FEDHV: LOW-OVERHEAD HYPERVOLUME WEIGHTING FOR FEDERATED MULTI-OBJECTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Task-wise federated multi-objective optimization (FedMOO) trains a shared model for competing prediction objectives under heterogeneous data, partial participation, and communication constraints. Existing methods commonly derive task weights from gradient or update geometry. This requires task-specific information or iterative server-side optimization. We introduce **FedHV**, which maps reference-relative objective slacks to closed-form inverse-slack weights. Each client optimizes one weighted loss and returns objective estimates with its model update. The protocol adds exactly $2m$ auxiliary scalars per participating client, yielding $\Theta(d + m)$ total per-client communication, compared with the $\Theta(md)$ task-specific communication of FSMGDA, and requires no additional synchronization stage. We analyze the resulting one-round-delayed weights under client heterogeneity, multi-step local updates, partial participation, and finite-sample objective reports. Under a fixed-horizon positive-slack reference condition, with the prescribed horizon-dependent step size and vanishing report error, FedHV achieves an $\mathcal{O}(T^{-1/2})$ rate for the average squared log-hypervolume gradient norm; persistent report error determines the resulting stationarity neighborhood. The same bound controls the squared Pareto-stationarity residual. Across six Dirichlet-partitioned non-IID settings from four vision benchmark families and three training seeds, FedHV exceeds FSMGDA and FedCMOO in mean accuracy in five settings. Among these methods and uniform scalarization, it achieves the highest worst-task accuracy in four settings and improves the difficult CIFAR-10 objective in both CIFAR10–MNIST settings.

1 INTRODUCTION

Federated learning (FL) trains models from decentralized data without sharing clients' raw examples. In many applications, however, one shared model must serve several prediction tasks whose objectives may conflict. Optimizing their average can favor easier tasks and leave a difficult objective with poor performance. *Task-wise federated multi-objective optimization* (FedMOO) instead seeks a shared model that balances multiple global task objectives (Yang et al., 2023; Askin et al., 2025). This setting differs from client-wise robust or fair FL, where the objectives represent clients or demographic groups rather than prediction tasks (Mohri et al., 2019; Li et al., 2020; Hu et al., 2022).

Adapting task trade-offs is difficult in FL because the server observes information from only a subset of heterogeneous clients, while multiple local updates introduce client drift. FSMGDA constructs a federated multi-gradient direction from objective-specific client updates, leading to $\Theta(md)$ per-client communication for m tasks and a d-dimensional model (Yang et al., 2023). FedCMOO reduces this communication through randomized low-rank estimation, but it still estimates task-update geometry and solves an iterative projected weight problem at the server (Askin et al., 2025). Objective values provide a lower-dimensional alternative: clients can report m scalar losses, allowing the server to adapt task weights without constructing or communicating task-specific d-dimensional updates.

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.

We propose **FedHV** (*Federated Hypervolume*), which forms task weights directly from these objective reports. Given global task losses $f_i(\theta)$ and fixed reference losses r_i, FedHV uses

$$w_i(\theta) \propto \big(r_i - f_i(\theta)\big)^{-1},$$

so that tasks closer to their references receive greater weight. This inverse-slack rule follows from centralized single-solution hypervolume maximization (Miranda & Von Zuben, 2016). FedHV obtains the reference vector from a short uniform-weight calibration phase, adds a positive margin, and freezes it before adaptive training.

At each round, selected clients evaluate their task losses at the common pre-update model and optimize one scalar loss using weights computed from the preceding round's reports. The new reports accompany the model updates and determine the next-round weights. FedHV therefore adds exactly $2m$ auxiliary scalars per participating client without an intermediate synchronization stage. The one-round delay is also central to the analysis: the implemented weights depend on noisy reports at the preceding global model rather than the current objectives. Our proof controls this discrepancy through the report error and the inter-round objective change, together with stochastic-gradient noise, partial-participation heterogeneity, and multi-step local drift.

Contributions.

- **Protocol.** We introduce a task-wise FedMOO method that computes closed-form inverse-slack weights from low-dimensional objective reports. FedHV uses one weighted local trajectory, adds $2m$ auxiliary scalars per participating client, and requires neither task-update geometry nor an additional synchronization stage.
- **Analysis.** We establish convergence for the one-round-delayed protocol under partial participation, client heterogeneity, stochastic multi-step local updates, and finite-sample objective reports. With vanishing report error and the prescribed step-size schedule, the average squared log-hypervolume gradient norm decreases as $\mathcal{O}(T^{-1/2})$. We also relate this quantity to the Pareto-stationarity residual and characterize reference-relative worst-task slack control.
- **Evaluation.** Across six Dirichlet-partitioned non-IID settings, FedHV exceeds FSMGDA and FedCMOO in mean accuracy in five settings and achieves the highest worst-task accuracy among all four evaluated methods in four. Ablations examine the effects of objective-report budget, client participation, and local computation.

2 Related Work

Pareto-oriented multi-objective optimization. Multi-objective optimization seeks solutions for which no objective can be improved without degrading another. MGDA computes a minimum-norm convex combination of task gradients and uses the resulting direction to approach Pareto-stationary points (Désidéri, 2012; Sener & Koltun, 2018). Pareto MTL extends this perspective by targeting several regions of the Pareto front rather than producing only one compromise solution (Lin et al., 2019). These methods explicitly define their updates through task-gradient geometry and Pareto-stationarity conditions.

Gradient balancing in multi-task learning. A related class of multi-task learning methods modifies task gradients to improve shared-model training without necessarily constructing an MGDA direction. GradNorm balances gradient magnitudes and relative training rates (Chen et al., 2018); PCGrad removes components of mutually conflicting gradients (Yu et al., 2020); and CAGrad balances average descent against gradient conflict (Liu et al., 2021). Although these methods address negative transfer among tasks, their update rules differ from methods explicitly designed around the Pareto-stationarity residual. They also require separate task-gradient information and were developed primarily for centralized optimization (Hu et al., 2023).

Federated multi-objective optimization. FMOL introduces FMGDA and FSMGDA for nonconvex federated multi-objective problems (Yang et al., 2023). FSMGDA supports multiple local steps but constructs its federated direction from objective-specific client updates, resulting in communication that scales with both the number of tasks and the model dimension. FedCMOO reduces this cost

using randomized low-rank estimation of task-update geometry and then optimizes the task weights through projected server-side iterations (Askin et al., 2025). FedHV does not estimate update geometry. It forms weights from an m-dimensional objective report and uses a single weighted local trajectory, replacing task-specific model information with objective values while retaining the usual model-update exchange.

Hypervolume-based optimization. Hypervolume is a reference-dependent and Pareto-compliant measure of joint objective quality (Zitzler et al., 2007; Auger et al., 2009; Guerreiro et al., 2021). It has been optimized for a single neural-network solution and for learned Pareto sets or parameterized solution families (Miranda & Von Zuben, 2016; Deist et al., 2023; Zhang et al., 2023). For a single solution, Miranda & Von Zuben (2016) derive the log-hypervolume objective and its inverse-slack weighting rule. FedHV uses this centralized identity as its scalarization rule; its contribution lies in the federated protocol and analysis. In particular, the weights are estimated from stochastic finite-sample client reports, held fixed during multiple local steps, and applied one round after the corresponding objectives are observed. Our analysis accounts jointly for this delay, partial participation, client heterogeneity, and local-update drift.

3 FEDHV

3.1 TASK-WISE FEDERATED MULTI-OBJECTIVE OPTIMIZATION

We consider N equal-size clients and m prediction objectives. Client k has local risk $f_i^{(k)}(\theta)$ for task i, and the corresponding global risk is $f_i(\theta) = N^{-1}\sum_{k=1}^{N} f_i^{(k)}(\theta)$. Task-wise FedMOO seeks a shared model $\theta \in \mathbb{R}^d$ that balances

$$\boldsymbol{F}(\theta) = \big(f_1(\theta), \dots, f_m(\theta)\big)^\top. \tag{1}$$

Let $\Delta_m := \{\lambda \in \mathbb{R}^m_{\geq 0} : \sum_i \lambda_i = 1\}$. Under the standard definition (Désidéri, 2012; Sener & Koltun, 2018), $\theta^\star$ is *Pareto-stationary* if some $\lambda \in \Delta_m$ satisfies

$$\sum_{i=1}^{m} \lambda_i \nabla f_i(\theta^\star) = \mathbf{0}. \tag{2}$$

3.2 LOG-HYPERVOLUME WEIGHTING

Let $\boldsymbol{r} = (r_1, \dots, r_m)^\top$ be a fixed vector of reference losses and define the reference-relative slack $s_i(\theta) := r_i - f_i(\theta)$. On the domain where all slacks are positive, FedHV minimizes the negative log-hypervolume

$$\Phi(\theta) = -\sum_{i=1}^{m} \log\big(r_i - f_i(\theta)\big). \tag{3}$$

The ideal weights computed from the true global objectives are

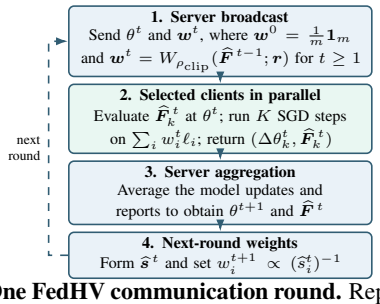
$$w_i^\star(\theta) = \frac{\big(r_i - f_i(\theta)\big)^{-1}}{\sum_{j=1}^{m}\big(r_j - f_j(\theta)\big)^{-1}}, \qquad i = 1, \dots, m. \tag{4}$$

Writing $c(\theta) := \sum_j (r_j - f_j(\theta))^{-1}$ gives

$$\nabla\Phi(\theta) = c(\theta)\sum_{i=1}^{m} w_i^\star(\theta)\nabla f_i(\theta). \tag{5}$$

Tasks closer to their references therefore receive greater weight, and normalization changes only the positive scale of the resulting gradient. The inverse-slack identity is due to Miranda & Von Zuben (2016); FedHV uses it within a federated protocol.

