(\tau_k)\Delta_k$. HERA applies a lagged coordinate-wise Root Mean Square (RMS) normalization:

$$q_k = \text{clip}\big(\sqrt{v_{k-1}} + \varepsilon_{\text{rms}}, q_{\min}, q_{\max}\big), \qquad r_k = r_k^{\text{base}} \oslash q_k. \tag{3}$$

Here v_{k-1} is available before the current update, $\varepsilon_{\text{rms}} > 0$, and clipping is coordinate-wise. After forming the committed step, the state is updated as

$$v_k = \beta v_{k-1} + (1 - \beta) r_k^{\text{base}} \odot r_k^{\text{base}}, \qquad v_{-1} = 0.$$

Taking $q_k \equiv \mathbf{1}$ disables RMS. Finite $q_{\min}$ and $q_{\max}$ bound the coordinate-wise rescaling; setting $q_{\max} = +\infty$ retains only the denominator floor.

Normalization shapes both the base update and the history fit. Without rescaling, coordinates with persistently large updates can dominate the Euclidean least-squares objective. RMS balances their contributions, while lagging the state separates the current update from its scale estimate.

3.2 Residual-history proposal

For a retained window of candidate–residual pairs (x_j, r_j), let $R_k = [\, r_j \,]_{j \in \mathcal{H}_k}$. The affine Anderson fit is

$$\alpha_k = \arg\min_{\mathbf{1}^\top \alpha = 1} \|R_k \alpha\|^2 + \lambda \|\alpha\|^2, \qquad \widetilde{y}_k = \sum_{j \in \mathcal{H}_k} \alpha_{k,j}(x_j + r_j), \tag{4}$$

where $\lambda > 0$. With $M_k = R_k^\top R_k + \lambda I$, one solves $M_k z = \mathbf{1}$ and sets $\alpha_k = z/(\mathbf{1}^\top z)$. Positive definiteness gives a unique solution. The residual term fits a combination of recent directions, and the ridge term controls coefficient size. The affine coefficients may be negative, allowing extrapolation while keeping each stored model paired with its residual.

The history policy selects retained pairs and refresh times. Appendix B.5 gives the equivalent difference-coordinate form of the fit. Candidate damping or norm clipping can be applied before the acceptance test; without either operation, $y_k = \widetilde{y}_k$.

Algorithm 1 HERA: stabilized base residual and anchored history correction

Input: asynchronous base updater, history policy, RMS parameters, κ_{dir}, Γ_{step}, nominal mixing ρ_k^0, and anchor budget Γ_k
Initialize x_0, $v_{-1} = 0$, and the history state
for server update k **do**
  Obtain r_k^{base} from the completed arrivals and base-updater state
  Compute q_k, r_k by (3); set $b_k \leftarrow x_k + r_k$ and $e_k \leftarrow 0$
  Refresh eligible history records according to the history policy
  if a history proposal is scheduled and sufficient records are available **then**
    Fit the retained history and form y_k; set $d_k \leftarrow y_k - x_k$
    if the numerical checks and (5) pass **then**
      Set ρ_k by (6)
      $e_k \leftarrow (1-\rho_k)(d_k - r_k)$
    end if
  end if
  $x_{k+1} \leftarrow b_k + e_k$
  $v_k \leftarrow \beta v_{k-1} + (1-\beta) r_k^{\mathrm{base}} \odot r_k^{\mathrm{base}}$
end for

3.3 Base-anchored acceptance

Write $d_k = y_k - x_k$ and choose $0 \le \kappa_{\mathrm{dir}} \le \Gamma_{\mathrm{step}}$. A proposal is eligible only if its coefficients and displacement are finite and

$$\langle d_k, r_k \rangle \ge \kappa_{\mathrm{dir}} \|r_k\|^2, \qquad \|d_k\| \le \Gamma_{\mathrm{step}} \|r_k\|. \tag{5}$$

These checks assess the complete displacement, including the history drift identified in Section 2. Rejection commits b_k. On acceptance, the server uses

$$x_{k+1} = (1-\rho_k) y_k + \rho_k b_k, \qquad (1-\rho_k)\|d_k - r_k\| \le \Gamma_k \|r_k\|.$$

Given a nominal mixing weight $\rho_k^0 \in [0,1]$, the anchor can be implemented by

$$1 - \rho_k = \min\left\{1 - \rho_k^0, \frac{\Gamma_k \|r_k\|}{\|d_k - r_k\|}\right\}, \tag{6}$$

with $\rho_k = \rho_k^0$ when $d_k = r_k$. Thus the anchor reduces the history contribution only when necessary.

The direction and step tests also imply

$$\|e_k\| \le (1-\rho_k)\sqrt{\Gamma_{\mathrm{step}}^2 + 1 - 2\kappa_{\mathrm{dir}}}\, \|r_k\|. \tag{7}$$

In particular, a fixed $\rho_k = 0.8$, a unit step cap, and positive alignment give $\|e_k\| \le (\sqrt{2}/5)\|r_k\|$. Thus fixed mixing induces an explicit relative anchor budget, while (6) directly supports a stepsize-scaled budget.

For the weighted local-SGD configuration, the base updater returns $w(\tau_k)\Delta_k$, the history retains recent pairs (x_k, r_k), and the fit follows (4). Dataset-specific choices are given in Appendix C.

All history operations take place at the server. Storing m model-sized records costs $O(dm)$ space, forming their Gram matrix costs $O(dm^2)$, and the coefficient solve has dimension m. Proposal construction and acceptance use the stored records and current base residual, without additional client queries.

4 Convergence analysis

We analyze HERA through the decomposition $u_k = r_k + e_k$ in (2). The global result controls the accumulated effects of the base residual, delay, and history correction; the local result relates residual fitting to proposal error near a contractive solution map.

4.1 GLOBAL CONVERGENCE

Assumption 4.1. *The global objective f is bounded from below by $f_* > -\infty$. Moreover, each client objective f_i is L-smooth:*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|, \qquad \forall i \in \{1, \dots, N\},\ \forall x, y.$$

Consequently, the averaged objective $f = N^{-1}\sum_{i=1}^N f_i$ is also L-smooth.

Assumption 4.2. *Let $\mathcal{I}_k$ contain the pre-update variables $x_k, x_{s_k}, s_k, v_{k-1}$, the local stepsize schedule, and the retained history before inserting (x_k, r_k). The uniform-sampling regime assumes $\Pr(i_k = i \mid \mathcal{I}_k) = 1/N$. For the local-SGD trajectory below, $\mathcal{F}_{k,t}$ contains $\mathcal{I}_k$, i_k, and the first t local mini-batch draws, so that $z_{k,t}$ is $\mathcal{F}_{k,t}$-measurable. Under this conditioning, for $t = 0, \dots, T_k - 1$,*

$$\mathbb{E}[g_{k,t} \mid \mathcal{F}_{k,t}] = \nabla f_{i_k}(z_{k,t}), \qquad \mathbb{E}\big[\|g_{k,t} - \nabla f_{i_k}(z_{k,t})\|^2 \mid \mathcal{F}_{k,t}\big] \le \sigma^2.$$

Assumption 4.3. *There exists $\zeta^2 \ge 0$ such that, for all x, $\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x) - \nabla f(x)\|^2 \le \zeta^2$.*

Assumption 4.4. *The staleness is bounded: $0 \le \tau_k \le \tau_{\max}$.*

For the weighted local-SGD base updater, server update k processes one client trajectory starting from the stale snapshot x_{s_k}:

$$z_{k,0} = x_{s_k}, \qquad z_{k,t+1} = z_{k,t} - \gamma_{k,t} g_{k,t}, \qquad t = 0, \dots, T_k - 1,$$

with nonnegative inner stepsizes and returned update $\Delta_k = z_{k,T_k} - x_{s_k}$. We write $\eta_k := \sum_{t=0}^{T_k-1} \gamma_{k,t}$ for the total inner stepsize. The base residual is $r_k = w(\tau_k) P_k \Delta_k$, with $P_k = \mathrm{Diag}(q_k)^{-1}$. The stepsize and anchor schedules are deterministic.

Theorem 4.1. *Consider Algorithm 1 with the weighted local-SGD base residual defined above. Suppose Assumptions 4.1–4.4 hold and the algorithmic parameters satisfy $0 < w_{\min} \le w(\tau) \le w_{\max} < \infty$ for $0 \le \tau \le \tau_{\max}$, $0 < q_{\min} \le q_{\max} < \infty$, finite Γ_{step}, and $\Gamma_k \le \Gamma_{\mathrm{anc}} \eta_k$ for some fixed $\Gamma_{\mathrm{anc}} > 0$. Then there exist constants $a, B_h, B_\tau > 0$ and $\eta_* > 0$, depending only on these fixed problem and algorithmic constants, such that if $\eta_k \le \eta_*$ for all k, then for every $K \ge 1$,*

$$\sum_{k=0}^{K-1} \eta_k\, \mathbb{E}\|\nabla f(x_{s_k})\|^2 \le \frac{2(f(x_0) - f_*)}{a} + \frac{2(B_h + B_\tau)}{a}(\sigma^2 + \zeta^2) \sum_{k=0}^{K-1} \eta_k^2.$$

Corollary 4.1. *Under the assumptions of Theorem 4.1, given $\gamma_0 > 0$ and $T_k = T$, choose $\gamma_{k,t} = \gamma_0 T^{-1} K^{-1/2}$, so that $\eta_k = \gamma_0 K^{-1/2}$, and assume $\gamma_0 K^{-1/2} \le \eta_*$. Then it holds that*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|^2 = O(K^{-1/2}).$$

Appendix A proves these results, including the transfer to current iterates. Proposition A.2 extends the argument to base residuals with conditional drift δ_k and a general anchor budget; uniform local SGD has $\delta_k = 0$.

Nonuniform arrivals. Completion order need not satisfy conditional uniformity. Write $p_{k,i} = \Pr(i_k = i \mid \mathcal{I}_k)$ and define the selection bias and sampling discrepancy

$$\mathfrak{b}_k(x) = \sum_{i=1}^N \left(p_{k,i} - \frac{1}{N}\right) \nabla f_i(x), \qquad \chi_k^2 = N \sum_{i=1}^N \left(p_{k,i} - \frac{1}{N}\right)^2. \tag{8}$$

Choose a deterministic $\Lambda \in [1, N]$ with $N \max_i p_{k,i} \le \Lambda$ almost surely, and set $V_\Lambda = \sigma^2 + \Lambda\zeta^2$ and $\bar{\chi}_K^2 = K^{-1}\sum_{k<K} \mathbb{E}[\chi_k^2]$. Assumption 4.3 gives $\|\mathfrak{b}_k(x)\|^2 \le \chi_k^2 \zeta^2$.

Theorem 4.2 (Nonuniform arrivals)**.** *Suppose Assumptions 4.1–4.4 hold, with the uniformity clause replaced by the conditional law $p_{k,i}$. Retain the bounds on w, q, and Γ_{step} in Theorem 4.1. There exist $a_b, c_b, \eta_b > 0$ such that, for $0 < \eta_k \equiv \eta \le \eta_b$ and $\Gamma_k \le \Gamma$ satisfying*

$$a_b - \sqrt{c_b}\Gamma - C_\tau^{\mathrm{res}} \eta \ge a_{\mathrm{res}} > 0, \qquad C_\tau^{\mathrm{res}} = L \max\{1, \Gamma_{\mathrm{step}}\}^2 c_b(\tau_{\max} + 1/2),$$

one has

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|^2 = O\left(\frac{f(x_0) - f_*}{K\eta} + (\eta + \Gamma)V_\Lambda + \zeta^2 \bar{\chi}_K^2\right). \tag{9}$$

The hidden constant is independent of K, η, Γ for a fixed positive margin a_{res}.

For $\eta = \gamma_0 K^{-1/2}$ and $\Gamma = O(\eta)$, the bound is $O(K^{-1/2} + \zeta^2 \bar{\chi}_K^2)$; a fixed Γ retains the term ΓV_Λ. Appendix A.5 proves the result with $\delta_k = C_{\mathrm{sel}} \|\mathfrak{b}_k(x_{s_k})\|^2$. No independence across arrivals or positive lower bound on $p_{k,i}$ is required.

For the CIFAR latency law, Proposition A.4 gives long-run completion fractions $\pi_i \propto [1 + (i \bmod 3)g + j/2]^{-1}$. These are not the conditional probabilities $p_{k,i}$: even balanced completion counts need not imply conditional uniformity. The theorem permits a queue-determined next identity by taking $\Lambda = N$. Importance correction requires positive conditional propensities and is a separate theoretical variant (Appendix A.6).

4.2 Local improvement

For arrival k, let $\Delta_k(x)$ be the update of the same local-SGD procedure started at x, with the current inner stepsizes and client-side sampling law, and define $\bar{\Delta}_k(x) := \mathbb{E}[\Delta_k(x) \mid \mathcal{I}_k]$. The corresponding exact mean local-SGD map is

$$\Psi_k(x) = x + w(\tau_k) P_k \bar{\Delta}_k(x), \qquad P_k = \mathrm{Diag}(q_k)^{-1}.$$

Let $b_k := x_k + r_k$ be the stabilized base candidate, used as the fallback when a proposal is rejected. On the accepted branch of Algorithm 1, one has

$$x_{k+1} = x_k + (1-\rho_k) d_k + \rho_k r_k = (1-\rho_k) y_k + \rho_k b_k, \qquad d_k = y_k - x_k.$$

We use the Euclidean norm throughout, consistent with the regularized residual fit in Algorithm 1.