The reference is constructed before adaptive training using uniform-weight calibration. Each two-task benchmark family uses one 20-round calibration run. For CIFAR10–MNIST, the reference obtained under the $\alpha = 0.30$ partition is reused for both reported settings. The CELEBA5 reference



Per-round method comparison

Property	FSMGDA	FedCMOO	FedHV
Task-specific geometry	✓	✓	**No**
Server-side weight optimization	MinNorm	Iter. proj.	**Closed form**
Single local scalar loss	No	✓	✓
Per-client communication	$\Theta(md)$	$\Theta(d)$	$\boldsymbol{\Theta(d+m)}$

Figure 1: **One FedHV communication round.** Reports from round $t-1$ determine the weights for round t and accompany the ordinary model exchange. Right: dominant per-round operations and per-client communication; Appendix D gives exact counts.

is constructed from three independent 20-round runs, corresponding to 60 sequential-round equivalents, and is shared across both CELEBA5 settings and all adaptive-training seeds. Adaptive models are reinitialized rather than continued from the calibration runs. Calibration is therefore a one-time preprocessing cost and does not add synchronization to the adaptive rounds. Appendix B.8 gives the construction, margins, and resulting references.

3.3 FEDERATED PROTOCOL

At round t, the server holds the global model θ^t and weight vector $\boldsymbol{w}^t$. The initial weights are $\boldsymbol{w}^0 = \frac{1}{m}\mathbf{1}_m$; subsequent weights are computed from the preceding aggregate objective report. The server samples $S_t \subseteq \{1, \ldots, N\}$ and broadcasts $(\theta^t, \boldsymbol{w}^t)$.

Before local optimization, each selected client $k \in S_t$ estimates $\widehat{\boldsymbol{F}}_k^t$ at the common pre-update model θ^t. The report may use B_{rep} minibatches or the client's complete local dataset. Client k then performs K local SGD steps on the scalarized loss

$$L_k^t(\theta; B) = \sum_{i=1}^{m} w_i^t \, \ell_i(\theta; B). \tag{6}$$

The server computes $\boldsymbol{w}^t$ once before local training, and each client holds it fixed throughout all K steps. Consequently, clients neither recompute the weights nor differentiate through the weight map during local SGD. This design permits one local training trajectory and avoids the task-specific gradient calculations that would be required to update the weights within the round.

Let $\Delta\theta_k^t := \theta_{k,K}^t - \theta^t$ denote the resulting model update. Client k returns $(\Delta\theta_k^t, \widehat{\boldsymbol{F}}_k^t)$ in one uplink, and the server averages both quantities:

$$\theta^{t+1} = \theta^t + \eta_g \frac{1}{|S_t|} \sum_{k \in S_t} \Delta\theta_k^t, \qquad \widehat{\boldsymbol{F}}^t = \frac{1}{|S_t|} \sum_{k \in S_t} \widehat{\boldsymbol{F}}_k^t. \tag{7}$$

The server forms the next weight vector by applying a numerical floor to the estimated slacks:

$$\begin{aligned} \widehat{s}_i^t &:= \max\left\{ r_i - \widehat{f}_i^t, \rho_{\text{clip}} \right\}, \qquad \rho_{\text{clip}} > 0, \\ w_i^{t+1} &:= \frac{(\widehat{s}_i^t)^{-1}}{\sum_{j=1}^m (\widehat{s}_j^t)^{-1}}. \end{aligned} \tag{8}$$

We denote this stabilized mapping by $W_{\rho_{\text{clip}}}(\widehat{\boldsymbol{F}}; \boldsymbol{r})$. The floor is a numerical failsafe for noisy or anomalous reports: it keeps the implemented weights finite but neither asserts nor enforces positivity of the true slacks $r_i - f_i(\theta)$. The convergence analysis instead requires the true objectives to remain inside a positive-slack domain. Reports collected in round t determine $\boldsymbol{w}^{t+1}$ rather than the weights used by the same client cohort. This one-round ordering fixes $\boldsymbol{w}^t$ before the current clients and minibatches are sampled, avoids an intermediate server–client exchange, and allows the weights to remain constant during local optimization. The weights accompany the model downlink and the reports accompany the model-update uplink, adding exactly $2m$ auxiliary scalars per participating client without an additional synchronization stage. Appendix E provides complete pseudocode.

4 THEORETICAL GUARANTEES

4.1 ASSUMPTIONS AND MAIN RESULT

We analyze the one-round-delayed protocol over a fixed horizon $T \geq 1$. Appendix A.2 defines the probability space and states the assumptions formally.

Standing assumptions. Clients have equal aggregation weights, and each round samples M of the N clients uniformly without replacement, with participation rate $p = M/N$. Local objectives are nonnegative and smooth. Minibatch gradients are conditionally unbiased with variance bounded by σ_i^2, and client gradient dissimilarity is bounded by ζ_i^2 along the analyzed trajectory. Multiple local steps introduce the corresponding client-drift term.

Reference-domain condition. The fixed reference must satisfy

$$r_i - f_i(\theta) \geq \rho > 0, \qquad i \in [m], \tag{9}$$

at every global, local, and auxiliary point used by the fixed-horizon proof. This condition keeps the log-hypervolume objective and its gradient well defined. The numerical floor $0 < \rho_{\mathrm{clip}} \leq \rho$ acts only on estimated slacks and does not establish Equation (9) for the true objectives.

Appendix A.5 gives a conditional high-probability certificate for preserving the safe domain. Its intuition is as follows. Suppose the initial model has a verified positive true-slack margin and the weighted stochastic local gradients are almost surely bounded. A round-excursion condition then limits the movement of every local, global, and auxiliary iterate, keeping an entire round inside a smaller positive-slack region whenever the round begins in the safe region. A horizon-dependent budget controls the cumulative change in Φ and implies that the global iterates remain safe with probability at least $1 - \delta$. These conditions explain how a safe initialization and bounded updates can preserve the reference margin; they are sufficient conditions and are not certified by the empirical calibration procedure alone.

Define the aggregate stochastic-gradient variance and the finite-population participation factor by

$$\Sigma^2 := \sum_{i=1}^{m} \sigma_i^2, \qquad \chi_N(M) := \frac{N - M}{M(N-1)} = \frac{1-p}{p(N-1)}. \tag{10}$$

The partial-participation heterogeneity term is

$$\mathcal{H}_N(p;\zeta) := \chi_N(M) \sum_{i=1}^{m} \zeta_i^2,$$

which vanishes under full participation.

Define the server-report error by

$$\Xi_{\mathrm{loss},T} := \max_{0 \leq t < T} \mathbb{E}\left[\left\| \widehat{\boldsymbol{F}}^{\,t} - \boldsymbol{F}(\theta^t) \right\|_\infty^2 \right]. \tag{11}$$

For the finite-population reporting procedure used by FedHV, Appendix A.7 establishes

$$\Xi_{\mathrm{loss},T} \leq \sum_{i=1}^{m} \left[\chi_N(M) \omega_{i,T}^2 + \tau_{i,T}^2 \right], \tag{12}$$

where $\omega_{i,T}^2$ measures between-client variation in objective i, and $\tau_{i,T}^2$ measures finite-sample error in the local objective reports. This decomposition gives two direct ways to tighten the report term. Increasing the report budget reduces $\tau_{i,T}^2$, and evaluating each selected client's complete local dataset makes this component zero. Increasing M reduces the client-subsampling coefficient $\chi_N(M)$, which becomes zero under full participation. Complete local evaluation therefore removes within-client reporting noise but does not remove the error caused by sampling only a subset of heterogeneous clients.

For $t \geq 1$, FedHV uses

$$\boldsymbol{w}^t = W_{\rho_{\mathrm{clip}}}\left(\widehat{\boldsymbol{F}}^{\,t-1}; \boldsymbol{r}\right),$$

whereas the ideal current weight is

$$\boldsymbol{w}^{\star,t} = W_{\rho_{\mathrm{clip}}}\big(\boldsymbol{F}(\theta^t); \boldsymbol{r}\big).$$

The Lipschitz property of the stabilized inverse-slack map, proved in Appendix A.6, gives

$$\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\|_1 \le \frac{2}{\rho_{\mathrm{clip}}}\Big(\|\widehat{\boldsymbol{F}}^{\,t-1} - \boldsymbol{F}(\theta^{t-1})\|_\infty + \|\boldsymbol{F}(\theta^{t-1}) - \boldsymbol{F}(\theta^t)\|_\infty\Big). \tag{13}$$

The first term is the preceding report error, while the second measures the inter-round change in the global objectives. This decomposition makes the effect of one-round staleness explicit rather than treating the implemented weights as if they were evaluated at θ^t.

Theorem 4.1 (FedHV stationarity)**.** *Under Assumptions A.1, A.2, A.3, A.4, A.9, and A.18, let* $\eta_{\mathrm{eff}} := \eta_g\eta_\ell K$*. There exist finite constants* $A_1, \dots, A_7$*, independent of* T*,* N*,* p*,* η_ℓ*,* η_g*, and* K*, such that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(\theta^t)\|^2 \le \frac{A_1\Delta_\Phi}{\eta_{\mathrm{eff}}T} + A_2\eta_{\mathrm{eff}}\Sigma^2 + A_3\eta_{\mathrm{eff}}\mathcal{H}_N(p;\zeta) + A_4\eta_\ell^2(K-1)(2K-1) + A_5\Xi_{\mathrm{loss},T} + A_6\eta_{\mathrm{eff}}^2\mathcal{V} + \frac{A_7}{T}, \tag{14}$$

where

$$\Delta_\Phi := \Phi(\theta^0) - \Phi_{\mathrm{lb}}, \qquad \Phi_{\mathrm{lb}} := -\sum_{i=1}^{m}\log r_i.$$

Here, Φ_{lb} *is a lower bound on* Φ*, not necessarily its global minimum, and* $\mathcal{V}$ *bounds the inter-round displacement that enters the stale-weight error. Appendix A.9 specifies* $\mathcal{V}$ *and the constants* $A_1, \dots, A_7$*.*

The bound separates the initial objective gap, stochastic-gradient variance, heterogeneity under partial participation, multi-step local drift, report error, and inter-round weight staleness. The A_7/T term accounts for the initial uniform weight vector. In particular, the nonvanishing neighborhood is not an unspecified error floor: its reporting component can be reduced through a larger report budget or greater client participation, subject to the corresponding communication and evaluation costs.

Corollary 4.2 (Rate and stationarity neighborhood)**.** *Under the conditions of Theorem 4.1, fix* $\eta_g > 0$ *and, for a target horizon* T*, choose*

$$\eta_\ell = \Theta\left(\frac{1}{\eta_g K\sqrt{T}}\right)$$

within the stability range of Assumption A.9. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(\theta^t)\|^2 \le \mathcal{O}(T^{-1/2}) + A_5\Xi_{\mathrm{loss},T}. \tag{15}$$

Thus, vanishing report error yields an $\mathcal{O}(T^{-1/2})$ *ergodic rate, whereas persistent report error determines a stationarity neighborhood. The schedule depends on the target horizon and differs from the fixed step sizes used in the experiments. Appendix A.9 gives the explicit rate and best-iterate statement.*

The proof first applies Equation (13) to separate report error from inter-round objective change. A local-SGD bias–variance–drift decomposition then controls the expected update and its second moment. Applying smoothness of Φ to the resulting one-round descent inequality and summing over t yields Theorem 4.1; Appendix A.12 gives the complete argument.

4.2 Relation to Pareto stationarity

Define the Pareto-stationarity residual

$$\mathcal{P}(\theta) := \min_{\lambda \in \Delta_m} \left\| \sum_{i=1}^{m} \lambda_i \nabla f_i(\theta) \right\|.$$

Proposition 4.3 (Quantitative Pareto relation)**.** *Let*

$$c(\theta) := \sum_{i=1}^{m} \frac{1}{r_i - f_i(\theta)}, \qquad c_{\min} := \sum_{i=1}^{m} \frac{1}{r_i}.$$

On the safe region,

$$\mathcal{P}(\theta) \le \frac{\|\nabla \Phi(\theta)\|}{c(\theta)} \le \frac{\|\nabla \Phi(\theta)\|}{c_{\min}}. \tag{16}$$

The normalized inverse-slack weights belong to Δ_m and are therefore a feasible choice in the definition of $\mathcal{P}(\theta)$. Nonnegative losses also imply $c(\theta) \ge c_{\min}$. Consequently, $\nabla \Phi(\theta) = \mathbf{0}$ implies Pareto stationarity, and Theorem 4.1 controls the corresponding squared Pareto residual. FedHV obtains this relation while each client differentiates only the scalar loss in Equation (6); the algorithm does not construct per-task gradient geometry, test pairwise gradient conflicts, or solve a minimum-norm problem. Appendix A.10 gives the exact and averaged forms.

4.3 Hypervolume-induced worst-task slack control

Proposition 4.4 (Minimum-slack control)**.** *Let $s_i(\theta) := r_i - f_i(\theta) > 0$, suppose $s_j(\theta) \le S_j$ for all $j \in [m]$, and define*

$$R_{\max}(\boldsymbol{S}) := \max_{i \in [m]} \prod_{j \ne i} S_j.$$

Then

$$\min_{i \in [m]} s_i(\theta) \ge \frac{e^{-\Phi(\theta)}}{R_{\max}(\boldsymbol{S})}. \tag{17}$$

For nonnegative losses, one may take $S_j = r_j$.

Hence, if $\Phi(\theta^t) \le B$, every task satisfies

$$f_i(\theta^t) \le r_i - \frac{e^{-B}}{R_{\max}(\boldsymbol{S})}.$$

This is a reference-relative loss-space guarantee and does not by itself imply a bound on worst-task test accuracy. Appendix A.4 gives the complete derivation.

5 Experiments

5.1 Experimental Setup

Benchmarks. We use six non-IID settings from the four benchmark families adopted by FedCMOO (Askin et al., 2025): MultiMNIST, MNIST–FMNIST, CIFAR10–MNIST, and CelebA5 (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009; Liu et al., 2015). The settings vary Dirichlet concentration, participation, and, for CelebA5, whether client allocation uses all attribute labels or only the first. Appendix B gives the complete constructions and partitions.

Training protocol and baselines. Within each setting, methods share the architecture, client partition, participation schedule, round budget, and number of local steps. We compare FedHV with FSMGDA (Yang et al., 2023), FedCMOO (Askin et al., 2025), and Uniform, which fixes $w_i = 1/m$ in the same single-trajectory update as FedHV. We report mean and sample standard deviation over seeds $\{0, 42, 2026\}$ on one fixed partition. Appendix B gives the architectures, optimization settings, and other configurations used by the implementations.

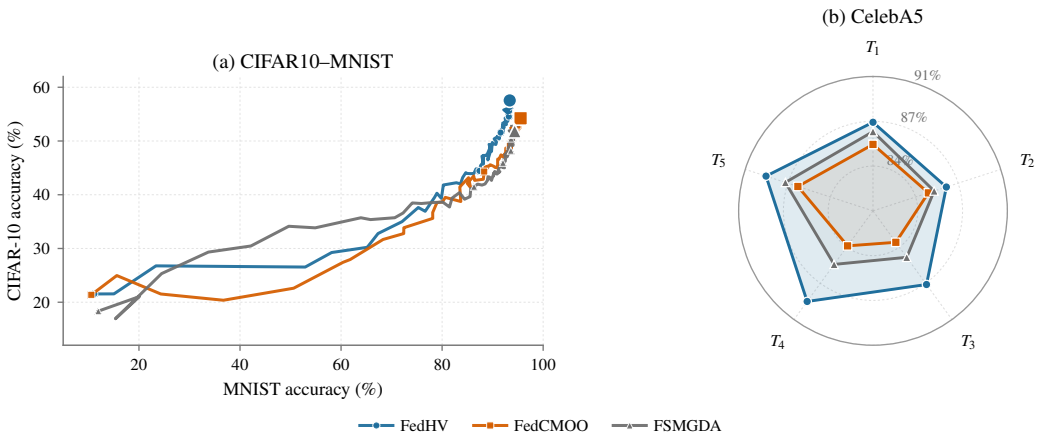


Figure 2: **Task-level accuracy, averaged over three runs.** **(a)** Evaluation trajectories on CIFAR10–MNIST-B ($\alpha = 0.30$); the terminal cross shows one sample standard deviation per task. **(b)** Mean accuracy for the five CelebA5-B objectives ($\alpha = 0.30$); the radial axis spans 80–91%. Appendix C.2 gives numerical values and uncertainty.

References and evaluation metrics. FedHV freezes a calibration-derived training reference $\boldsymbol{r}^{\text{train}}$; every method uses $\boldsymbol{r}^{\text{eval}} = \mathbf{31}_m$ for evaluation hypervolume. We report $\text{MeanAcc} = \frac{1}{m}\sum_i a_i$ and $\text{WorstAcc} = \min_i a_i$; Appendix C.1 reports mean loss and $\text{HV} = \prod_i (r_i^{\text{eval}} - \ell_i)$.