Assumption 4.5. *Consider an accepted arrival k. There exists a neighborhood $\mathcal{B}$ of a local minimizer $x^\star$ such that $\Psi_k \in C^1(\mathcal{B})$ and*

$$\|\Psi_k(x) - x^\star\|_2 \le \nu \|x - x^\star\|_2, \qquad \forall x \in \mathcal{B},$$

for some $\nu \in (0,1)$. In addition, $x_j,\ b_j := x_j + r_j \in \mathcal{B},\ \forall j \in \mathcal{H}_k \cup \{k\},\ k \in \mathcal{H}_k$.

On the retained window we write the base candidates as $b_j = \Psi_k(x_j) + \varepsilon_{j,k},\ j \in \mathcal{H}_k \cup \{k\}$, which defines the evaluation errors $\varepsilon_{j,k}$ relative to the current mean local-SGD map Ψ_k. Also write $\bar{\varepsilon}_k := \max_{j \in \mathcal{H}_k \cup \{k\}} \|\varepsilon_{j,k}\|_2$. For the affine history fit (4), define the observable residual-fit ratio

$$\Gamma_k^\lambda := \frac{\left(\|R_k \alpha_k\|_2^2 + \lambda \|\alpha_k\|_2^2\right)^{1/2}}{\left(\|r_k\|_2^2 + \lambda\right)^{1/2}}, \qquad R_k = [\, r_j \,]_{j \in \mathcal{H}_k}.$$

Since $k \in \mathcal{H}_k$, the current residual column is a feasible comparator, so $\Gamma_k^\lambda \le 1$.

The errors $\varepsilon_{j,k}$ measure departures of the stored pairs from the current contractive mean map. The next result compares the accepted proposal with the base candidate.

Theorem 4.3. *Under Assumption 4.5, the base candidate satisfies*

$$\|b_k - x^\star\|_2 \le \nu \|x_k - x^\star\|_2 + \bar{\varepsilon}_k.$$

If the history proposal is accepted and $b_k \ne x^\star$, define $\theta_k := \frac{\|y_k - x^\star\|_2}{\|b_k - x^\star\|_2}$. Then the accepted HERA update satisfies

$$\|x_{k+1} - x^\star\|_2 \le \left[\rho_k + (1-\rho_k)\theta_k\right] \|b_k - x^\star\|_2.$$

Therefore, whenever $\theta_k < 1$ and $\rho_k < 1$, the accepted step has smaller error than the base candidate at the same arrival.

The solution-dependent ratio θ_k is not an acceptance test. The next result bounds proposal error using residual fitting, map conditioning, and evaluation mismatch.

Corollary 4.2. *Under the assumptions of Theorem 4.3, consider the affine fit (4). Suppose the segments from $x^\star$ to the retained points lie in $\mathcal{B}$, $D\Psi_k$ is L_Ψ-Lipschitz there, and $\|\alpha_k\|_1 \le A_{\mathcal{B}}$. Define*

$$\delta_{\mathcal{B},k} := \max_{j \in \mathcal{H}_k \cup \{k\}} \|x_j - x^\star\|_2.$$

Let $\omega_k = \|y_k - \widetilde{y}_k\|_2$ be the change due to candidate processing. Then $\|y_k - x^\star\|_2 \le U_k$, where

$$U_k := \frac{\nu}{1-\nu} \|R_k \alpha_k\|_2 + \frac{A_{\mathcal{B}}}{1-\nu} \left(\frac{L_\Psi}{2} \delta_{\mathcal{B},k}^2 + \bar{\varepsilon}_k \right) + \omega_k.$$

On the accepted branch,

$$\|x_{k+1} - x^\star\|_2 \le \rho_k\big(\nu\|x_k - x^\star\|_2 + \bar{\varepsilon}_k\big) + (1-\rho_k)U_k.$$

For $\rho_k < 1$, this error bound is smaller than the base-candidate bound whenever $U_k < \nu\|x_k - x^\star\|_2 + \bar{\varepsilon}_k$; Theorem 4.3 gives the corresponding condition for actual error reduction.

The residual-fit ratio gives $\|R_k\alpha_k\|_2 \le \Gamma_k^\lambda(\|r_k\|_2^2+\lambda)^{1/2}$. Proposal quality therefore depends jointly on residual fitting, map conditioning through $1/(1-\nu)$, regularization, and evaluation mismatch, consistent with residual-based local analyses of Anderson acceleration (Toth et al., 2017; Evans et al., 2020). Appendix B proves the bound and decomposes $\bar{\varepsilon}_k$.

5 Experiments

The experiments separate mechanism and end-to-end performance. MNIST provides a four-way ablation of RMS stabilization and history correction, while CIFAR-10 and CIFAR-100 compare HERA with six asynchronous baselines across ten combinations of client count, data partition, and delay.

Ablation on MNIST. Table 1 compares first-order asynchronous updates, RMS alone, history alone, and their combination. RMS improves the first-order baseline, and adding history to RMS raises mean final accuracy in all five settings. The largest additional gains occur for label-sorted and Dirichlet non-IID data: 10.55 and 9.08 percentage points, respectively. History alone gives smaller changes, indicating that stabilization and history reuse are most effective together in these runs.

Table 1: MNIST ablation across five asynchronous settings. Final accuracy (%) is the mean ± population standard deviation over ten runs.

Setting	First-order async	RMS only	History only	HERA
Base IID	89.28 ± 1.27	97.19 ± 1.06	89.36 ± 1.16	**98.20 ± 0.42**
Label-Sorted Non-IID	74.16 ± 9.25	82.90 ± 15.86	73.73 ± 9.26	**93.45 ± 4.10**
Dirichlet Non-IID	79.72 ± 9.09	87.71 ± 20.88	81.25 ± 6.16	**96.79 ± 1.21**
High Delay IID	89.64 ± 2.49	95.17 ± 2.68	90.29 ± 0.89	**98.17 ± 0.41**
IID 20 Workers	81.47 ± 2.36	96.62 ± 1.36	81.89 ± 2.33	**97.61 ± 0.72**

CIFAR setup. All methods use a CIFAR-style ResNet18 with GroupNorm and the same 45,000/5,000 training/validation split, client partitions, and evaluation seeds. Each client dispatch performs one mini-batch SGD step with batch size 64. Training spans 200 nominal epochs on CIFAR-10 and 300 on CIFAR-100, with client updates processed in completion order. We report mean final validation accuracy and sample standard deviation over three runs. Appendix C.2 gives the training settings and learning rates.

We compare HERA with Delay-SGD, FedAsync, FedBuff, FADAS, CA2FL, and FedAC. The nominal base weight $\rho_k^0 = 0.8$ and geometric checks induce the fixed envelope $\Gamma = \sqrt{2}/5$. These configurations use the residual-level fixed-budget regime of Corollary A.1. Applying that bound to the calibrated base update is conditional on the residual moments and positive descent margin in Proposition A.2. Appendix C.2 details these conditions and their scope; Corollary 4.1 treats the separate stepsize-scaled anchor regime.

Results across ten scenarios. HERA achieves the highest mean final validation accuracy in eight of the ten scenarios, including all six non-IID settings (Table 2). Under joint data and delay heterogeneity, its gains over the strongest baseline are 2.23 percentage points on CIFAR-10 (M5) and 3.12 on CIFAR-100 (M10). In the twenty-client IID settings, FedAsync leads HERA by 0.64 points on CIFAR-10 (M2), and CA2FL leads by 1.04 points on CIFAR-100 (M7).

Figure 2 shows the twenty-client non-IID settings M5 and M10 with heterogeneous delays. HERA separates more clearly late in training and reaches the highest final mean in both.

Table 2: CIFAR comparison with the strongest baseline in each scenario. Accuracy is the mean ± sample standard deviation over three runs; Δ is HERA's mean minus the baseline mean, in percentage points. In M10, CA2FL ties FedBuff in mean accuracy. Appendix C.2 gives all seven methods and the scenario definitions.

Scenario	HERA	Strongest baseline	Baseline accuracy	Δ
CIFAR-10				
M1	89.660 ± 0.610	FedAsync	88.160 ± 0.790	$+1.500$
M2	83.253 ± 0.768	FedAsync	83.893 ± 1.123	-0.640
M3	87.673 ± 1.621	CA2FL	85.727 ± 1.389	$+1.946$
M4	78.860 ± 2.478	FedBuff	77.800 ± 3.632	$+1.060$
M5	75.047 ± 0.503	CA2FL	72.813 ± 1.580	$+2.234$
CIFAR-100				
M6	70.580 ± 0.832	FedBuff	70.313 ± 0.401	$+0.267$
M7	64.080 ± 0.956	CA2FL	65.120 ± 1.265	-1.040
M8	70.567 ± 0.446	CA2FL	70.087 ± 0.574	$+0.480$
M9	64.287 ± 1.230	FedBuff	63.767 ± 0.280	$+0.520$
M10	62.473 ± 1.118	FedBuff	59.353 ± 1.318	$+3.120$

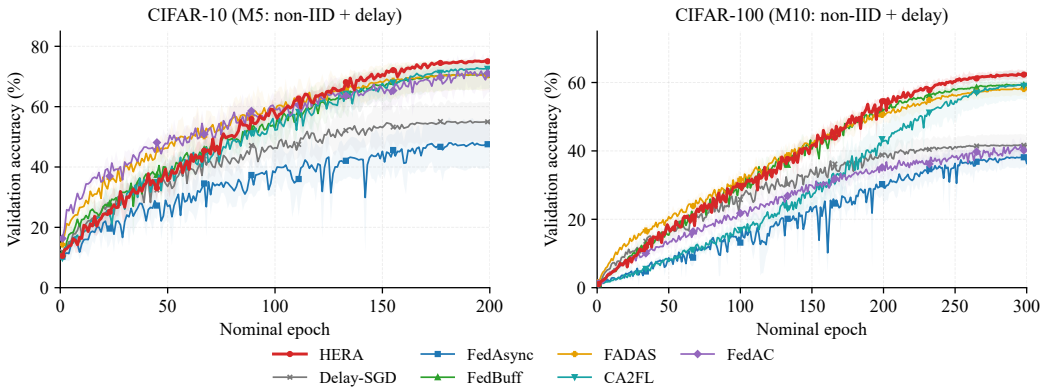

Figure 2: Learning curves for HERA and six asynchronous baselines on CIFAR-10 (M5, left) and CIFAR-100 (M10, right), with twenty non-IID clients and heterogeneous delays. The horizontal axis shows nominal epochs; lines and bands indicate mean validation accuracy and sample standard deviation over three runs.

Cross-setting pattern. The MNIST ablation supports combining stabilization with history reuse, while the largest CIFAR margins occur under joint data and delay heterogeneity. The late-stage separation is qualitatively consistent with the local analysis, where residual fitting becomes more useful in a coherent neighborhood. The contraction condition remains an analytical assumption.

6 Conclusion

HERA reuses asynchronous client updates through a stabilized residual history and a base-anchored server correction. The analysis separates base-update drift, delay, history amplitude, and selection bias. Uniform local SGD with a stepsize-scaled anchor attains $O(K^{-1/2})$ stationarity; nonuniform arrivals and fixed anchors retain explicit residual terms. The local analysis connects proposal quality to residual fitting, map conditioning, and evaluation mismatch. On MNIST, the ablation shows that history reuse is most effective when paired with RMS stabilization. Across the ten CIFAR scenarios, HERA reaches the highest mean final validation accuracy in eight, including all six non-IID settings.

AI USE STATEMENT

Generative AI tools assisted with literature review, experimental design, code development and testing, analysis of experimental records, mathematical checks and proof revisions, and manuscript editing. The authors are responsible for the methods, proofs, results, and final manuscript.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies the residual-history update and its acceptance rule. Section 4 states the assumptions for the convergence results. Appendices A and B provide the proofs. Appendix C describes the model architectures, data partitions, asynchronous scheduler, training settings, and evaluation statistics.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Donald G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12 (4):547–560, 1965.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 344–353, 2019.

Wei Bian, Xiaojun Chen, and C. T. Kelley. Anderson acceleration for a class of nonsmooth fixed-point problems. *SIAM Journal on Scientific Computing*, 43(5):S1–S20, 2021.

Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous online federated learning for edge devices with non-IID data. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 15–24, 2020. doi: 10.1109/BigData50022.2020.9378161.

Chen Dun, Mirian Hipolito, Chris Jermaine, Dimitrios Dimitriadis, and Anastasios Kyrillidis. Efficient and light-weight federated learning via asynchronous distributed dropout. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 6630–6660, 2023.

Claire Evans, Sara Pollock, Leo G. Rebholz, and Mengying Xiao. A proof that anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically). *SIAM Journal on Numerical Analysis*, 58(1):788–810, 2020.

Mathieu Even, Anastasia Koloskova, and Laurent Massoulié. Asynchronous sgd on graphs: A unified framework for asynchronous decentralized and federated optimization. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 64–72, 2024.

Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009.

Xue Feng, M. Paul Laiu, and Thomas Strohmer. Fedosaa: Improving federated learning with one-step anderson acceleration. In *Conference on Parsimony and Learning*, pp. 985–1006, 2025.

Rustem Islamov, Mher Safaryan, and Dan Alistarh. AsGrad: A sharp unified analysis of asynchronous-sgd algorithms. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 649–657, 2024.

Zhifeng Jiang, Wei Wang, Baochun Li, and Bo Li. Pisces: Efficient federated learning via guided asynchronous training. In *Proceedings of the 13th ACM Symposium on Cloud Computing*, pp. 370–385, 2022. doi: 10.1145/3542929.3563463.

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*, 2020.

Louis Leconte, Matthieu Jonckheere, Sergey Samsonov, and Eric Moulines. Queuing dynamics of asynchronous federated learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 1711–1719, 2024.

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*, 2020.

Zilinghan Li, Pranshu Chaturvedi, Shilan He, Han Chen, Gagandeep Singh, Volodymyr Kindratenko, Eliu A. Huerta, Kibaek Kim, and Ravi Madduri. FedCompass: Efficient cross-silo federated learning on heterogeneous client devices using a computing power-aware scheduler. In *International Conference on Learning Representations*, 2024.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2015.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3043–3052, 2018.

Hailiang Liu, Jia-Hao He, and Xuping Tian. Anderson acceleration of gradient methods with energy for optimization problems. *Communications on Applied Mathematics and Computation*, 6(2): 1299–1318, 2024a.

Ji Liu, Juncheng Jia, Tianshi Che, Chao Huo, Jiaxiang Ren, Yang Zhou, Huaiyu Dai, and Dejing Dou. Fedasmu: Efficient asynchronous federated learning with dynamic staleness-aware model update. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13900–13908, 2024b.

Vien V. Mai and Mikael Johansson. Anderson acceleration of proximal gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6620–6629, 2020.

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*, 2017.

John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Michael Rabbat, Mani Malek, and Dzmitry Huba. Federated learning with buffered asynchronous aggregation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

Wenqing Ouyang, Yang Liu, and Andre Milzarek. Descent properties of an anderson accelerated gradient method with restarting. *SIAM Journal on Optimization*, 34(1):336–365, 2024.

Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. ARock: An algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.

Sara Pollock and Leo G. Rebholz. Filtering for anderson acceleration. *SIAM Journal on Scientific Computing*, 45(4):A1571–A1590, 2023.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Damien Scieur. Adaptive quasi-Newton and Anderson acceleration framework with explicit global (accelerated) convergence rates. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 883–891, 2024.

Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. In *Advances in Neural Information Processing Systems 29*, pp. 712–720, 2016.

Ningxin Su and Baochun Li. How asynchronous can federated learning be? In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pp. 1–11, 2022. doi: 10.1109/IWQoS54832.2022.9812885.

Sheng Sun, Zengqi Zhang, Quyang Pan, Min Liu, Yuwei Wang, Tianliu He, Yali Chen, and Zhiyuan Wu. Staleness-controlled asynchronous federated learning: Accuracy and efficiency tradeoff. *IEEE Transactions on Mobile Computing*, 23(12):12621–12634, 2024.

Ziyuan Tang, Tianshi Xu, Huan He, Yousef Saad, and Yuanzhe Xi. Anderson acceleration with truncated gram–schmidt. *SIAM Journal on Matrix Analysis and Applications*, 45(4):1850–1872, 2024.

Alex Toth and C. T. Kelley. Convergence analysis for anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015.

Alex Toth, J. Austin Ellis, Tom Evans, Steven Hamilton, C. T. Kelley, Roger Pawlowski, and Stuart Slattery. Local improvement results for anderson acceleration with inaccurate function evaluations. *SIAM Journal on Scientific Computing*, 39(5):S47–S65, 2017.

Leonie von Wahl, Niklas Heidenreich, Prasenjit Mitra, Michael Nolting, and Nicolas Tempelmeier. Data disparity and temporal unavailability aware asynchronous federated learning for predictive maintenance on transportation fleets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15420–15428, 2024. doi: 10.1609/aaai.v38i14.29467.

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

Yujia Wang, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen. Tackling the data heterogeneity in asynchronous federated learning with cached update calibration. In *International Conference on Learning Representations*, 2024a.

Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. FADAS: Towards federated adaptive asynchronous optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Fuchao Wei, Chenglong Bao, and Yang Liu. Stochastic anderson mixing for nonconvex stochastic optimization. In *Advances in Neural Information Processing Systems*, 2021.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

Wenjing Yan, Xiangyu Zhong, Xiaolu Wang, and Ying-Jun Angela Zhang. Momentum-driven adaptivity: Towards tuning-free asynchronous federated learning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yu Zang, Zhe Xue, Shilong Ou, Lingyang Chu, Junping Du, and Yunfei Long. Efficient asynchronous federated learning with prospective momentum aggregation and fine-grained correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16642–16650, 2024.

Junzi Zhang, Brendan O'Donoghue, and Stephen Boyd. Globally convergent type-i anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization*, 30(4):3170–3197, 2020.

Tuo Zhang, Lei Gao, Sunwoo Lee, Mi Zhang, and Salman Avestimehr. Timelyfl: Heterogeneity-aware asynchronous federated learning with adaptive partial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5064–5073, 2023.

Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, and Tie-Yan Liu. Asynchronous stochastic gradient descent with delay compensation. In *International Conference on Machine Learning*, pp. 4120–4129, 2017.

A GLOBAL CONVERGENCE PROOFS

We use the notation introduced in Section 4. In particular,

$$u_k = x_{k+1} - x_k, \qquad r_k = w(\tau_k) P_k \Delta_k, \qquad P_k = \mathrm{Diag}(q_k)^{-1},$$
$$q_k = \mathrm{clip}\big(\sqrt{v_{k-1}} + \varepsilon_{\mathrm{rms}},\ q_{\min}, q_{\max}\big), \qquad s_k = k - \tau_k.$$

For the multi-step local-SGD realization, arrival k starts from the stale snapshot x_{s_k} and follows

$$z_{k,0} = x_{s_k}, \qquad z_{k,t+1} = z_{k,t} - \gamma_{k,t} g_{k,t}, \qquad t = 0, \dots, T_k - 1,$$

with returned displacement

$$\Delta_k = z_{k,T_k} - x_{s_k}, \qquad \eta_k = \sum_{t=0}^{T_k - 1} \gamma_{k,t}.$$

By Assumption 4.2, conditioned on $(\mathcal{I}_k, i_k)$ and the first t local mini-batch draws, $g_{k,t}$ is unbiased at $z_{k,t}$ and has conditional variance bounded by σ^2. Since q_k depends only on v_{k-1}, the matrix P_k is $\mathcal{I}_k$-measurable. The clipping bounds on q_k and the staleness bound $0 \le \tau_k \le \tau_{\max}$ give

$$\underline{p} I \preceq P_k \preceq \bar{p} I, \qquad \underline{p} = q_{\max}^{-1}, \quad \bar{p} = q_{\min}^{-1},$$

and

$$w_{\min} = \min_{0 \le \tau \le \tau_{\max}} w(\tau) > 0, \qquad w_{\max} = \max_{0 \le \tau \le \tau_{\max}} w(\tau).$$

These bounds will be used throughout the appendix.

A.1 AUXILIARY ESTIMATES

Lemma A.1. *Fix an arrival k and condition on $(\mathcal{I}_k, i_k)$. If $\eta_k \le 1/(4L)$, then*

$$\max_{0 \le t \le T_k} \mathbb{E}\left[\|z_{k,t} - x_{s_k}\|^2 \mid \mathcal{I}_k, i_k\right] \le 3\eta_k^2 \big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big),$$

$$\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k, i_k] + \eta_k \nabla f_{i_k}(x_{s_k})\| \le \sqrt{3}\, L \eta_k^2 \sqrt{\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2},$$

and

$$\mathbb{E}\left[\|\Delta_k\|^2 \mid \mathcal{I}_k, i_k\right] \le 3\eta_k^2 \big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big).$$

Proof. For brevity, condition on $(\mathcal{I}_k, i_k)$ throughout. Let

$$D_{k,t} = \mathbb{E}\left[\|z_{k,t} - x_{s_k}\|^2\right], \qquad M_{k,t} = \mathbb{E}\left[\|g_{k,t}\|^2\right].$$

Because

$$z_{k,t} - x_{s_k} = -\sum_{\ell=0}^{t-1} \gamma_{k,\ell} g_{k,\ell},$$

weighted Cauchy gives

$$D_{k,t} \le \left(\sum_{\ell=0}^{t-1} \gamma_{k,\ell}\right) \sum_{\ell=0}^{t-1} \gamma_{k,\ell} M_{k,\ell} \le \eta_k \sum_{\ell=0}^{t-1} \gamma_{k,\ell} M_{k,\ell}.$$

By the local-SGD unbiasedness/variance conditions and the L-smoothness of f_{i_k},

$$M_{k,t} \le \mathbb{E}\left[\|\nabla f_{i_k}(z_{k,t})\|^2\right] + \sigma^2 \le 2\|\nabla f_{i_k}(x_{s_k})\|^2 + 2L^2 D_{k,t} + \sigma^2$$
$$\le 2\big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big) + 2L^2 D_{k,t}.$$

Therefore,

$$D_{k,t} \le 2\eta_k^2 \big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big) + 2L^2 \eta_k^2 \max_{0 \le \ell \le t-1} D_{k,\ell}.$$

Taking the maximum over $t \le T_k$ and using $\eta_k \le 1/(4L)$ yields

$$\max_{0 \le t \le T_k} D_{k,t} \le \frac{2}{1 - 2L^2\eta_k^2} \eta_k^2 \big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big) \le 3\eta_k^2 \big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big).$$

This proves the first claim.

For the mean local displacement,

$$\mathbb{E}[\Delta_k] + \eta_k \nabla f_{i_k}(x_{s_k}) = -\sum_{t=0}^{T_k-1} \gamma_{k,t}\, \mathbb{E}\left[\nabla f_{i_k}(z_{k,t}) - \nabla f_{i_k}(x_{s_k})\right].$$

Using L-smoothness, Jensen's inequality, and the radius bound just proved,

$$\|\mathbb{E}[\Delta_k] + \eta_k \nabla f_{i_k}(x_{s_k})\| \le L \sum_{t=0}^{T_k-1} \gamma_{k,t}\, \mathbb{E}\|z_{k,t} - x_{s_k}\| \le L\eta_k \max_{0\le t\le T_k} \sqrt{D_{k,t}}$$

$$\le \sqrt{3}\, L\eta_k^2 \sqrt{\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2}.$$

This proves the second claim.

Finally,

$$\mathbb{E}\|\Delta_k\|^2 = \mathbb{E}\left\|\sum_{t=0}^{T_k-1} \gamma_{k,t} g_{k,t}\right\|^2 \le \eta_k \sum_{t=0}^{T_k-1} \gamma_{k,t} M_{k,t}.$$

Using the bound on $M_{k,t}$ together with the first claim,

$$M_{k,t} \le 2\big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big) + 2L^2 \cdot 3\eta_k^2\big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big)$$

$$\le 3\big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big),$$

where the last step again uses $\eta_k \le 1/(4L)$. Therefore,

$$\mathbb{E}\|\Delta_k\|^2 \le 3\eta_k^2\big(\|\nabla f_{i_k}(x_{s_k})\|^2 + \sigma^2\big).$$

This proves the third claim. □

Lemma A.2. *Suppose that Assumptions 4.2 and 4.3 hold in the multi-step local-SGD sense stated above. If $\eta_k \le 1/(4L)$, then the returned local update satisfies*

$$\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k \nabla f(x_{s_k})\| \le \sqrt{3}\, L\eta_k^2 \sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2},$$

and

$$\mathbb{E}\left[\|\Delta_k\|^2 \mid \mathcal{I}_k\right] \le 3\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

Proof. By Lemma A.1, for every fixed client index $i \in \{1, \dots, N\}$,

$$\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k, i_k = i] + \eta_k \nabla f_i(x_{s_k})\| \le \sqrt{3}\, L\eta_k^2 \sqrt{\|\nabla f_i(x_{s_k})\|^2 + \sigma^2},$$

and

$$\mathbb{E}\left[\|\Delta_k\|^2 \mid \mathcal{I}_k, i_k = i\right] \le 3\eta_k^2\big(\|\nabla f_i(x_{s_k})\|^2 + \sigma^2\big).$$

Taking conditional expectations over the uniform draw of i_k and using

$$\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_{s_k}) = \nabla f(x_{s_k}),$$

we obtain

$$\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k \nabla f(x_{s_k})\| \le \frac{1}{N}\sum_{i=1}^{N} \|\mathbb{E}[\Delta_k \mid \mathcal{I}_k, i_k = i] + \eta_k \nabla f_i(x_{s_k})\|$$

$$\le \sqrt{3}\, L\eta_k^2 \left(\frac{1}{N}\sum_{i=1}^{N}\big(\|\nabla f_i(x_{s_k})\|^2 + \sigma^2\big)\right)^{1/2}.$$

Assumption 4.3 implies

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x_{s_k})\|^2 \le \|\nabla f(x_{s_k})\|^2 + \zeta^2,$$

so

$$\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k \nabla f(x_{s_k})\| \le \sqrt{3}\, L\eta_k^2 \sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2}.$$

Similarly,

$$\mathbb{E}\left[\|\Delta_k\|^2 \mid \mathcal{I}_k\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\|\Delta_k\|^2 \mid \mathcal{I}_k, i_k = i\right]$$

$$\le 3\eta_k^2 \left(\frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x_{s_k})\|^2 + \sigma^2\right) \le 3\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

□

To streamline the next bounds, define

$$a_0 = w_{\min}\underline{p}, \qquad c_0 = 3w_{\max}^2\bar{p}^2, \qquad d_0 = \sqrt{3}\, L w_{\max}\bar{p}, \qquad B_h = \frac{d_0}{2}.$$

Lemma A.3. *Suppose that Assumptions 4.2–4.4 hold. If*

$$\eta_k \le \min\left\{\frac{1}{4L}, \frac{a_0}{2d_0}\right\},$$

then the residual r_k satisfies

$$\mathbb{E}\left[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k\right] \le -\frac{a_0}{2}\eta_k\|\nabla f(x_{s_k})\|^2 + B_h\eta_k^2(\sigma^2 + \zeta^2),$$

and

$$\mathbb{E}\left[\|r_k\|^2 \mid \mathcal{I}_k\right] \le c_0\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