Across the reported standard runs, all retained server-aggregated objective reports have positive raw reference slack and the numerical slack floor is never activated; Appendix B.8 reports the minimum observed margins. This is a report-level diagnostic rather than a certificate of the true-slack condition in (9).

5.2 Comparison with Federated Multi-Objective Baselines

Table 1 shows that FedHV exceeds FSMGDA and FedCMOO in MeanAcc and WorstAcc in five of six settings; FedCMOO leads these methods on MNIST–FMNIST. Among all four methods, FedHV has the highest MeanAcc in three settings and WorstAcc in four, whereas Uniform has the highest MeanAcc on both CelebA5 settings and WorstAcc on CelebA5-B. The largest and most consistent separation occurs on the mixed-difficulty CIFAR10–MNIST benchmark. Using the reported means, FedHV raises WorstAcc from the next-best 51.96% to 53.03% in setting A and from 56.24% to 57.55% in setting B, gains of 1.07 and 1.31 percentage points. The corresponding MeanAcc margins are smaller (0.42 and 0.13 points), indicating that the main benefit is improved performance on the weaker task rather than a uniform increase across both objectives.

Figure 2 clarifies these trade-offs. On CIFAR10–MNIST-B, FedHV attains higher final CIFAR-10 accuracy than FSMGDA and FedCMOO but lower MNIST accuracy than FedCMOO; its aggregate gain therefore comes from the harder objective, not uniform improvement. On CelebA5-B, FedHV remains comparatively balanced across the five attribute groups. Complete per-task results, including the alternate partition, are in Appendix C.2.

Table 1: Final MeanAcc and WorstAcc (%), reported as mean ± sample standard deviation over three training runs on one fixed client partition. Bold denotes the best result and underlining denotes the second-best result among the four methods. Rankings are computed from the unrounded means.

	Uniform		FSMGDA		FedCMOO		FedHV	
Setting	Mean ↑	Worst ↑	Mean ↑	Worst ↑	Mean ↑	Worst ↑	Mean ↑	Worst ↑
MultiMNIST	90.17 ± 0.18	88.14 ± 0.34	88.78 ± 1.38	86.15 ± 2.53	89.89 ± 0.16	88.79 ± 0.57	**90.72 ± 0.65**	**88.94 ± 0.61**
MNIST–FMNIST	86.41 ± 0.51	76.98 ± 0.77	84.84 ± 0.87	75.58 ± 1.05	**86.69 ± 0.18**	**78.58 ± 0.66**	86.35 ± 0.45	77.03 ± 0.70
CelebA5-A	**88.08 ± 0.11**	86.30 ± 0.19	86.98 ± 0.06	85.77 ± 0.08	86.60 ± 0.31	85.00 ± 0.29	88.06 ± 0.21	**86.44 ± 0.16**
CelebA5-B	**87.94 ± 0.07**	**86.32 ± 0.05**	85.88 ± 0.07	84.70 ± 0.05	84.68 ± 1.94	82.11 ± 3.55	87.88 ± 0.05	86.32 ± 0.11
CIFAR10–MNIST-A	72.72 ± 0.11	51.33 ± 0.18	72.76 ± 0.22	51.96 ± 0.55	71.54 ± 0.48	47.98 ± 1.15	**73.18 ± 0.14**	**53.03 ± 0.35**
CIFAR10–MNIST-B	75.33 ± 0.20	56.24 ± 0.32	73.10 ± 0.66	51.85 ± 1.00	74.88 ± 0.51	54.23 ± 0.88	**75.46 ± 0.15**	**57.55 ± 0.20**

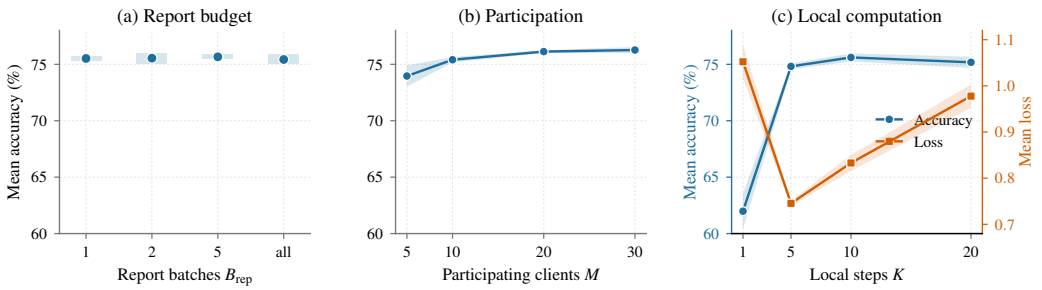


Figure 3: **FedHV ablations on CIFAR10–MNIST-B.** Mean ± std. dev. over three runs for **(a)** report budget B_{rep}, **(b)** participating clients M, and **(c)** local steps K. Panel (c) includes MeanLoss on the right axis.

FedHV obtains these results while each client trains on one scalar loss and communicates only an m-dimensional objective report in addition to its model update. Table 1 and Figure 2 show that objective-value weighting can match or exceed the geometry-based methods without task-specific model updates or a server-side weight solver.

Accuracy and loss-space metrics do not rank the methods identically. FedHV has the lowest MeanLoss and largest evaluation HV on MULTIMNIST; FedCMOO leads on MNIST–FMNIST; and Uniform leads on both CELEBA5 settings. On CIFAR10–MNIST, FSMGDA leads these metrics at $\alpha = 0.10$ and FedCMOO at $\alpha = 0.30$, although FedHV has the highest mean and worst-task accuracy in both settings. These results indicate a loss–accuracy trade-off rather than dominance across all criteria; Appendix C.1 gives the complete values. Beyond task accuracy, FedHV reduces the observed computational and communication overhead of multi-objective optimization. Empirical profiling on CIFAR10–MNIST (detailed in Appendix D) confirms that FedHV reduces end-to-end training time by 24.7% compared to FedCMOO and 32.9% compared to FSMGDA, while exhibiting lower observed across-seed runtime variability (± 5.69s vs. ± 33.31s). Furthermore, by replacing task-update geometry with low-dimensional scalar reports, FedHV eliminates asymmetric uplink bottlenecks, effectively cutting the required client uplink payload in half compared to geometry-based baselines. Crucially, these end-to-end efficiency gains fully account for FedHV's one-time preprocessing calibration cost.

5.3 ABLATION STUDIES

On CIFAR10–MNIST-B, we vary one factor from the default $(B_{\text{rep}}, M, K) = (\text{all}, 10, 10)$: $B_{\text{rep}} \in \{1, 2, 5, \text{all}\}$, $M \in \{5, 10, 20, 30\}$, or $K \in \{1, 5, 10, 20\}$.

Sparse reporting suffices: final mean accuracy varies by $\leq 0.23\%$ across report budgets, confirming that FedHV's low overhead does not require full local dataset evaluation. Increasing participation M improves accuracy but raises per-round cost. For local computation, $K = 1$ makes little progress, while increasing K beyond 10 degrades MeanLoss without accuracy gains, exhibiting a classic local-progress versus drift trade-off. Detailed numerical values are provided in Appendix C.5.

6 DISCUSSION

FedHV achieves adaptive task weighting without task-update geometry, iterative server-side solvers, or extra synchronization. Empirically, it excels under imbalanced task difficulty—improving worst-task accuracy on CIFAR10–MNIST—and selects a robust, distinct trade-off across evaluation metrics. Furthermore, the observed ablation trends are qualitatively consistent with the error, participation, and drift terms characterized in Theorem 4.1. Theoretically, our convergence analysis relies on a fixed reference maintaining strictly positive true slacks; Appendix A.5 details sufficient conditions for this reference safety, leaving the broader impacts of calibration margins and objective scaling for future exploration.

AI USE STATEMENT

Generative AI tools were used to assist several stages of the research and writing workflow. Specifically, they were used for literature retrieval and the discovery of related work; brainstorming and refining research ideas; assisting with the implementation of FedHV and the evaluated baseline methods; and supporting the design and implementation of the experimental pipeline, including non-IID client-data partitioning. They were additionally used to review selected sections of the manuscript and suggest revisions to improve their organization, clarity, and readability. All suggested revisions were evaluated and finalized by the authors.

For the theoretical component, generative AI tools assisted in exploring proof strategies, drafting and refining proofs of lemmas and theoretical claims, and improving the presentation and wording of certain assumptions.

All cited works were manually inspected and verified by the authors. All AI-assisted code was reviewed and tested, and the baseline implementations were checked against their original methodological descriptions. The authors independently verified the mathematical assumptions, derivations, lemmas, and proof steps. All figures were generated from the authors' experimental outputs, and their underlying numerical values were checked for correctness. Generative AI was not used to generate synthetic datasets or to fabricate or alter experimental results. The authors reviewed all AI-assisted material and take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses established public benchmark datasets and does not collect new human-subject data. CelebA contains facial images and appearance-related attributes and may reflect demographic and representational biases in its original collection and annotations. Consequently, the reported results should not be interpreted as evidence of demographic fairness or deployment readiness. Our federated experiments simulate decentralized training and do not provide formal privacy guarantees; in particular, objective reports may reveal aggregate task-level information. The proposed method is intended as an optimization technique rather than a complete privacy or fairness mechanism.