Proof. By definition,

$$r_k = w(\tau_k)P_k\Delta_k.$$

Since P_k is $\mathcal{I}_k$-measurable, Lemma A.2 implies

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] = w(\tau_k)\left\langle \nabla f(x_{s_k}), P_k\mathbb{E}[\Delta_k \mid \mathcal{I}_k]\right\rangle.$$

Using the identity

$$\mathbb{E}[\Delta_k \mid \mathcal{I}_k] = -\eta_k\nabla f(x_{s_k}) + \Big(\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k\nabla f(x_{s_k})\Big),$$

we get

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] = -w(\tau_k)\eta_k\langle \nabla f(x_{s_k}), P_k\nabla f(x_{s_k})\rangle$$
$$+w(\tau_k)\left\langle \nabla f(x_{s_k}), P_k\Big(\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k\nabla f(x_{s_k})\Big)\right\rangle.$$

Using $w(\tau_k) \ge w_{\min}$ and $P_k \succeq \underline{p}I$ for the first term, and $w(\tau_k) \le w_{\max}$, $\|P_k\| \le \bar{p}$, and Lemma A.2 for the second term, we obtain

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -a_0\eta_k\|\nabla f(x_{s_k})\|^2 + d_0\eta_k^2\|\nabla f(x_{s_k})\|\sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2}.$$

Since

$$\|\nabla f(x_{s_k})\|\sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2} \le \|\nabla f(x_{s_k})\|^2 + \frac{\sigma^2 + \zeta^2}{2},$$

and $\eta_k \le \min\left\{\frac{1}{4L}, \frac{a_0}{2d_0}\right\} \le a_0/(2d_0)$, it follows that

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -\left(a_0 - d_0\eta_k\right)\eta_k\|\nabla f(x_{s_k})\|^2 + \frac{d_0}{2}\eta_k^2(\sigma^2 + \zeta^2)$$
$$\le -\frac{a_0}{2}\eta_k\|\nabla f(x_{s_k})\|^2 + B_h\eta_k^2(\sigma^2 + \zeta^2).$$

For the second moment,

$$\|r_k\|^2 = w(\tau_k)^2\|P_k\Delta_k\|^2 \le w_{\max}^2\bar{p}^2\|\Delta_k\|^2.$$

Taking conditional expectations and applying Lemma A.2 yields

$$\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k] \le 3w_{\max}^2\bar{p}^2\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big) = c_0\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

□

Lemma A.4. *For every arrival k, the exact committed step can be written as*

$$u_k = r_k + e_k, \qquad e_k = \mathbf{1}_{\mathrm{acc},k}(1-\rho_k)(d_k - r_k),$$

where $\mathbf{1}_{\mathrm{acc},k}$ denotes the indicator of the accepted-history branch. Moreover, given $\Gamma_{\mathrm{anc}} > 0$, if $\Gamma_k \le \Gamma_{\mathrm{anc}}\eta_k$ and $\Gamma_u = \max\{1, \Gamma_{\mathrm{step}}\}$, then

$$\|e_k\| \le \Gamma_{\mathrm{anc}}\eta_k\|r_k\|, \qquad \|u_k\| \le \Gamma_u\|r_k\|.$$

Proof. If the proposal is rejected, or if the history is insufficient, Algorithm 1 commits $u_k = r_k$. In that case $e_k = 0$, and both estimates are immediate.

On the accepted branch,

$$u_k = (1-\rho_k)d_k + \rho_k r_k = r_k + (1-\rho_k)(d_k - r_k),$$

which gives the stated decomposition. By setting $\Gamma_k \le \Gamma_{\mathrm{anc}}\,\eta_k$ and the anchor safeguard in Algorithm 1,

$$\|e_k\| = \|(1-\rho_k)(d_k - r_k)\| \le \Gamma_k\|r_k\| \le \Gamma_{\mathrm{anc}}\eta_k\|r_k\|.$$

In addition, the step-ratio safeguard gives $\|d_k\| \le \Gamma_{\mathrm{step}}\|r_k\|$. Therefore

$$\|u_k\| = \|(1-\rho_k)d_k + \rho_k r_k\| \le (1-\rho_k)\|d_k\| + \rho_k\|r_k\| \le \big((1-\rho_k)\Gamma_{\mathrm{step}} + \rho_k\big)\|r_k\| \le \Gamma_u\|r_k\|.$$

The same bound is trivial on fallback, because then $\|u_k\| = \|r_k\| \le \Gamma_u\|r_k\|$. □

For the HERA correction and the delay transfer, let

$$a = \frac{a_0}{4}, \quad B_e = \frac{\Gamma_{\mathrm{anc}}\sqrt{c_0}}{2}, \quad c = \max\{1, \Gamma_{\mathrm{step}}\}^2 c_0, \quad C_\tau = Lc\left(\tau_{\max} + \frac{1}{2}\right), \quad B_\tau = B_e + C_\tau,$$

and define the global stepsize threshold

$$\eta_* = \min\left\{\frac{1}{4L}, \frac{a_0}{2d_0}, \frac{a_0}{4\Gamma_{\mathrm{anc}}\sqrt{c_0}}, \frac{a}{2C_\tau}\right\}.$$

Proposition A.1. *Suppose that Assumptions 4.2–4.4 hold. If $\eta_k \le \eta_*$ for all k, then the exact committed step u_k satisfies*

$$\mathbb{E}\left[\langle \nabla f(x_{s_k}), u_k\rangle \mid \mathcal{I}_k\right] \le -a\eta_k\|\nabla f(x_{s_k})\|^2 + (B_h + B_e)\eta_k^2(\sigma^2 + \zeta^2),$$

and

$$\mathbb{E}\left[\|u_k\|^2 \mid \mathcal{I}_k\right] \le c\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

Proof. By Lemma A.3,

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -\frac{a_0}{2}\eta_k\|\nabla f(x_{s_k})\|^2 + B_h\eta_k^2(\sigma^2 + \zeta^2).$$

Next, Lemma A.4, Cauchy–Schwarz, Jensen's inequality, and Lemma A.3 give

$$\mathbb{E}[\langle \nabla f(x_{s_k}), e_k\rangle \mid \mathcal{I}_k] \le \|\nabla f(x_{s_k})\|\,\mathbb{E}[\|e_k\| \mid \mathcal{I}_k] \le \Gamma_{\mathrm{anc}}\eta_k\|\nabla f(x_{s_k})\|\,\mathbb{E}[\|r_k\| \mid \mathcal{I}_k]$$
$$\le \Gamma_{\mathrm{anc}}\eta_k\|\nabla f(x_{s_k})\|\sqrt{\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k]} \le \Gamma_{\mathrm{anc}}\sqrt{c_0}\,\eta_k^2\|\nabla f(x_{s_k})\|\sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2}.$$

Using again

$$\|\nabla f(x_{s_k})\|\sqrt{\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2} \le \|\nabla f(x_{s_k})\|^2 + \frac{\sigma^2 + \zeta^2}{2},$$

we obtain

$$\mathbb{E}[\langle \nabla f(x_{s_k}), e_k\rangle \mid \mathcal{I}_k] \le \Gamma_{\mathrm{anc}}\sqrt{c_0}\,\eta_k^2\|\nabla f(x_{s_k})\|^2 + B_e\eta_k^2(\sigma^2 + \zeta^2).$$

Therefore,

$$\mathbb{E}[\langle \nabla f(x_{s_k}), u_k\rangle \mid \mathcal{I}_k] = \mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] + \mathbb{E}[\langle \nabla f(x_{s_k}), e_k\rangle \mid \mathcal{I}_k]$$
$$\le -\left(\frac{a_0}{2} - \Gamma_{\mathrm{anc}}\sqrt{c_0}\,\eta_k\right)\eta_k\|\nabla f(x_{s_k})\|^2 + (B_h + B_e)\eta_k^2(\sigma^2 + \zeta^2).$$

If $\eta_k \le \eta_*$, then $\frac{a_0}{2} - \Gamma_{\mathrm{anc}}\sqrt{c_0}\,\eta_k \ge a_0/4 = a$, which proves the descent estimate.

For the second moment, Lemma A.4 and Lemma A.3 imply

$$\mathbb{E}[\|u_k\|^2 \mid \mathcal{I}_k] \le \Gamma_u^2\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k]$$
$$\le \Gamma_u^2 c_0\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big)$$
$$= c\eta_k^2\big(\|\nabla f(x_{s_k})\|^2 + \sigma^2 + \zeta^2\big).$$

□

A.2 Proof of Theorem 4.1

For brevity, write

$$S_k = \|\nabla f(x_{s_k})\|^2.$$

Lemma A.5. *Suppose that Assumptions 4.1–4.4 hold and that the size restrictions in Proposition A.1 are satisfied. Then, for every $k \geq 0$,*

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - a\eta_k \mathbb{E}[S_k] + (B_h + B_e)\eta_k^2(\sigma^2 + \zeta^2) + \frac{Lc(\tau_{\max}+1)}{2}\eta_k^2 \mathbb{E}[S_k + \sigma^2 + \zeta^2] + \frac{L}{2}\sum_{j=k-\tau_{\max}}^{k-1} \mathbb{E}\|u_j\|^2,$$

where terms with negative indices are interpreted as zero.

Proof. By L-smoothness of f,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), u_k \rangle + \frac{L}{2}\|u_k\|^2.$$

Add and subtract $\nabla f(x_{s_k})$ to obtain

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_{s_k}), u_k \rangle + \langle \nabla f(x_k) - \nabla f(x_{s_k}), u_k \rangle + \frac{L}{2}\|u_k\|^2.$$

Apply smoothness and Young's inequality separately to each intervening update:

$$\begin{aligned}\langle \nabla f(x_k) - \nabla f(x_{s_k}), u_k \rangle &\leq L\|x_k - x_{s_k}\|\,\|u_k\| \\ &\leq L\sum_{j=s_k}^{k-1} \|u_j\|\,\|u_k\| \\ &\leq \frac{L}{2}\sum_{j=s_k}^{k-1} \|u_j\|^2 + \frac{L\tau_k}{2}\|u_k\|^2.\end{aligned}$$

This pathwise estimate also holds when $\tau_k = 0$, with an empty sum. Since $\tau_k \leq \tau_{\max}$,

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_{s_k}), u_k \rangle + \frac{L(\tau_{\max}+1)}{2}\|u_k\|^2 + \frac{L}{2}\sum_{j=k-\tau_{\max}}^{k-1} \|u_j\|^2.$$

Taking expectations and using Proposition A.1 completes the proof. □

Lemma A.6. *For any nonnegative sequence $\{q_j\}_{j\geq 0}$,*

$$\sum_{k=0}^{K-1}\sum_{j=k-\tau_{\max}}^{k-1} q_j \leq \tau_{\max}\sum_{j=0}^{K-1} q_j,$$

where $q_j = 0$ for $j < 0$.

Proof. Fix $j \in \{0, \dots, K-1\}$. The term q_j appears in the inner sum exactly when $j \in \{k - \tau_{\max}, \dots, k-1\}$, that is, when $j+1 \leq k \leq j+\tau_{\max}$. There are at most $\tau_{\max}$ such values of k. Summing over j proves the claim. □

Proof of Theorem 4.1. Start from Lemma A.5 and sum from $k = 0$ to $K-1$:

$$\mathbb{E}[f(x_K)] \leq f(x_0) - a\sum_{k=0}^{K-1} \eta_k \mathbb{E}[S_k] + (B_h + B_e)(\sigma^2 + \zeta^2)\sum_{k=0}^{K-1}\eta_k^2 + \frac{Lc(\tau_{\max}+1)}{2}\sum_{k=0}^{K-1}\eta_k^2 \mathbb{E}[S_k + \sigma^2 + \zeta^2] + \frac{L}{2}\sum_{k=0}^{K-1}\sum_{j=k-\tau_{\max}}^{k-1} \mathbb{E}\|u_j\|^2.$$

By Lemma A.6,

$$\sum_{k=0}^{K-1}\sum_{j=k-\tau_{\max}}^{k-1}\mathbb{E}\|u_j\|^2 \le \tau_{\max}\sum_{j=0}^{K-1}\mathbb{E}\|u_j\|^2.$$

Applying Proposition A.1 once more,

$$\mathbb{E}\|u_j\|^2 \le c\eta_j^2\mathbb{E}[S_j+\sigma^2+\zeta^2].$$

Hence

$$\mathbb{E}[f(x_K)] \le f(x_0) - a\sum_{k=0}^{K-1}\eta_k\mathbb{E}[S_k] + (B_h+B_e)(\sigma^2+\zeta^2)\sum_{k=0}^{K-1}\eta_k^2$$
$$+\frac{Lc(\tau_{\max}+1)}{2}\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[S_k+\sigma^2+\zeta^2] + \frac{Lc\tau_{\max}}{2}\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[S_k+\sigma^2+\zeta^2].$$

Collecting the $\eta_k^2\mathbb{E}[S_k+\sigma^2+\zeta^2]$ terms gives

$$\mathbb{E}[f(x_K)] \le f(x_0) - \sum_{k=0}^{K-1}\big(a\eta_k - C_\tau\eta_k^2\big)\mathbb{E}[S_k] + (B_h+B_\tau)(\sigma^2+\zeta^2)\sum_{k=0}^{K-1}\eta_k^2.$$