REPRODUCIBILITY STATEMENT

Appendix A presents the formal assumptions, auxiliary results, conditional reference-safety certificate, explicit convergence analysis, and complete proofs of the theoretical claims. Appendix B documents the benchmark and objective construction, model architectures, federated data partitioning, training and evaluation protocol, baseline implementations, method-specific learning rates, FedHV hyperparameters, reference-calibration procedure, and random seeds. Appendix C reports the complete loss, hypervolume, and task-level results; learning curves; ablation studies; weight trajectories; additional diagnostics; and uncertainty across training seeds. Appendix D provides analytical communication and computational complexity comparisons together with empirical wall-clock and communication measurements. Finally, Appendix E gives the complete round-level pseudocode for FedHV. The client-partition seed, training seeds, optimization settings, participation and local-update configurations, objective-report settings, and frozen training references used in the experiments are explicitly reported.

REFERENCES

Baris Askin, Pranay Sharma, Gauri Joshi, and Carlee Joe-Wong. Federated communication-efficient multi-objective optimization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4627–4635. PMLR, 2025.

Anne Auger, Johannes Bader, Dimo Brockhoff, and Eckart Zitzler. Theory of the hypervolume indicator: Optimal μ-distributions and the choice of the reference point. In *Proceedings of the 10th ACM SIGEVO Workshop on Foundations of Genetic Algorithms*, pp. 87–102. ACM, 2009. doi: 10.1145/1527125.1527138.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Timo M. Deist, Monika Grewal, Frank J. W. M. Dankers, Tanja Alderliesten, and Peter A. N. Bosman. Multi-objective learning using HV maximization. In *Evolutionary Multi-Criterion Optimization*, volume 13970 of *Lecture Notes in Computer Science*, pp. 103–117. Springer, 2023. doi: 10.1007/978-3-031-27250-9_8.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5–6):313–318, 2012. doi: 10.1016/j.crma.2012.03.014.

Andreia P. Guerreiro, Carlos M. Fonseca, and Luís Paquete. The hypervolume indicator: Computational problems and algorithms. *ACM Computing Surveys*, 54(6), 2021. doi: 10.1145/3453474.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *NeurIPS Workshop on Federated Learning*, 2019.

Yuzheng Hu, Ruicheng Xian, Qilong Wu, Qiuling Fan, Lang Yin, and Han Zhao. Revisiting scalarization in multi-task learning: A theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multiobjective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051, 2022.

Frank P. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8(1):33–37, 1997. doi: 10.1002/ett.4460080106.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Msc thesis, Department of Computer Science, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Conrado S. Miranda and Fernando J. Von Zuben. Single-solution hypervolume maximization and its use for improving generalization of neural networks. *arXiv preprint arXiv:1602.01164*, 2016.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: 10.2307/1907266.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong, and Michinari Momma. Federated multi-objective learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39602–39625, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xiaoyuan Zhang, Xi Lin, Bo Xue, Yifan Chen, and Qingfu Zhang. Hypervolume maximization: A geometric view of pareto set learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38902–38929, 2023.

Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization*, volume 4403 of *Lecture Notes in Computer Science*, pp. 862–876. Springer, 2007. doi: 10.1007/978-3-540-70928-2_64.

A Full Theoretical Development

A.1 Probability Space and Round Filtration

Objective and round-wise notation. We analyze FedHV under the log-hypervolume objective introduced in Section 3. Throughout this section, we denote the slack of objective i by

$$s_i(\theta) := r_i - f_i(\theta),$$

and write

$$\boldsymbol{F}(\theta) := \big(f_1(\theta), \dots, f_m(\theta)\big)^\top.$$

Throughout the theoretical section, $\boldsymbol{F}(\theta)$ denotes this global objective vector, while $f_i(\theta)$ denotes its scalar components. Likewise, $\boldsymbol{r}$ is the fixed reference vector and r_i is its i-th component. The reference vector $\boldsymbol{r}$ is assumed to be chosen safely so that all true slacks remain positive along the optimization trajectory.

For every round t, define the ideal current hypervolume weights

$$w_i^{\star,t} := w_i^\star(\theta^t) = \frac{\big(r_i - f_i(\theta^t)\big)^{-1}}{\sum_{j=1}^m \big(r_j - f_j(\theta^t)\big)^{-1}}, \qquad i = 1, \dots, m.$$

We use $\boldsymbol{w}^{\star,t}$ to denote the corresponding vector of ideal weights. These weights are used only as an analytical reference.

The actual weight vector used by FedHV in communication round t is denoted by $\boldsymbol{w}^t$. Following Algorithm 1,

$$\boldsymbol{w}^0 = \frac{1}{m}\mathbf{1}_m,$$

and, for $t \geq 1$,

$$\boldsymbol{w}^t = W_{\rho_{\text{clip}}}\left(\widehat{\boldsymbol{F}}^{\,t-1}; \boldsymbol{r}\right),$$

where $W_{\rho_{\text{clip}}}$ denotes the stabilized inverse-slack normalization defined in (8). Hence, $\boldsymbol{w}^t$ is computed from objective estimates collected in round $t-1$, whereas $\boldsymbol{w}^{\star,t}$ is the ideal hypervolume weight associated with the current model θ^t.

Round history and client sampling. Let $\mathcal{F}_t$ denote the sigma-field generated by all randomness revealed before the participating client set S_t is sampled in round t. In particular, the current global model θ^t, all objective estimates collected through round $t-1$, and the stale weight vector $\boldsymbol{w}^t$ are $\mathcal{F}_t$-measurable.

Conditioned on $\mathcal{F}_t$, the participating set S_t is sampled uniformly without replacement from the N clients, with

$$|S_t| = M = pN,$$

and is independent of the current-round minibatch randomness. Consequently, because $\boldsymbol{w}^t$ is fixed conditional on $\mathcal{F}_t$, for any collection of client vectors $\{a_k(\boldsymbol{w}^t)\}_{k=1}^N$ that is $\mathcal{F}_t$-measurable,

$$\mathbb{E}\left[\frac{1}{M}\sum_{k\in S_t} a_k(\boldsymbol{w}^t)\,\middle|\,\mathcal{F}_t\right] = \frac{1}{N}\sum_{k=1}^N a_k(\boldsymbol{w}^t).$$

Let $\mathcal{U}_t^{\text{rep}}$ denote all randomness used to form the objective reports after S_t is selected and before local SGD starts. For local step s, define the within-round filtration

$$\mathcal{G}_{t,s} := \sigma\left(\mathcal{F}_t, S_t, \mathcal{U}_t^{\text{rep}}, \{B_{k,u}^t : k \in S_t,\ 0 \leq u < s\}\right),$$

where $B_{k,u}^t$ denotes the minibatch used at local step u in round t. The current local iterates $\theta_{k,s}^t$ and the weight vector $\boldsymbol{w}^t$ are $\mathcal{G}_{t,s}$-measurable. Conditionally on $\mathcal{G}_{t,s}$, the current minibatch $B_{k,s}^t$ is sampled afresh with replacement from the fixed empirical distribution on $\mathcal{D}_k$. In particular, it is independent of previous local minibatches and report randomness conditional on the fixed client dataset. The next-round history $\mathcal{F}_{t+1}$ contains the current client sample, $\mathcal{U}_t^{\text{rep}}$, all minibatches, client updates, and reported objective estimates. No independence between objective-report errors and client sampling is required by the objective-estimation bound.

A.2 Formal Assumptions

For the convergence result, fix a deterministic target horizon $T \geq 1$. All trajectory-wise assumptions and all convergence statements below concern the communication rounds $t = 0, \ldots, T-1$.

Assumption A.1 (Uniform client averaging)**.** For the theoretical result, all clients have equal aggregation weight, that is, $n_k = n/N$ and

$$f_i(\theta) = \frac{1}{N} \sum_{k=1}^{N} f_i^{(k)}(\theta).$$

The server uses the simple average of the selected-client updates and objective reports, and S_t is sampled uniformly without replacement with $|S_t| = M = pN$. Thus, conditional on $\mathcal{F}_t$, every selected-client average of an $\mathcal{F}_t$-measurable client vector is an unbiased estimator of its population average. The theorem does not cover unequal client sizes or data-weighted aggregation.

Assumption A.2 (Local smoothness and stochastic gradients)**.** For every client $k \in [N]$ and objective $i \in [m]$, the local objective $f_i^{(k)}$ is nonnegative and L_i-smooth; that is, for all $\theta, \theta' \in \mathbb{R}^d$,

$$\left\| \nabla f_i^{(k)}(\theta) - \nabla f_i^{(k)}(\theta') \right\| \leq L_i \left\| \theta - \theta' \right\|. \tag{18}$$

Consequently, the global objective

$$f_i(\theta) = \frac{1}{N} \sum_{k=1}^{N} f_i^{(k)}(\theta)$$

is also L_i-smooth.

For every round t, local step s, participating client $k \in S_t$, and objective i, let $B_{k,s}^t$ be the minibatch sampled and used at that step. Specifically, $B_{k,s}^t$ contains b indices drawn independently and uniformly with replacement from $\{1, \ldots, n_k\}$, and these draws are fresh conditional on $\mathcal{G}_{t,s}$. The same minibatch is used for all objectives, which allows their stochastic gradients to be correlated. If $g_i^{(k)}(\theta; B)$ denotes the average of the per-example gradients over B, then for the $\mathcal{G}_{t,s}$-measurable local iterate $\theta_{k,s}^t$, conditional unbiasedness follows from the sampling rule:

$$\mathbb{E}\left[g_i^{(k)}\big(\theta_{k,s}^t; B_{k,s}^t\big) \mid \mathcal{G}_{t,s} \right] = \nabla f_i^{(k)}\big(\theta_{k,s}^t\big), \tag{19}$$

This identity is exact for the stated sampling protocol. We assume only the usual uniform conditional variance bound: there exist finite constants σ_i^2 such that

$$\mathbb{E}\left[\left\| g_i^{(k)}\big(\theta_{k,s}^t; B_{k,s}^t\big) - \nabla f_i^{(k)}\big(\theta_{k,s}^t\big) \right\|^2 \mid \mathcal{G}_{t,s} \right] \leq \sigma_i^2. \tag{20}$$

The task-gradient noises produced from the same minibatch may be arbitrarily correlated across objectives; no independence across objectives is assumed. The fresh-sampling rule is stronger than the martingale-difference property needed in the proofs; the proofs do not condition on future iterates and do not require independence of all noise variables across local steps.

Assumption A.3 (Trajectory-wise client heterogeneity)**.** For each objective $i \in [m]$, define the population gradient-dissimilarity function

$$\mathsf{H}_i(\theta) := \frac{1}{N} \sum_{k=1}^{N} \left\| \nabla f_i^{(k)}(\theta) - \nabla f_i(\theta) \right\|^2.$$

There exists a finite constant ζ_i^2 such that, almost surely, for every round $t = 0, \ldots, T-1$,

$$\mathsf{H}_i(\theta^t) \leq \zeta_i^2, \qquad \mathsf{H}_i(v_{t,s}) \leq \zeta_i^2, \quad s = 0, \ldots, K,$$

where $v_{t,s}$ is the virtual trajectory in (60). This is a trajectory-wise condition on the fixed horizon: it is imposed only at the global and virtual iterates that the proof actually evaluates, not uniformly over every $\theta \in \mathcal{D}$. The actual local iterates and all connecting segments are covered by the fixed-horizon domain condition in Assumption A.4. Their deviation from the virtual trajectory is controlled by Lemma A.30, so no separate global-in-$\mathcal{D}$ heterogeneity bound is required at arbitrary local points.

Assumption A.4 (Safe reference on the fixed analysis horizon)**.** Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a convex analysis region containing the points that appear in the convergence proof. Define the safe subset

$$\mathcal{D}_\rho := \left\{ \theta \in \mathcal{D} : \min_{i \in [m]} \big(r_i - f_i(\theta)\big) \geq \rho \right\}, \qquad \rho > 0.$$

The initialization is assumed safe:

$$\theta^0 \in \mathcal{D}_\rho.$$

For a communication round t, let $\mathcal{B}_t$ be the union of the line segments used by the proof,

$$\mathcal{B}_t := [\theta^t, \theta^{t+1}] \ \cup \bigcup_{k \in S_t} \bigcup_{s=0}^{K-1} [\theta_{k,s}^t, v_{t,s}] \ \cup \bigcup_{s=0}^{K-1} [v_{t,s}, \theta^t],$$

and let

$$\mathcal{R}_t := \{\theta^t, \theta^{t+1}\} \cup \{\theta_{k,s}^t : k \in S_t,\ 0 \leq s \leq K\} \cup \{v_{t,s} : 0 \leq s \leq K\} \cup \mathcal{B}_t.$$

Here $[a, b]$ denotes the line segment joining a and b; the actual local and virtual iterates are defined in the convergence proof. For every communication round $t = 0, \ldots, T-1$, assume, almost surely,

$$\mathcal{R}_t \subseteq \mathcal{D}_\rho.$$

Equivalently, every global iterate, local iterate, virtual iterate, and connecting segment used by the proof remains in the domain on the entire deterministic horizon. This is a domain condition for the logarithm, not an independent bounded-gradient assumption. The same fixed reference vector $\boldsymbol{r}$ is used in every round. The implementation uses a separate numerical clipping floor $0 < \rho_{\text{clip}} \leq \rho$. When the server uses the round-t estimated global objective vector $\widehat{\boldsymbol{F}}^{\,t}$, the estimated slacks used to compute the next task weights are stabilized according to

$$\widehat{s}_i^{\,t} := \max\left\{ r_i - \widehat{f}_i^{\,t}, \rho_{\text{clip}} \right\}. \tag{21}$$

The task losses and their estimates are nonnegative:

$$f_i(\theta) \geq 0, \qquad \widehat{f}_i^{\,t} \geq 0,$$

and the fixed reference satisfies

$$r_i \geq \rho > 0, \qquad \forall i \in [m].$$

Remark A.5 (Safe reference versus slack stabilization). The set $\mathcal{D}_\rho$ makes explicit the domain on which the true slacks $r_i - f_i(\theta)$ are bounded below by ρ. This condition ensures that the log-hypervolume objective

$$\Phi(\theta) = -\sum_{i=1}^{m} \log\big(r_i - f_i(\theta)\big)$$

is well defined and smooth on every segment used on the fixed horizon.

The clipping operation in (21) has a different role: the numerical floor ρ_{clip} stabilizes task weights computed from estimated losses and prevents their inverse-slack factors from becoming numerically unbounded. The safe margin ρ, by contrast, is a property of the *true* slacks required by the analysis. Because $0 < \rho_{\text{clip}} \leq \rho$, clipping is inactive for the true global objective vector on $\mathcal{D}_\rho$. Numerical clipping does not enforce the safe-reference condition for the true objective slacks. Accordingly, the convergence analysis is a fixed-horizon result under Assumption A.4. The practical fixed-reference selection used in the experiments is an empirical calibration rule and does not by itself verify this almost-sure domain condition; it is described in Appendix B.8. Appendix A.5 gives separate sufficient conditions that convert a verified initial margin into a high-probability trajectory certificate.

A.3 Basic Properties

Lemma A.6 (Self-bounded global gradients)**.** *Under Assumptions A.1 and A.2, whenever* $f_i(\theta) \leq r_i$*,*

$$\|\nabla f_i(\theta)\|^2 \leq 2L_i f_i(\theta) \leq 2L_i r_i =: G_i^2.$$

Consequently, with $G_{\max} := \max_i G_i$, every global scalarized gradient with weights in Δ_m satisfies

$$\left\| \sum_{i=1}^{m} w_i \nabla f_i(\theta) \right\| \le G_{\max}.$$

This bound is derived from nonnegativity and smoothness; it is not an independent bounded-gradient assumption. Its ρ-free form avoids circular constants when Appendix A.5 derives a safe margin rather than assuming one.

Weighted stochastic-gradient noise. At local step s of communication round t, participating client $k \in S_t$ uses the stale weight vector $\boldsymbol{w}^t$, which is $\mathcal{F}_t$-measurable, and forms the weighted stochastic gradient

$$g_{k,s}^t := \sum_{i=1}^{m} w_i^t \, g_i^{(k)}\big(\theta_{k,s}^t; B_{k,s}^t\big).$$

The corresponding exact local weighted gradient is

$$\nabla L_{\boldsymbol{w}^t}^{(k)}\big(\theta_{k,s}^t\big) = \sum_{i=1}^{m} w_i^t \nabla f_i^{(k)}\big(\theta_{k,s}^t\big).$$

In the implementation this is the gradient of

$$\sum_{i=1}^{m} \mathrm{sg}\big(w_i^t\big) \; \ell_i(\theta; B_{k,s}^t),$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Thus the weights are held fixed during all K local steps and no derivative through the weight-estimation map is taken. We define the weighted stochastic-gradient noise by

$$\xi_{k,s}^t := g_{k,s}^t - \nabla L_{\boldsymbol{w}^t}^{(k)}\big(\theta_{k,s}^t\big).$$

Writing $\varepsilon_{i,k,s}^t := g_i^{(k)}(\theta_{k,s}^t; B_{k,s}^t) - \nabla f_i^{(k)}(\theta_{k,s}^t)$, we have $\xi_{k,s}^t = \sum_i w_i^t \varepsilon_{i,k,s}^t$. Since $\boldsymbol{w}^t \in \Delta_m$, Jensen's inequality gives

$$\mathbb{E}\left[\|\xi_{k,s}^t\|^2 \middle| \mathcal{G}_{t,s} \right] \le \sum_{i=1}^{m} w_i^t \sigma_i^2 \le \Sigma^2,$$

without requiring the task noises generated by a shared minibatch to be uncorrelated. For the average over the participating clients, let κ_M denote any valid bound satisfying

$$\mathbb{E}\left[\left\| \frac{1}{M} \sum_{k \in S_t} \xi_{k,s}^t \right\|^2 \middle| \mathcal{G}_{t,s} \right] \le \kappa_M \Sigma^2.$$

Jensen's inequality always gives $\kappa_M = 1$, which is the value used in the main theorem. Under the additional cross-client condition (CCO-grad) stated in Remark A.23, the usual federated variance reduction gives $\kappa_M = 1/M$.