Since $f(x_K)\ge f_*$, it follows that

$$\sum_{k=0}^{K-1}\big(a\eta_k - C_\tau\eta_k^2\big)\mathbb{E}[S_k] \le \Delta_0 + (B_h+B_\tau)(\sigma^2+\zeta^2)\sum_{k=0}^{K-1}\eta_k^2,$$

where $\Delta_0 = f(x_0)-f_*$. Because $\eta_k\le\eta_*\le a/(2C_\tau)$, one has

$$a\eta_k - C_\tau\eta_k^2 \ge \frac{a}{2}\eta_k.$$

Therefore

$$\frac{a}{2}\sum_{k=0}^{K-1}\eta_k\mathbb{E}[S_k] \le \Delta_0 + (B_h+B_\tau)(\sigma^2+\zeta^2)\sum_{k=0}^{K-1}\eta_k^2.$$

Multiplying by $2/a$ yields

$$\sum_{k=0}^{K-1}\eta_k\,\mathbb{E}\|\nabla f(x_{s_k})\|^2 \le \frac{2\Delta_0}{a} + \frac{2(B_h+B_\tau)}{a}(\sigma^2+\zeta^2)\sum_{k=0}^{K-1}\eta_k^2.$$

This is exactly the claim of Theorem 4.1. □

A.3 Proof of Corollary 4.1

Proof. Under the stated schedule,

$$\eta_k \equiv \eta_K = \gamma_0 K^{-1/2}.$$

Applying Theorem 4.1 gives

$$\eta_K\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_{s_k})\|^2 \le \frac{2(f(x_0)-f_*)}{a} + \frac{2(B_h+B_\tau)}{a}(\sigma^2+\zeta^2)K\eta_K^2.$$

Since $K\eta_K = \gamma_0\sqrt{K}$ and $K\eta_K^2=\gamma_0^2$, division by $K\eta_K$ yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_{s_k})\|^2 = O(K^{-1/2}).$$

To pass from the stale point x_{s_k} to the current iterate x_k, use $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$ and the L-smoothness of f:

$$\|\nabla f(x_k)\|^2 \le 2\|\nabla f(x_{s_k})\|^2 + 2L^2\|x_k - x_{s_k}\|^2.$$

By Cauchy's inequality applied to $x_k - x_{s_k} = \sum_{j=s_k}^{k-1} u_j$,

$$\|x_k - x_{s_k}\|^2 \le \tau_{\max} \sum_{j=k-\tau_{\max}}^{k-1} \|u_j\|^2.$$

Summing over k and applying Lemma A.6 gives

$$\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|^2 \le 2 \sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_{s_k})\|^2 + 2L^2\tau_{\max}^2 \sum_{j=0}^{K-1} \mathbb{E}\|u_j\|^2.$$

Write $\bar{S}_K = K^{-1}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_{s_k})\|^2$ and $V = \sigma^2 + \zeta^2$. For constant η_K, Proposition A.1 gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|^2 \le (2 + 2L^2 c\tau_{\max}^2\eta_K^2)\bar{S}_K + 2L^2 c\tau_{\max}^2\eta_K^2 V.$$

The stepsize restriction $\eta_K \le a/[2Lc(\tau_{\max} + 1/2)]$ implies

$$2L^2 c\tau_{\max}^2\eta_K^2 \le \frac{a^2}{2c}, \qquad 2L^2 c\tau_{\max}^2\eta_K^2 \le aL\tau_{\max}\eta_K.$$

Consequently,

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(x_k)\|^2 \le \left(2 + \frac{a^2}{2c}\right)\bar{S}_K + aL\tau_{\max}\eta_K V = O(K^{-1/2}).$$

The claimed bound is proved. □

A.4 Residual-level descent and anchor schedules

The following result expresses the descent argument in terms of the base residual itself. It allows the base updater to maintain a cache or aggregate arrivals, with its conditional descent and second moment appearing as explicit hypotheses. The retained-history representation enters through the committed correction, rather than through the coefficient solver.

Proposition A.2 (Residual-level descent transfer)**.** *Let f be L-smooth and bounded below by f_*, and suppose $0 \le k - s_k \le \tau_{\max}$. Let $\eta_k > 0$ and $\Gamma_k \ge 0$ be deterministic schedules. Write $S_k = \|\nabla f(x_{s_k})\|^2$, where x_k, x_{s_k} are measurable with respect to the pre-update information $\mathcal{I}_k$. Suppose that, for constants $a_b, c_b > 0$, $B_b, V \ge 0$, and a nonnegative, integrable, $\mathcal{I}_k$-measurable drift term δ_k,*

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -a_b\eta_k S_k + B_b\eta_k^2 V + \eta_k\delta_k, \tag{10}$$

$$\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k] \le c_b\eta_k^2(S_k + V). \tag{11}$$

For HERA's committed step, assume $\|e_k\| \le \Gamma_k\|r_k\|$ and the step cap (5). Define

$$\Gamma_u = \max\{1, \Gamma_{\text{step}}\}, \qquad C_\tau^{\text{res}} = L\Gamma_u^2 c_b\left(\tau_{\max} + \frac{1}{2}\right).$$

Then, for every $K \ge 1$,

$$\sum_{k=0}^{K-1} \eta_k\left(a_b - \sqrt{c_b}\Gamma_k - C_\tau^{\text{res}}\eta_k\right)\mathbb{E}[S_k]$$

$$\le f(x_0) - f_* + (B_b + C_\tau^{\text{res}})V\sum_{k=0}^{K-1}\eta_k^2 + \frac{\sqrt{c_b}}{2}V\sum_{k=0}^{K-1}\Gamma_k\eta_k + \sum_{k=0}^{K-1}\eta_k\mathbb{E}[\delta_k]. \tag{12}$$

Proof. The anchor inequality and conditional Cauchy–Schwarz give

$$\begin{aligned}\mathbb{E}[\langle \nabla f(x_{s_k}), e_k\rangle \mid \mathcal{I}_k] &\le \Gamma_k\sqrt{S_k}\sqrt{\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k]} \\ &\le \sqrt{c_b}\Gamma_k\eta_k\sqrt{S_k}\sqrt{S_k + V} \\ &\le \sqrt{c_b}\Gamma_k\eta_k(S_k + V/2).\end{aligned}$$

Since $u_k = r_k + e_k$, adding (10) yields

$$\mathbb{E}[\langle \nabla f(x_{s_k}), u_k\rangle \mid \mathcal{I}_k] \le -(a_b - \sqrt{c_b}\Gamma_k)\eta_k S_k + B_b\eta_k^2 V + \frac{\sqrt{c_b}}{2}\Gamma_k\eta_k V + \eta_k\delta_k.$$

On acceptance, $u_k = (1-\rho_k)d_k + \rho_k r_k$ with $\rho_k \in [0,1]$; on fallback $u_k = r_k$. Thus

$$\|u_k\| \le \Gamma_u\|r_k\|, \qquad \mathbb{E}[\|u_k\|^2 \mid \mathcal{I}_k] \le \Gamma_u^2 c_b\eta_k^2(S_k + V).$$

Smoothness and the pathwise delay estimate in Lemma A.5 imply

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_{s_k}), u_k\rangle + \frac{L(\tau_{\max}+1)}{2}\|u_k\|^2 + \frac{L}{2}\sum_{j=k-\tau_{\max}}^{k-1}\|u_j\|^2.$$

Take expectations and sum over k. By Lemma A.6, each squared step in the last double sum occurs at most $\tau_{\max}$ times. The two squared-step contributions are therefore at most

$$C_\tau^{\mathrm{res}}\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[S_k + V].$$

Substitute the conditional descent estimate, telescope the objective values, and use $f(x_K) \ge f_*$. Rearrangement gives (12). □

Local-SGD residuals. Lemma A.3 verifies (10)–(11) with

$$a_b = a_0/2, \qquad B_b = B_h, \qquad c_b = c_0, \qquad V = \sigma^2 + \zeta^2, \qquad \delta_k = 0.$$

For $\Gamma_k \le \Gamma_{\mathrm{anc}}\eta_k$ and sufficiently small stepsizes, the history contribution in (12) is of order $V\sum_k \eta_k^2$. This is the mechanism behind Theorem 4.1.

Corollary A.1 (Fixed-budget residual bound)**.** *Under the hypotheses of Proposition A.2, let $\eta_k \equiv \eta > 0$ and $\Gamma_k \le \Gamma$, with*

$$a_b - \sqrt{c_b}\Gamma - C_\tau^{\mathrm{res}}\eta \ge a_{\mathrm{res}} > 0. \tag{13}$$

Define $\bar{S}_K = K^{-1}\sum_{k=0}^{K-1}\mathbb{E}[S_k]$ and $\bar{\delta}_K = K^{-1}\sum_{k=0}^{K-1}\mathbb{E}[\delta_k]$. Then

$$\bar{S}_K \le \frac{1}{a_{\mathrm{res}}}\left[\frac{f(x_0) - f_*}{K\eta} + (B_b + C_\tau^{\mathrm{res}})\eta V + \frac{\sqrt{c_b}}{2}\Gamma V + \bar{\delta}_K\right], \tag{14}$$

and the current iterates satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_k)\|^2 \le (2 + 2L^2\Gamma_u^2 c_b\tau_{\max}^2\eta^2)\bar{S}_K + 2L^2\Gamma_u^2 c_b\tau_{\max}^2\eta^2 V. \tag{15}$$

In particular, for $\eta = \gamma_0 K^{-1/2}$ and a margin a_{res} bounded away from zero uniformly in K, the current-iterate bound is

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_k)\|^2 = O\Big(K^{-1/2} + \Gamma V + \bar{\delta}_K\Big).$$

Proof. The pathwise anchor bound also holds with the deterministic envelope Γ. Substitute this envelope and $\eta_k = \eta$ in (12), use (13), and divide by $K\eta$ to obtain (14). Smoothness and the bounded-delay telescoping estimate give

$$\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(x_k)\|^2 \le 2\sum_{k=0}^{K-1}\mathbb{E}[S_k] + 2L^2\tau_{\max}^2\sum_{k=0}^{K-1}\mathbb{E}\|u_k\|^2.$$

Since $\mathbb{E}\|u_k\|^2 \le \Gamma_u^2 c_b\eta^2\mathbb{E}[S_k + V]$, division by K proves (15). Substituting $\eta = \gamma_0 K^{-1/2}$ proves the last statement. □

The finite-budget inequalities retain both the stochastic contribution ηV and the fixed-anchor contribution ΓV; the $K^{-1/2}$ form additionally uses the stated horizon-dependent stepsize. Equation (7) converts fixed nominal mixing and geometric caps into an anchor envelope, but does not ensure the positive margin (13). A residual term in this upper bound is not a matching lower bound on the limiting gradient norm.

A.5 NONUNIFORM ARRIVALS

We now replace only the conditional uniformity clause of Assumption 4.2 by $\Pr(i_k = i \mid \mathcal{I}_k) = p_{k,i}$. The local gradient conditions remain imposed after conditioning on $\mathcal{I}_k$ and i_k. In particular, marginal unbiasedness before completion-based selection does not suffice if the selection mechanism depends on the unrevealed local gradient noise. The conditional laws may otherwise vary with the server state and be dependent across arrivals.

Use $\mathfrak{b}_k$, χ_k, Λ, and V_Λ from (8) and Section 4.1.

Lemma A.7 (Selection bias and local-update moments)**.** *Under the modified sampling assumption, Assumptions 4.1 and 4.3, and $\eta_k \le 1/(4L)$, one has*

$$\|\mathfrak{b}_k(x)\|^2 \le \chi_k^2 \zeta^2, \qquad 0 \le \chi_k^2 \le \Lambda - 1, \tag{16}$$

for every x. Moreover, writing $S_k = \|\nabla f(x_{s_k})\|^2$,

$$\big\|\mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k\big(\nabla f(x_{s_k}) + \mathfrak{b}_k(x_{s_k})\big)\big\| \le \sqrt{6} L \eta_k^2 \sqrt{S_k + V_\Lambda}, \tag{17}$$

$$\mathbb{E}[\|\Delta_k\|^2 \mid \mathcal{I}_k] \le 6\eta_k^2 (S_k + V_\Lambda). \tag{18}$$

Proof. Because $\sum_i (p_{k,i} - 1/N) = 0$, subtracting $\nabla f(x)$ from each client gradient does not change $\mathfrak{b}_k(x)$. Cauchy–Schwarz and Assumption 4.3 yield

$$\begin{aligned}\|\mathfrak{b}_k(x)\|^2 &= \left\| \sum_{i=1}^N \left(p_{k,i} - \frac{1}{N}\right)(\nabla f_i(x) - \nabla f(x)) \right\|^2 \\ &\le \left[N \sum_{i=1}^N \left(p_{k,i} - \frac{1}{N}\right)^2 \right] \left[\frac{1}{N} \sum_{i=1}^N \|\nabla f_i(x) - \nabla f(x)\|^2 \right] \le \chi_k^2 \zeta^2.\end{aligned}$$

Also, $\chi_k^2 = N \sum_i p_{k,i}^2 - 1 \le N \max_i p_{k,i} - 1 \le \Lambda - 1$, proving (16).