Lemma A.7 (Martingale-difference weighted noise)**.** *Under Assumption A.2, the weighted noise defined above satisfies, for every $k \in S_t$ and local step s:*

$$\mathbb{E}\left[\xi_{k,s}^t \mid \mathcal{G}_{t,s} \right] = \mathbf{0},$$

and

$$\mathbb{E}\left[\|\xi_{k,s}^t\|^2 \mid \mathcal{G}_{t,s} \right] \le \Sigma^2, \qquad \Sigma^2 := \sum_{i=1}^{m} \sigma_i^2.$$

This is a consequence of the fresh conditional minibatch sampling rule and is not an additional stochastic-gradient assumption. No independence across objectives, clients, or local steps is required for the martingale-difference statement. The client-averaged bound with $\kappa_M = 1$ follows directly from Jensen's inequality. Under (CCO-grad) in Remark A.23, one may sharpen this bound to $\kappa_M = 1/M$. This optional condition is the only place where a $1/M$ stochastic-variance factor is used.

Lemma A.8 (Smoothness of the log-hypervolume objective)**.** *Under Assumptions A.2 and A.4, together with Lemma A.6, the log-hypervolume objective is L_Φ-smooth on every line segment contained in $\mathcal{D}_\rho$; specifically, for any $\theta, \theta' \in \mathcal{D}$ with $[\theta, \theta'] \subseteq \mathcal{D}_\rho$,*

$$\|\nabla\Phi(\theta) - \nabla\Phi(\theta')\| \le L_\Phi\|\theta - \theta'\|, \qquad L_\Phi \le \sum_{i=1}^m \frac{L_i}{\rho} + \sum_{i=1}^m \frac{G_i^2}{\rho^2}.$$

The proof is given in Appendix A.12.3.

Assumption A.9 (Stepsizes)**.** Let

$$L_w := \sum_{i=1}^m L_i, \qquad c^t := \sum_{i=1}^m \big(r_i - f_i(\theta^t)\big)^{-1},$$

and define

$$c_{\min} := \sum_{i=1}^m \frac{1}{r_i}, \qquad c_{\max} := \frac{m}{\rho}.$$

Let $\eta_\ell > 0$ and $\eta_g > 0$. The local and effective stepsizes satisfy

$$\eta_\ell K \le \frac{c_1}{L_w}, \qquad \eta_{\text{eff}} := \eta_g\eta_\ell K \le \frac{c_2 c_{\min}^2}{c_{\max} L_\Phi}, \tag{22}$$

where, throughout the theorem and its proof, we fix the numerical values

$$c_1 := \frac{1}{4}, \qquad c_2 := \frac{1}{48}.$$

Thus there is no unspecified "sufficiently small" constant in the statement. The value of c_1 gives $\eta_\ell L_w s \le 1/4$ for every $s \le K$, which is the condition used in Lemma A.30. Since the numerical coefficient C_0 in Lemma A.32 is $C_0 = 6$, the value $c_2 = 1/(8C_0) = 1/48$ leaves the explicit absorption margin used in the one-round descent proof. Define for later reference

$$\bar{\eta}_{\text{eff}} := \frac{c_2 c_{\min}^2}{c_{\max} L_\Phi}.$$

Lemma A.10 (Log-hypervolume stationarity implies Pareto stationarity)**.** *If $\nabla\Phi(\theta^\star) = 0$, then there exists $\lambda \in \Delta_m$ such that*

$$\sum_{i=1}^m \lambda_i \nabla f_i(\theta^\star) = 0.$$

Therefore, every stationary point of the log-hypervolume objective is Pareto-stationary. The proof is given in Appendix A.12.1.

A.4 Barrier Control of the Minimum Task Slack

The log-hypervolume objective is also a logarithmic barrier for the feasible box $\{\theta : f_i(\theta) < r_i,\ i \in [m]\}$.

Lemma A.11 (Barrier level set and slack floor)**.** *Let $s_i(\theta) := r_i - f_i(\theta) > 0$, and suppose $s_j(\theta) \le S_j$ for constants $S_j > 0$. Define*

$$R_{\max}(\boldsymbol{S}) := \max_{i\in[m]} \prod_{j\ne i} S_j, \qquad \psi(\theta) := -\log \min_i s_i(\theta).$$

Then

$$\min_i s_i(\theta) \ge \frac{e^{-\Phi(\theta)}}{R_{\max}(\boldsymbol{S})}, \tag{23}$$

and

$$\psi(\theta) - \log R_{\max}(\boldsymbol{S}) \le \Phi(\theta) \le m\psi(\theta). \tag{24}$$

Because every objective is nonnegative, $S_j = r_j$ is always a valid choice.

Proof. For any i, the identity $\sum_j \log s_j(\theta) = -\Phi(\theta)$ gives

$$\log s_i(\theta) = -\Phi(\theta) - \sum_{j \neq i} \log s_j(\theta) \geq -\Phi(\theta) - \log R_{\max}(\boldsymbol{S}).$$

Exponentiating and minimizing over i proves (23). For the right inequality in (24), each $-\log s_i \leq \psi$. For the left inequality, choose $i^\star \in \arg\max_i(-\log s_i)$ and use

$$\Phi(\theta) = \psi(\theta) + \sum_{j \neq i^\star} -\log s_j(\theta) \geq \psi(\theta) - \log R_{\max}(\boldsymbol{S}).$$

□

Taking $\Phi(\theta) \leq B$ in (23) proves Proposition 4.4: every task obeys $f_i(\theta) \leq r_i - e^{-B}/R_{\max}(\boldsymbol{S})$. The guarantee concerns reference-relative objective losses. It does not by itself convert to a worst-task test-accuracy guarantee.

A.5 Conditional Reference-Safety Certificate

The base convergence theorem conditions on Assumption A.4. The following separate result shows how an initial true-slack margin can certify that domain condition with high probability under additional bounded-update requirements. It does not claim that the empirical calibration rule alone verifies them.

Assumption A.12 (Almost-surely bounded local updates)**.** There is $G_{\mathrm{sg}} \geq G_{\max}$ such that the weighted local stochastic gradient satisfies $\|g_{k,s}^t\| \leq G_{\mathrm{sg}}$ almost surely for every t, k, s.

Lemma A.13 (Round-excursion buffer)**.** *Let*

$$\Delta_{\max} := \eta_\ell K G_{\mathrm{sg}} \max\{1, \eta_g\},$$

and suppose

$$G_{\max}\Delta_{\max} + \frac{L_{\max}}{2}\Delta_{\max}^2 \leq \frac{\rho}{2}, \qquad L_{\max} := \max_i L_i. \tag{25}$$

If $\theta^t \in \mathcal{D}_\rho$, then every global, local, and virtual iterate and every connecting segment in $\mathcal{R}_t$ lies in $\mathcal{D}_{\rho/2}$ almost surely.

Proof. For any θ' with $\|\theta' - \theta^t\| \leq \Delta_{\max}$, smoothness and Lemma A.6 give

$$f_i(\theta') \leq f_i(\theta^t) + G_{\max}\Delta_{\max} + \frac{L_{\max}}{2}\Delta_{\max}^2 \leq f_i(\theta^t) + \frac{\rho}{2}.$$

Thus the closed ball of radius $\Delta_{\max}$ around θ^t is contained in $\mathcal{D}_{\rho/2}$. Assumption A.12 gives $\|\theta_{k,s}^t - \theta^t\| \leq \eta_\ell K G_{\mathrm{sg}}$ and $\|\theta^{t+1} - \theta^t\| \leq \eta_g \eta_\ell K G_{\mathrm{sg}}$. The same induction for the virtual trajectory uses $\|\nabla L_{\boldsymbol{w}^t}(v_{t,s})\| \leq G_{\max}$. All these points lie in the ball, and convexity of the ball contains the required segments. □

Lemma A.14 (Maximal Azuma inequality)**.** *If $(Y_t)_{t \geq 0}$ is a supermartingale with $Y_0 = 0$ and $|Y_{t+1} - Y_t| \leq c$ almost surely, then*

$$\mathbb{P}\left(\max_{0 \leq t \leq T} Y_t \geq \lambda\right) \leq \exp\left(-\frac{\lambda^2}{2Tc^2}\right).$$

Proof. Hoeffding's lemma makes $\exp(\gamma Y_t - \gamma^2 c^2 t/2)$ a nonnegative supermartingale. Ville's maximal inequality therefore bounds the displayed probability by $\exp(-\gamma\lambda + \gamma^2 c^2 T/2)$; minimizing at $\gamma = \lambda/(c^2 T)$ proves the result. □