The nonuniform second moment must be bounded separately from the mean bias. Since $p_{k,i} \le \Lambda/N$,

$$\sum_{i=1}^N p_{k,i} \|\nabla f_i(x) - \nabla f(x)\|^2 \le \Lambda \zeta^2,$$

so

$$\sum_{i=1}^N p_{k,i} \big(\|\nabla f_i(x)\|^2 + \sigma^2\big) \le 2\|\nabla f(x)\|^2 + 2\Lambda\zeta^2 + \sigma^2 \le 2\big(\|\nabla f(x)\|^2 + V_\Lambda\big). \tag{19}$$

The fixed-client estimates in Lemma A.1 hold for every client with $p_{k,i} > 0$. Averaging its mean estimate with weights $p_{k,i}$ and applying Jensen's inequality gives

$$\begin{aligned}&\left\| \mathbb{E}[\Delta_k \mid \mathcal{I}_k] + \eta_k \sum_{i=1}^N p_{k,i} \nabla f_i(x_{s_k}) \right\| \\ &\qquad \le \sqrt{3} L \eta_k^2 \left(\sum_{i=1}^N p_{k,i} \big(\|\nabla f_i(x_{s_k})\|^2 + \sigma^2\big) \right)^{1/2} \le \sqrt{6} L \eta_k^2 \sqrt{S_k + V_\Lambda}.\end{aligned}$$

The identity $\sum_i p_{k,i} \nabla f_i = \nabla f + \mathfrak{b}_k$ proves (17). Averaging the second-moment estimate in the same lemma and using (19) proves (18). Zero-probability clients contribute nothing to these sums. □

Lemma A.8 (Nonuniform base residual). *Under the hypotheses of Lemma A.7 and the bounds on w and q in Theorem 4.1, choose*

$$a_b = \frac{3a_0}{4}, \qquad c_b = 2c_0, \qquad B_b = \frac{d_0}{\sqrt{2}},$$
$$C_{\mathrm{sel}} = \frac{2w_{\max}^2\bar{p}^2}{a_0}, \qquad \eta_b = \min\left\{\frac{1}{4L}, \frac{a_0}{8\sqrt{2}d_0}\right\}.$$

For $\eta_k \le \eta_b$, the residual hypotheses (10)–(11) hold with $V = V_\Lambda$ and

$$\delta_k = C_{\mathrm{sel}}\|\mathfrak{b}_k(x_{s_k})\|^2 \le C_{\mathrm{sel}}\zeta^2\chi_k^2. \tag{20}$$

Here $a_0, c_0, d_0, \bar{p}$ are the constants already defined in Appendix A.

Proof. Both $w(\tau_k)$ and P_k are $\mathcal{I}_k$-measurable. Using $P_k \succeq \underline{p}I$ and (17),

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -a_0\eta_k S_k + w_{\max}\bar{p}\,\eta_k\sqrt{S_k}\|\mathfrak{b}_k(x_{s_k})\| + \sqrt{2}d_0\eta_k^2\sqrt{S_k}\sqrt{S_k + V_\Lambda}.$$

Young's inequality gives

$$w_{\max}\bar{p}\sqrt{S_k}\|\mathfrak{b}_k(x_{s_k})\| \le \frac{a_0}{8}S_k + \frac{2w_{\max}^2\bar{p}^2}{a_0}\|\mathfrak{b}_k(x_{s_k})\|^2.$$

Together with $\sqrt{S_k}\sqrt{S_k + V_\Lambda} \le S_k + V_\Lambda/2$, this yields

$$\mathbb{E}[\langle \nabla f(x_{s_k}), r_k\rangle \mid \mathcal{I}_k] \le -\left(\frac{7a_0}{8} - \sqrt{2}d_0\eta_k\right)\eta_k S_k + \frac{d_0}{\sqrt{2}}\eta_k^2 V_\Lambda + C_{\mathrm{sel}}\eta_k\|\mathfrak{b}_k(x_{s_k})\|^2.$$

The stepsize restriction makes the descent coefficient at least $3a_0/4 = a_b$. The second moment follows directly from (18):

$$\mathbb{E}[\|r_k\|^2 \mid \mathcal{I}_k] \le w_{\max}^2\bar{p}^2\mathbb{E}[\|\Delta_k\|^2 \mid \mathcal{I}_k] \le 2c_0\eta_k^2(S_k + V_\Lambda).$$

Finally, δ_k is $\mathcal{I}_k$-measurable and bounded by $C_{\mathrm{sel}}(\Lambda - 1)\zeta^2$, so it satisfies the integrability hypothesis of Proposition A.2. □

Proof of Theorem 4.2. Use Lemma A.8 in Corollary A.1. Since $\bar{\delta}_K \le C_{\mathrm{sel}}\zeta^2\bar{\chi}_K^2$, its stale-iterate bound is

$$\bar{S}_K \le \frac{1}{a_{\mathrm{res}}}\left[\frac{f(x_0) - f_*}{K\eta} + (B_b + C_\tau^{\mathrm{res}})\eta V_\Lambda + \frac{\sqrt{c_b}}{2}\Gamma V_\Lambda + C_{\mathrm{sel}}\zeta^2\bar{\chi}_K^2\right]. \tag{21}$$

Equation (15) transfers this bound to current iterates. The multiplicative factor is uniformly bounded for $\eta \le \eta_b$, and its remaining $O(\eta^2 V_\Lambda)$ term is absorbed into $O(\eta V_\Lambda)$. This proves (9), with constants independent of K, η, Γ for a fixed positive margin. Substitution of $\eta = \gamma_0 K^{-1/2}$ yields the two anchor regimes stated after the theorem. □

Conditional uniformity gives $\chi_k = 0$ and $\Lambda = 1$, recovering the order of Theorem 4.1; its original proof has sharper uniform-sampling constants. Identical client gradients also remove the bias term, whereas an IID partition of finite datasets does not by itself imply $\zeta = 0$. The general bound permits $p_{k,i} = 0$ and uses no independence across arrivals, beyond the stated conditional local-gradient assumptions. It is a worst-case conditional bound: when the queue reveals the next identity, $\chi_k^2 = N - 1$ even if long-run participation is balanced.

A.6 Importance correction as a modified base updater

Proposition A.3 (Positive-propensity correction). *Replace the uniformity clause of Assumption 4.2 by the conditional law $p_{k,i}$, and suppose $p_{k,i} \ge p_{\min} > 0$ almost surely for every client. Define*

$$\widehat{\Delta}_k = \frac{\Delta_k}{Np_{k,i_k}}, \qquad r_k = w(\tau_k)P_k\widehat{\Delta}_k.$$

Under the remaining hypotheses of Theorem 4.1, including stepsize-scaled anchoring, its conclusion and Corollary 4.1 hold with modified constants and stepsize threshold.

Proof. Positive conditional propensities give

$$\mathbb{E}[\widehat{\Delta}_k \mid \mathcal{I}_k] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\Delta_k \mid \mathcal{I}_k, i_k = i].$$

Averaging Lemma A.1 uniformly over clients therefore recovers the mean bound

$$\left\| \mathbb{E}[\widehat{\Delta}_k \mid \mathcal{I}_k] + \eta_k \nabla f(x_{s_k}) \right\| \le \sqrt{3} L \eta_k^2 \sqrt{S_k + \sigma^2 + \zeta^2}.$$

For the second moment,

$$\begin{aligned}\mathbb{E}[\|\widehat{\Delta}_k\|^2 \mid \mathcal{I}_k] &= \frac{1}{N^2}\sum_{i=1}^{N} \frac{\mathbb{E}[\|\Delta_k\|^2 \mid \mathcal{I}_k, i_k = i]}{p_{k,i}} \\ &\le \frac{3\eta_k^2}{N p_{\min}}(S_k + \sigma^2 + \zeta^2).\end{aligned}$$

The proof of Lemma A.3 thus gives the same descent bound and a second-moment constant $c_0/(Np_{\min})$. With this replacement, the correction and delay arguments proving Theorem 4.1 and Corollary 4.1 apply unchanged. The correction cancels the leading client-selection bias; the $O(\eta_k^2)$ multi-step local-SGD approximation error remains. □

This proposition concerns a modified updater, not the CIFAR implementation. A fully exposed completion queue may have zero conditional propensities, precluding its hypothesis. Moreover, replacing $p_{k,i}$ by a long-run completion fraction does not establish the conditional expectation identity used in the proof.

A.7 Completion-driven sampling

Proposition A.4 (Completion frequencies and bounded delay)**.** *Suppose all clients are initially dispatched at simulated time zero, each has one task in flight, and each is redispatched immediately upon completion. Let the successive latencies of client i be*

$$d_{i,\ell} = 1 + (i \bmod 3)g + \epsilon_{i,\ell}, \qquad \epsilon_{i,\ell} \sim U(0, j), \qquad g, j \ge 0,$$

with independent dispatch jitter (zero when $j = 0$) and no simulated-time server processing delay. For simulated time t, let $A_i(t)$ count client i's completed tasks, and write $\mu_i = 1 + (i \bmod 3)g + j/2$. Then, almost surely,

$$\lim_{t\to\infty} \frac{A_i(t)}{\sum_{\ell=1}^{N} A_\ell(t)} = \pi_i := \frac{\mu_i^{-1}}{\sum_{\ell=1}^{N} \mu_\ell^{-1}}. \tag{22}$$

For single-arrival server commits, staleness is bounded, for example, by

$$\tau_{\max} = (N-1)\big(\lceil d_{\max}/d_{\min}\rceil + 1\big), \qquad d_{\min} = 1, \quad d_{\max} = 1 + 2g + j.$$

Proof. Immediate redispatch makes client i's m-th completion time $t_{i,m} = \sum_{\ell=1}^{m} d_{i,\ell}$. The strong law gives $t_{i,m}/m \to \mu_i > 0$ almost surely. Inverting $t_{i,A_i(t)} \le t < t_{i,A_i(t)+1}$ yields $A_i(t)/t \to \mu_i^{-1}$. Dividing by the sum of these limits proves (22). The same fractions hold along the merged completion stream: processing a simultaneous set one arrival at a time changes each count by at most N, which is negligible relative to the total number of completions.

For the delay bound, a task is in flight for at most $d_{\max}$ simulated time. Each other client has completion gaps at least $d_{\min}$, so it contributes at most $\lceil d_{\max}/d_{\min}\rceil + 1$ completions during that interval, including possible boundary ties. The in-flight client cannot complete another task before its own return. Counting the other clients' commits proves the bound. □

If the μ_i differ, the limiting fractions are nonuniform. Conditional uniformity at every single-arrival commit would instead make every expected empirical client fraction equal to $1/N$; bounded convergence would then force $\pi_i = 1/N$, a contradiction. Conversely, equal limiting fractions do not establish conditional uniformity: equal deterministic latencies with a fixed tie order yield balanced

counts but a predictable next identity. Including the full event-queue state in $\mathcal{I}_k$ likewise makes the next identity measurable, with $p_{k,i} \in \{0, 1\}$; this is allowed in Theorem 4.2, but not Proposition A.3.

Equation (22) characterizes completion counts, not a conditional sampling law or an optimization limit. In particular, it does not justify substituting π_i for $p_{k,i}$ in (8), or asserting convergence to $\sum_i \pi_i f_i$ in the presence of correlated delays, weighting, preconditioning, and cached updates.

B Proof of Theorem 4.3 and Corollary 4.2

We use the notation of Section 4.2. In particular,

$$\Psi_k(x) := x + w(\tau_k) P_k \bar{\Delta}_k(x), \qquad \bar{\Delta}_k(x) := \mathbb{E}[\Delta_k(x) \mid \mathcal{I}_k].$$

B.1 Proof of Theorem 4.3

Proof. By Assumption 4.5,

$$\|\Psi_k(x_k) - x^\star\|_2 \le \nu \, \|x_k - x^\star\|_2.$$

Using $b_k = \Psi_k(x_k) + \varepsilon_{k,k}$ therefore gives

$$\|b_k - x^\star\|_2 \le \nu \, \|x_k - x^\star\|_2 + \|\varepsilon_{k,k}\|_2 \le \nu \, \|x_k - x^\star\|_2 + \bar{\varepsilon}_k.$$

On the accepted branch of Algorithm 1,

$$x_{k+1} = x_k + (1 - \rho_k) d_k + \rho_k r_k, \qquad d_k = y_k - x_k, \qquad b_k = x_k + r_k.$$

Therefore

$$x_{k+1} = (1 - \rho_k) y_k + \rho_k b_k.$$

Subtracting $x^\star$ and using the triangle inequality in $\|\cdot\|_2$ gives

$$\|x_{k+1} - x^\star\|_2 \le (1 - \rho_k)\|y_k - x^\star\|_2 + \rho_k \|b_k - x^\star\|_2.$$

If $b_k \neq x^\star$, define

$$\theta_k := \frac{\|y_k - x^\star\|_2}{\|b_k - x^\star\|_2}.$$

Then

$$\|x_{k+1} - x^\star\|_2 \le \big[\rho_k + (1 - \rho_k)\theta_k\big] \|b_k - x^\star\|_2.$$

In particular, whenever $\theta_k < 1$, the history term improves the comparison factor, and this improvement is strict as soon as $\rho_k < 1$. □

B.2 Proof of Corollary 4.2

Proof. Fix the current map Ψ_k and write $J = D\Psi_k(x^\star)$ and $e_j = x_j - x^\star$. Assumption 4.5 implies $\Psi_k(x^\star) = x^\star$ and $\|J\|_2 \le \nu$. Taylor's formula along the segment from $x^\star$ to x_j gives

$$\Psi_k(x_j) = x^\star + J e_j + h_j, \qquad \|h_j\|_2 \le \frac{L_\Psi}{2} \|e_j\|_2^2.$$

Since $b_j = \Psi_k(x_j) + \varepsilon_{j,k}$ and $r_j = b_j - x_j$,

$$r_j = (J - I) e_j + h_j + \varepsilon_{j,k}.$$

Set $z = \sum_j \alpha_{k,j} e_j$ and $c = \sum_j \alpha_{k,j}(h_j + \varepsilon_{j,k})$. The affine constraint $\sum_j \alpha_{k,j} = 1$ yields

$$R_k \alpha_k = (J - I) z + c, \qquad \widetilde{y}_k - x^\star = J z + c,$$

and

$$\|c\|_2 \le A_{\mathcal{B}} \left(\frac{L_\Psi}{2} \delta_{\mathcal{B},k}^2 + \bar{\varepsilon}_k \right).$$

Because $\|J\|_2 \le \nu < 1$, the matrix $J - I$ is invertible and $\|(J - I)^{-1}\|_2 \le (1 - \nu)^{-1}$. Eliminating z gives the exact identity

$$\widetilde{y}_k - x^\star = J(J - I)^{-1} R_k \alpha_k - (J - I)^{-1} c.$$

Taking norms bounds the fitted candidate $\widetilde{y}_k$. The triangle inequality $\|y_k - x^\star\|_2 \le \|\widetilde{y}_k - x^\star\|_2 + \omega_k$ then gives $\|y_k - x^\star\|_2 \le U_k$, including the candidate-processing term in the corollary. Combining this bound with $x_{k+1} - x^\star = \rho_k(b_k - x^\star) + (1-\rho_k)(y_k - x^\star)$ and the base-candidate bound in Theorem 4.3 proves the committed-step bound. The stated comparison follows by subtracting the base-candidate upper bound.

Finally, the current-column selector is feasible in the ridge problem, so

$$\|R_k\alpha_k\|_2^2 + \lambda\|\alpha_k\|_2^2 \le \|r_k\|_2^2 + \lambda.$$

This proves $\Gamma_k^\lambda \le 1$ and the claimed bound on the fitted residual. The inverse-map factor and regularization term enter the proposal bound through U_k. □

B.3 Evaluation mismatch and special cases

For the weighted local-SGD base residual, write $w_j = w(\tau_j)$, $P_j = \mathrm{Diag}(q_j)^{-1}$, and $\bar{\Delta}_j(x) = \mathbb{E}[\Delta_j(x) \mid \mathcal{I}_j]$. The mismatch admits the exact decomposition

$$\begin{aligned}\varepsilon_{j,k} = {} & w_j P_j\big(\Delta_j - \bar{\Delta}_j(x_{s_j})\big) \\ & + w_j P_j\big(\bar{\Delta}_j(x_{s_j}) - \bar{\Delta}_j(x_j)\big) \\ & + (w_j P_j - w_k P_k)\bar{\Delta}_j(x_j) \\ & + w_k P_k\big(\bar{\Delta}_j(x_j) - \bar{\Delta}_k(x_j)\big).\end{aligned}$$

The four terms respectively separate the realized local-update noise, the stale-snapshot displacement, the change in delay weighting and RMS normalization, and the change in the mean local-update map.

For a quantitative conditional bound, suppose $\bar{\Delta}_j$ is $L_{\Delta,j}$-Lipschitz on the segment joining x_{s_j} and x_j. With $\xi_j = \|\Delta_j - \bar{\Delta}_j(x_{s_j})\|_2$ and $W = w_{\max}/q_{\min}$,

$$\begin{aligned}\|\varepsilon_{j,k}\|_2 \le {} & W\big(\xi_j + L_{\Delta,j}\|x_{s_j} - x_j\|_2\big) \\ & + \|w_j P_j - w_k P_k\|_2\,\|\bar{\Delta}_j(x_j)\|_2 \\ & + W\|\bar{\Delta}_j(x_j) - \bar{\Delta}_k(x_j)\|_2.\end{aligned}$$

The stale displacement is bounded by $\|x_{s_j} - x_j\|_2 \le \sum_{\ell=s_j}^{j-1}\|x_{\ell+1} - x_\ell\|_2$. For a single mean-gradient step $\bar{\Delta}_j(x) = -\eta_j\nabla f(x)$, one can take $L_{\Delta,j} = \eta_j L$. For multi-step local SGD, $L_{\Delta,j}$ is the assumed Lipschitz constant of the corresponding mean local-update map.

Synchronous arrivals eliminate the stale-snapshot term. Exact evaluation of the conditional mean update eliminates ξ_j; deterministic gradients on a randomly selected heterogeneous client can still differ from this mean. If all retained candidates are exact evaluations of one frozen map, then $b_j = \Psi_k(x_j)$ and $\bar{\varepsilon}_k = 0$. With no candidate processing ($\omega_k = 0$), $\lambda = 0$, and $\rho_k = 0$, the accepted proposal is the standard affine Anderson update. For positive ridge and safeguarded mixing, the bounds above retain the corresponding regularization and anchoring effects.

B.4 A single-step strongly convex example

The following single-step strongly convex example makes the frozen-map assumption and its error terms explicit. Consider one local gradient step with the conditional uniform client sampling of Assumption 4.2, so that $\bar{\Delta}_j(x) = -\eta_j\nabla f(x)$. Suppose the RMS denominator is scalar on the retained window, $P_j = q_j^{-1}I$, and write $a_j = w(\tau_j)\eta_j/q_j$. On a convex ball $\mathcal{B}$ containing $x^\star$, assume $\nabla f(x^\star) = 0$,

$$0 < \mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le M\|x - y\|_2.$$

For $0 < a_k < 2/L$, the current frozen map is $\Psi_k(x) = x - a_k\nabla f(x)$, with

$$\nu_k = \max\{|1 - a_k\mu|, |1 - a_k L|\} < 1, \qquad L_\Psi = a_k M.$$

Indeed, integrating the Hessian along the segment from $x^\star$ to x gives a symmetric average Hessian with spectrum in $[\mu, L]$, proving the contraction bound. Differentiating Ψ_k gives the stated Jacobian-Lipschitz constant. If $\nu_k = 0$, any larger constant in $(0, 1)$ can be used in the preceding corollary. As required by Assumption 4.5, the retained points and base candidates lie in $\mathcal{B}$.

Let $e_j = g_j - \nabla f_{i_j}(x_{s_j})$ be the realized local gradient noise. Assume ∇f_i is L_c-Lipschitz on the relevant segments, $\|\nabla f_i(x) - \nabla f(x)\|_2 \le \zeta_\infty$ on $\mathcal{B}$, and $\|\nabla f(x)\|_2 \le G$ there. Since $b_j = x_j - a_j g_j$, direct subtraction yields

$$\|\varepsilon_{j,k}\|_2 \le a_j\big(\|e_j\|_2 + \zeta_\infty + L_c\|x_{s_j} - x_j\|_2\big) + |a_j - a_k|G.$$

Thus, if stale displacements are at most $\tau_{\max}S$, where S bounds each intervening server step, then

$$\bar{\varepsilon}_k \le a_{\max}\big(e_{\max} + \zeta_\infty + L_c\tau_{\max}S\big) + G\max_j |a_j - a_k|.$$

All maxima here are over the retained window; stale snapshots and connecting segments are assumed to remain in the stated region. The pointwise heterogeneity bound can use $\zeta_\infty = \sqrt{N}\,\zeta$ under Assumption 4.3. For a fixed window of at most m arrivals, the variance bound in Assumption 4.2 and a union bound give $e_{\max} \le \sigma\sqrt{m/\delta}$ with probability at least $1-\delta$, before conditioning on the acceptance decision.

Substitution into U_k makes the local error bound depend explicitly on curvature, gradient noise, client heterogeneity, delay, and changes in effective step size. For quadratic f, $M = 0$. The resulting U_k enters the conditional comparison in Corollary 4.2.

B.5 Residual history in difference coordinates

Consider the affine fit (4) on a retained window containing the current pair (x_k, r_k). Enumerate the other $p = |\mathcal{H}_k| - 1$ pairs as (x_{j_ℓ}, r_{j_ℓ}), and define

$$D_x = [\,x_{j_\ell} - x_k\,]_{\ell=1}^{p}, \qquad D_r = [\,r_{j_\ell} - r_k\,]_{\ell=1}^{p}.$$

For $\gamma \in \mathbb{R}^p$, set $\alpha_{j_\ell} = \gamma_\ell$ and $\alpha_k = 1 - \mathbf{1}^\top\gamma$. This parameterizes the affine constraint exactly and gives

$$R_k\alpha = r_k + D_r\gamma, \qquad \widetilde{y}_k = b_k + (D_x + D_r)\gamma. \tag{23}$$

Consequently, the coefficient problem in these coordinates is

$$\min_\gamma \ \|r_k + D_r\gamma\|^2 + \lambda\left(\|\gamma\|^2 + (1 - \mathbf{1}^\top\gamma)^2\right), \tag{24}$$

with normal equations

$$\left[D_r^\top D_r + \lambda(I + \mathbf{1}\mathbf{1}^\top)\right]\gamma = -D_r^\top r_k + \lambda\mathbf{1}.$$

The fitted candidate is therefore the base candidate plus a correction formed from differences of complete candidate–residual pairs. Substitution into (2) yields, before any candidate processing,

$$e_k = \mathbf{1}_{\mathrm{acc},k}(1 - \rho_k)(D_x + D_r)\gamma_k.$$

For a fixed retained window, the affine and difference expressions are exact coordinate representations of the same fit, including the transformed ridge penalty. Block formation, smoothing, and age filtering determine which records are supplied to a history fit; these operations belong to the history policy. Their settings are specified separately from the common committed-update rule.

C Experimental details and additional results

C.1 MNIST experiments

The MNIST experiments use a small convolutional network, 12,000 training examples, 1,000 test examples, batch size 600, 60 epochs, and learning rate 0.01. The network has two 3×3 convolutional layers with 32 and 64 channels, a 2×2 max-pooling layer, a fully connected layer with 128 hidden units, and a 10-way output layer. ReLU activations follow the convolutional and hidden fully connected layers; the output uses log-softmax.

The base setting uses ten clients, an IID split, and delay parameters $(g, j) = (0.5, 0.1)$ as defined below. Label-sorted shards and a Dirichlet split with concentration 0.05 provide two non-IID settings. The high-delay setting uses $(g, j) = (1, 0.2)$, and the high-concurrency setting increases the number of clients to twenty. The first-order baseline in Table 1 uses a staleness-weighted gradient step. In the

four variants of the base setting, this step is implemented as the equivalent FedAsync update with unit scaling.

Table 3 reports the full MNIST comparison under the same five settings. HERA achieves the highest mean final accuracy and the lowest mean final loss in each setting.

Table 3: MNIST comparison across five asynchronous settings. Entries are means $\pm$ population standard deviations over ten runs; final accuracy is in percent.

Setting	Method	Final Acc. $\uparrow$	Final Loss $\downarrow$
Base IID	HERA	$\mathbf{98.20 \pm 0.42}$	$\mathbf{0.0675 \pm 0.0110}$
Base IID	FedAsync	89.27 ± 1.27	0.3712 ± 0.0335
Base IID	FedBuff	91.25 ± 1.54	0.2904 ± 0.0371
Base IID	FedAC	83.95 ± 13.73	0.7481 ± 0.9845
Label-Sorted Non-IID	HERA	$\mathbf{93.45 \pm 4.10}$	$\mathbf{0.2609 \pm 0.2513}$
Label-Sorted Non-IID	FedAsync	74.16 ± 9.25	0.7834 ± 0.2217
Label-Sorted Non-IID	FedBuff	88.25 ± 1.43	0.3829 ± 0.0492
Label-Sorted Non-IID	FedAC	80.49 ± 8.25	0.6446 ± 0.3097
Dirichlet Non-IID	HERA	$\mathbf{96.79 \pm 1.21}$	$\mathbf{0.1167 \pm 0.0482}$
Dirichlet Non-IID	FedAsync	79.72 ± 9.09	0.6261 ± 0.2730
Dirichlet Non-IID	FedBuff	88.97 ± 3.13	0.3500 ± 0.0786
Dirichlet Non-IID	FedAC	73.60 ± 15.33	1.4917 ± 1.2444
High Delay IID	HERA	$\mathbf{98.17 \pm 0.41}$	$\mathbf{0.0831 \pm 0.0271}$
High Delay IID	FedAsync	89.64 ± 2.49	0.3495 ± 0.0509
High Delay IID	FedBuff	92.09 ± 0.89	0.2749 ± 0.0242
High Delay IID	FedAC	77.30 ± 12.29	0.8158 ± 0.5376
IID 20 Workers	HERA	$\mathbf{97.61 \pm 0.72}$	$\mathbf{0.0750 \pm 0.0204}$
IID 20 Workers	FedAsync	81.47 ± 2.36	0.9989 ± 0.2217
IID 20 Workers	FedBuff	87.19 ± 2.89	0.4850 ± 0.2237
IID 20 Workers	FedAC	83.78 ± 17.91	0.6621 ± 1.0153

C.2 CIFAR EXPERIMENTS

Data and architecture. For each CIFAR dataset, we divide the 50,000 training images into 45,000 distributed training images and 5,000 validation images. Within each scenario, all methods share the split, client partition, and random-seed set. Random-number generators are seeded before model construction. All methods use a CIFAR-style ResNet18 with a 3×3 stem, four residual stages of widths $64, 128, 256, 512$, GroupNorm with eight groups, global average pooling, and a dataset-specific classifier. Table 4 specifies the client counts, partitions, and latency parameters.

Table 4: CIFAR scenarios. Dir. denotes a Dirichlet partition with concentration 0.05, and (g, j) are the latency parameters.

Scenario	Dataset	Clients	Partition	(g, j)
M1	CIFAR-10	10	IID	$(0, 0)$
M2	CIFAR-10	20	IID	$(0, 0)$
M3	CIFAR-10	10	Dir.	$(0, 0)$
M4	CIFAR-10	20	Dir.	$(0, 0)$
M5	CIFAR-10	20	Dir.	$(5, 0.1)$
M6	CIFAR-100	10	IID	$(0.1, 0.02)$
M7	CIFAR-100	20	IID	$(0.1, 0.02)$
M8	CIFAR-100	10	Dir.	$(0.1, 0.02)$
M9	CIFAR-100	20	Dir.	$(0.1, 0.02)$
M10	CIFAR-100	20	Dir.	$(5, 0.1)$

Asynchronous execution. The event queue assigns client i latency $d_i = 1 + (i \bmod 3)g + \epsilon$, with jitter $\epsilon \sim U(0, j)$ generated for each client dispatch. Each client has at most one update in flight and

trains from its dispatched server snapshot. On the earliest completion, the server processes the returned update and immediately redispatches that client. Each method's buffering rule determines when aggregation updates the server model. Equal latencies in M1–M4 still produce stale updates because other arrivals can change the server model while a client is training. Proposition A.4 characterizes the long-run completion fractions under independent dispatch jitter. They are generally nonuniform, and neither these fractions nor uniform marginal counts establish the conditional uniformity required by Theorem 4.1. The conditional law in Theorem 4.2 allows dependent completion identities, including probabilities equal to zero; no inverse-propensity weighting is used in these experiments.

Training spans 200 nominal epochs on CIFAR-10 and 300 on CIFAR-100. One nominal epoch counts a number of completed arrivals equal to the sum of the client data-loader lengths. The corresponding number of processed examples and simulated duration depend on client shard sizes and aggregation-release rules.

Training configurations. Each client dispatch computes one mini-batch gradient at its dispatched server snapshot and returns a single SGD update. All methods use batch size 64, weight decay 5×10^{-4}, and zero client-side momentum. FEDAC includes its control-variate correction in this step. Learning rates are selected in preliminary runs and kept fixed across three evaluation runs with new random seeds.

Table 5 records the scenario-specific HERA learning-rate and staleness-weight settings.

Table 5: HERA scenario-specific parameters. η_l is the client learning-rate parameter, and q_{up} is the exponent in $(1+\tau)^{-q_{\mathrm{up}}}$.

Scenario	η_l	q_{up}
M1	.05	0
M2	.05	.5
M3	.05	0
M4	.05	.5
M5	.05	0
M6	.20	0
M7	.10	0
M8	.20	0
M9	.15	0
M10	.10	0

Server update. The base displacement is calibrated using stored client updates. The history fit retains up to four displacement-difference columns and uses exponential smoothing with coefficient 0.9. Fitting starts with one eligible column; records older than 40 history refreshes are discarded, and rejected proposals leave the history unchanged. Let t index history refreshes and k_t the associated server update.

At refresh t, let $\bar{x}_t$ be the reference model, $\bar{r}_t$ the associated base displacement, and h_t the fitted history correction. The base candidate is $\bar{b}_t = \bar{x}_t + \bar{r}_t$. With history-correction coefficient 0.3 and base-residual coefficient 1, candidate construction and norm clipping are

$$\widetilde{d}_t = \bar{r}_t + 0.3 h_t, \qquad d_t = \min\left\{1, \frac{\|\bar{r}_t\|}{\|\widetilde{d}_t\|}\right\} \widetilde{d}_t, \qquad y_t = \bar{x}_t + d_t, \tag{25}$$

where $d_t = 0$ if $\widetilde{d}_t = 0$. The numerical checks require finite fitting quantities and a finite candidate; the directional check requires $\langle d_t, \bar{r}_t \rangle > 0$. Writing I_t for this acceptance indicator, the committed candidate is

$$z_t = \bar{b}_t + I_t \, 0.2 \, (y_t - \bar{b}_t). \tag{26}$$

Thus the nominal base weight is $\rho^0_{k_t} = 0.8$. Under the unit norm cap and positive-alignment test, (7) gives

$$\|z_t - \bar{b}_t\| \le \frac{\sqrt{2}}{5} \|\bar{r}_t\|,$$

corresponding to the constant anchor envelope $\Gamma_{k_t} = \sqrt{2}/5$ in (2).

For this fixed envelope, Corollary A.1 gives a finite-budget bound retaining the anchor term ΓV and the average base-update drift. Its application to the calibrated base update is conditional on the residual estimates (10)–(11) and a positive margin $a_b - \sqrt{c_b}\Gamma - C_\tau^{\mathrm{res}}\eta$. These requirements involve the cache mismatch, conditional second moment, and both lower and upper bounds on the RMS denominator. The stepsize-scaled anchor in Corollary 4.1 defines the separate regime with a vanishing stationarity bound. Appendix E summarizes the scope of these guarantees for the experimental configuration.

The server RMS state uses $\beta = 0.9$ and $\varepsilon_{\mathrm{rms}} = 10^{-8}$. The denominator floor is $q_{\min} = 0.1$ on CIFAR-10 and 0.2 on CIFAR-100. These server-side history, mixing, and RMS settings are shared across M1–M10.

Baselines. We compare with Delay-SGD, FEDASYNC, FedBuff, FADAS, CA2FL, and FEDAC. FedBuff weights each uploaded update by $(1+\tau)^{-1/2}$ before its five-arrival average. FADAS and CA2FL also use buffers of five arrivals, while FEDAC uses a buffer size equal to the number of clients. Tables 6 and 7 report the learning rates for all seven methods.

Table 6: CIFAR-10 learning rates by scenario. Entries give the client rate; FADAS and FEDAC use client/server pairs. The server rate is 1 for HERA, FedBuff, and CA2FL.

Scenario	HERA	Delay-SGD	FEDASYNC	FedBuff	FADAS	CA2FL	FEDAC
M1	.05	.10	.30	.05	$.01/3\times10^{-4}$	.05	$.10/1\times10^{-4}$
M2	.05	.05	.20	.20	$.01/5\times10^{-5}$	.05	$.10/1\times10^{-4}$
M3	.05	.10	.30	.05	$.01/3\times10^{-4}$	.05	$.10/1\times10^{-4}$
M4	.05	.05	.40	.20	$.01/1\times10^{-4}$	.05	$.10/1\times10^{-4}$
M5	.05	.025	.10	.10	$.01/5\times10^{-5}$	.05	$.10/1\times10^{-4}$

Table 7: CIFAR-100 learning rates by scenario. Entries give the client rate; FADAS and FEDAC use client/server pairs. The server rate is 1 for HERA, FedBuff, and CA2FL.

Scenario	HERA	Delay-SGD	FEDASYNC	FedBuff	FADAS	CA2FL	FEDAC
M6	.20	.05	.30	.30	$.01/1\times10^{-4}$	.20	$.10/1\times10^{-4}$
M7	.10	.05	.30	.20	$.01/1\times10^{-4}$	.20	$.10/3\times10^{-4}$
M8	.20	.05	.30	.30	$.01/1\times10^{-4}$	.20	$.10/1\times10^{-4}$
M9	.15	.05	.30	.20	$.01/1\times10^{-4}$	.20	$.10/3\times10^{-4}$
M10	.10	.05	.30	.20	$.01/1\times10^{-4}$	.20	$.10/3\times10^{-4}$

Evaluation. Table 8 reports final validation accuracy as the mean and sample standard deviation over three runs. Within each scenario, the strongest baseline is the method with the highest mean accuracy among the six baselines. Figure 2 shows the learning curves for M5 and M10.

Table 8: Final validation accuracy (%, mean ± sample standard deviation over three runs) after 200 nominal epochs on CIFAR-10 and 300 on CIFAR-100. Bold indicates the highest mean in each scenario.

Scenario	HERA	Delay-SGD	FEDASYNC	FedBuff
CIFAR-10				
M1	$\mathbf{89.660 \pm 0.610}$	86.960 ± 1.628	88.160 ± 0.790	87.413 ± 0.232
M2	83.253 ± 0.768	78.380 ± 1.306	$\mathbf{83.893 \pm 1.123}$	82.073 ± 1.593
M3	$\mathbf{87.673 \pm 1.621}$	84.200 ± 1.126	85.540 ± 1.906	85.687 ± 0.718
M4	$\mathbf{78.860 \pm 2.478}$	71.240 ± 2.873	75.860 ± 2.541	77.800 ± 3.632
M5	$\mathbf{75.047 \pm 0.503}$	54.847 ± 6.514	47.113 ± 7.843	70.607 ± 4.851
CIFAR-100				
M6	$\mathbf{70.580 \pm 0.832}$	63.900 ± 0.413	66.413 ± 0.061	70.313 ± 0.401
M7	64.080 ± 0.956	50.993 ± 0.594	57.313 ± 0.186	63.207 ± 1.468
M8	$\mathbf{70.567 \pm 0.446}$	62.247 ± 0.422	64.267 ± 0.647	70.027 ± 0.234
M9	$\mathbf{64.287 \pm 1.230}$	48.527 ± 1.465	52.993 ± 1.026	63.767 ± 0.280
M10	$\mathbf{62.473 \pm 1.118}$	41.647 ± 3.103	36.360 ± 0.330	59.353 ± 1.318

Scenario	FADAS	CA2FL	FEDAC
CIFAR-10			
M1	85.440 ± 0.436	87.260 ± 0.492	85.273 ± 0.880
M2	81.920 ± 0.720	80.027 ± 1.268	79.833 ± 1.925
M3	83.307 ± 1.706	85.727 ± 1.389	82.947 ± 1.102
M4	75.333 ± 3.078	75.080 ± 1.590	72.733 ± 2.027
M5	70.473 ± 4.908	72.813 ± 1.580	70.167 ± 5.527
CIFAR-100			
M6	64.560 ± 1.232	69.973 ± 0.340	58.940 ± 1.109
M7	58.060 ± 0.035	$\mathbf{65.120 \pm 1.265}$	44.880 ± 1.683
M8	64.433 ± 0.326	70.087 ± 0.574	58.520 ± 1.448
M9	57.667 ± 0.567	63.287 ± 0.938	45.420 ± 0.832
M10	58.127 ± 2.111	59.353 ± 4.360	40.367 ± 2.220

D EXTENDED RELATED WORK

Client heterogeneity and server optimization. Several federated methods modify the local objective or update direction to address statistical heterogeneity. FedProx penalizes departure from the dispatched model through a proximal term (Li et al., 2020). SCAFFOLD uses control variates to correct client drift (Karimireddy et al., 2020), while FedDyn introduces a dynamic local regularizer to align local and global solutions (Acar et al., 2021). Adaptive federated optimization applies server-side variants of Adagrad, Adam, and Yogi to aggregated client updates (Reddi et al., 2021). These methods provide context for the choice of base updater and scaling. HERA focuses on the subsequent use of residual history and the size of its correction relative to that base update.

Scheduling and resource-aware participation. The workload assigned to clients affects both update delay and participation. FedCompass adjusts local work using a computing-power-aware scheduler so that groups of clients finish at similar times (Li et al., 2024). TimelyFL adapts partial training to the resources available at each client (Zhang et al., 2023), whereas AsyncDrop distributes submodels to heterogeneous devices (Dun et al., 2023). Pisces jointly guides participant selection and aggregation timing (Jiang et al., 2022). These approaches act on how updates are produced and collected; the residual-history correction studied here acts on the resulting server trajectory.

Online updates and staleness control. Asynchronous learning has also been developed for online edge data in ASO-Fed (Chen et al., 2020) and for data disparity and temporal unavailability in predictive maintenance through DAAFL (von Wahl et al., 2024). PORT examines the degree of asynchrony in federated training (Su & Li, 2022). FedASMU adapts model updates to staleness (Liu et al., 2024b), and explicit staleness-control mechanisms study the tradeoff between accuracy and efficiency (Sun et al., 2024). These settings emphasize how data availability and update timing jointly shape the information received by the server.

Communication structure and arrival dynamics. AD-PSGD and stochastic gradient push study distributed optimization with decentralized communication (Lian et al., 2018; Assran et al., 2019). Graph-based asynchronous SGD provides a unified framework for decentralized and federated updates (Even et al., 2024). At a central server, queueing models make client service rates and sampling choices explicit (Leconte et al., 2024), while AsGrad analyzes asynchronous SGD under heterogeneous data and computation or communication speeds (Islamov et al., 2024). These perspectives motivate distinguishing the delay of an update from the distribution of the client that produced it. Our analysis represents these effects through separate delay and conditional-selection terms.

Other uses of optimization history. History can enter an optimizer through mechanisms beyond an Anderson residual fit. MasFL uses momentum to combine historical descent directions, and AdaMasFL adds local gradient normalization (Yan et al., 2025). Energy-based gradient methods have also been combined with Anderson acceleration (Liu et al., 2024a). In HERA, the mixing coefficients are selected by a regularized residual fit, and the resulting proposal is checked against the current base displacement. This separates the construction of the history proposal from control of the committed server step.

E LIMITATIONS

The asymptotic-to-zero result assumes local-SGD residuals, stepsize-scaled anchoring, bounded RMS denominators and delay, conditional uniform client sampling, and finite variance. The nonuniform-arrival theorem relaxes conditional uniformity but retains the conditional gradient assumptions and incurs explicit selection-bias and second-moment terms. Its discrepancy can be conservative under a queue that determines the next identity; long-run completion frequencies cannot replace conditional probabilities without further analysis. The fixed CIFAR anchor has a residual-level fixed-budget interpretation, not an asymptotic-to-zero guarantee. The conditional residual moments and positive descent margin have not been verified for the calibrated CIFAR base updater. Local improvement depends on residual-fit quality and evaluation error within a contractive neighborhood. Experiments use simulated asynchronous execution on MNIST and CIFAR and compare accuracy at nominal-epoch budgets. Cross-device evaluation, matched-time comparisons, and measurements of the server history overhead would characterize the computational and communication costs more directly.

F BROADER IMPACTS

HERA is an application-agnostic optimization method for asynchronous federated learning. Potential benefits include better use of heterogeneous compute resources and reduced idle waiting. Its societal effects depend on the downstream models, datasets, and applications. Lower training costs could support both beneficial and harmful uses, making privacy protection, system security, and responsible deployment important considerations for downstream users.