Theorem A.15 (Conditional self-certification of reference safety)**.** *Assume $s_i(\theta^0) > 0$, Assumption A.12, and* (25)*. Let $R_r := \max_i \prod_{j \neq i} r_j$, $\eta_{\mathrm{eff}} := \eta_g \eta_\ell K$, and*

$$L_\Phi(\varrho) := \sum_i \frac{L_i}{\varrho} + \sum_i \frac{G_i^2}{\varrho^2}.$$

For any $\mathcal{F}_t$-measurable weights $\boldsymbol{w}^t \in \Delta_m$, set

$$\mu(\rho) := \frac{mG_{\max}}{\rho}\left(\eta_{\text{eff}}G_{\max} + \frac{1}{2}\eta_g\eta_\ell^2 K^2 L_w G_{\text{sg}}\right) + \frac{1}{2}L_\Phi(\rho/2)\eta_{\text{eff}}^2 G_{\text{sg}}^2,$$

$$D(\rho) := \frac{2mG_{\max}}{\rho}\eta_{\text{eff}}G_{\text{sg}} + \frac{1}{2}L_\Phi(\rho/2)\eta_{\text{eff}}^2 G_{\text{sg}}^2.$$

If, for $\delta \in (0,1)$,

$$\Phi(\theta^0) + T\mu(\rho) + \big(D(\rho) + \mu(\rho)\big)\sqrt{2T\log(1/\delta)} \le \log\frac{1}{\rho R_r}, \tag{26}$$

then, with probability at least $1-\delta$, $\theta^t \in \mathcal{D}_\rho$ for every $t \le T$ and $\mathcal{R}_t \subseteq \mathcal{D}_{\rho/2}$ for every $t < T$. Hence the domain condition of Assumption A.4 holds at level $\rho/2$ on this event.

Proof. Let $B := \log(1/(\rho R_r))$. By Lemma A.11, $\{\Phi \le B\} \subseteq \mathcal{D}_\rho$. Define the stopping time $\tau := \min\{t \ge 0 : \Phi(\theta^t) > B\}$, with $\min \varnothing = \infty$. Condition (26) implies $\tau \ge 1$.

On $\{t < \tau\}$, Lemma A.13 places the whole round in $\mathcal{D}_{\rho/2}$. Smoothness of Φ on that set, the identity

$$\mathbb{E}[\Delta^t \mid \mathcal{F}_t] = -\eta_\ell K \nabla L_{\boldsymbol{w}^t}(\theta^t) + b_t, \qquad \Delta^t := \frac{1}{M}\sum_{k\in S_t}\Delta\theta_k^t,$$

and local smoothness give

$$\|b_t\| \le \frac{1}{2}L_w G_{\text{sg}}\eta_\ell^2 K^2, \qquad \|\Delta^t\| \le \eta_\ell K G_{\text{sg}}.$$

Together with $\|\nabla\Phi(\theta^t)\| \le mG_{\max}/\rho$, these bounds yield

$$\mathbb{E}[\Phi(\theta^{t+1}) - \Phi(\theta^t) \mid \mathcal{F}_t] \le \mu(\rho),$$
$$|\Phi(\theta^{t+1}) - \Phi(\theta^t)| \le D(\rho)$$

on $\{t < \tau\}$. The second inequality uses the sharper bound $\|\nabla\Phi\| \le 2mG_{\max}/\rho$ along the safe segment.

Now set $X_t := \Phi(\theta^{t\wedge\tau})$ and $Y_t := X_t - X_0 - (t\wedge\tau)\mu(\rho)$. Then (Y_t) is a supermartingale with increments bounded by $D(\rho)+\mu(\rho)$. Lemma A.14 and (26) imply, with probability at least $1-\delta$,

$$\Phi(\theta^{t\wedge\tau}) \le \Phi(\theta^0) + T\mu(\rho) + \big(D(\rho)+\mu(\rho)\big)\sqrt{2T\log(1/\delta)} \le B$$

simultaneously for all $t \le T$. If $\tau \le T$, substituting $t = \tau$ contradicts the definition of τ. Thus $\tau > T$, which proves the claimed safety statements. □

Remark A.16 (Scope of the certificate). The theorem starts from a true-slack-safe initialization and requires almost-surely bounded local stochastic gradients plus the explicit buffer and budget inequalities. These conditions can in principle be checked after selecting a calibration margin, but they are not verified numerically for the present experiments. The result therefore narrows the theory–practice gap without claiming that the reported calibration heuristic automatically certifies Assumption A.4. Moreover, Theorem 4.1 retains its original expectation statement; the high-probability safety event is not used as an implicit conditioning step in that theorem.

Finite-population participation factor and heterogeneity. Using the number of participating clients

$$M := |S_t| = pN,$$

define

$$\chi_N(M) := \frac{N-M}{M(N-1)} = \frac{1-p}{p(N-1)}.$$

The gradient-side partial-participation term is

$$\mathcal{H}_N(p;\zeta) := \chi_N(M)\sum_{i=1}^m \zeta_i^2.$$

In particular,

$$\mathcal{H}_N(p;\zeta) \le \frac{1-p}{p}\sum_{i=1}^m \zeta_i^2.$$

Here ζ_i^2 is the trajectory-wise envelope from Assumption A.3 on the fixed analysis horizon; it is not a supremum over all points in $\mathcal{D}$.

A.6 WEIGHT-MAP STABILITY

Notation (HV weight map and stale weights). For any nonnegative objective vector $z = (z_1, \dots, z_m)^\top$, define the stabilized inverse-slack weight map $W_{\rho_{\mathrm{clip}}}(\cdot; \boldsymbol{r})$ componentwise by

$$[W_{\rho_{\mathrm{clip}}}(z; \boldsymbol{r})]_i := \frac{\big(\max\{r_i - z_i, \rho_{\mathrm{clip}}\}\big)^{-1}}{\sum_{j=1}^m \big(\max\{r_j - z_j, \rho_{\mathrm{clip}}\}\big)^{-1}}, \qquad i = 1, \dots, m.$$

Under Assumption A.4 and $0 < \rho_{\mathrm{clip}} \le \rho$, clipping is inactive for the true global objective vector $\boldsymbol{F}(\theta)$. Hence, the ideal hypervolume weight vector can be written as

$$\boldsymbol{w}^\star(\theta) = W_{\rho_{\mathrm{clip}}}(\boldsymbol{F}(\theta); \boldsymbol{r}).$$

In particular,

$$\boldsymbol{w}^{\star,t} := \boldsymbol{w}^\star(\theta^t) = W_{\rho_{\mathrm{clip}}}(\boldsymbol{F}(\theta^t); \boldsymbol{r}).$$

The actual FedHV weight used in round t is

$$\boldsymbol{w}^0 = \frac{1}{m}\mathbf{1}_m, \qquad \boldsymbol{w}^t = W_{\rho_{\mathrm{clip}}}(\widehat{\boldsymbol{F}}^{\,t-1}; \boldsymbol{r}), \quad t \ge 1.$$

Thus, $\boldsymbol{w}^t$ is measurable with respect to the round history $\mathcal{F}_t$ and is fixed before S_t is sampled.

We write $\|\cdot\|_\infty$ for the sup norm on $\mathbb{R}^m$ and $\|\cdot\|_1$ for the ℓ_1 norm.

Under Assumptions A.4 and A.9 and the nonnegativity of the losses,

$$c^t \in [c_{\min}, c_{\max}].$$

Lemma A.17 (Sharp calculus of the stabilized HV weight map)**.** *For $a \in (0,\infty)^m$, define $V_i(a) := a_i^{-1} / \sum_j a_j^{-1}$ and $a_{\min} := \min_i a_i$. Then*

$$\sum_{i,j} \left|\frac{\partial V_i}{\partial a_j}\right| \le \frac{2}{a_{\min}}, \qquad \sum_{i,j,k} \left|\frac{\partial^2 V_i}{\partial a_j \partial a_k}\right| \le \frac{8}{a_{\min}^2}. \tag{27}$$

Consequently, for all $z, z' \in \mathbb{R}^m$,

$$\|W_{\rho_{\mathrm{clip}}}(z; \boldsymbol{r}) - W_{\rho_{\mathrm{clip}}}(z'; \boldsymbol{r})\|_1 \le \frac{2}{\rho_{\mathrm{clip}}}\|z - z'\|_\infty. \tag{28}$$

Thus the notation used below may take $C(\rho_{\mathrm{clip}}, \boldsymbol{r}) := 2/\rho_{\mathrm{clip}}$, which is dimension-free and independent of $\boldsymbol{r}$. If the whole segment $[z, z']$ has unclipped slacks at least $\widetilde{\rho} \ge \rho_{\mathrm{clip}}$, the sharper local constant $2/\widetilde{\rho}$ holds. If $\widetilde{\rho} > \rho_{\mathrm{clip}}$, then, with $J_W(z') := \partial_z W_{\rho_{\mathrm{clip}}}(z'; \boldsymbol{r})$,

$$\|W(z) - W(z') - J_W(z')(z - z')\|_1 \le \frac{4}{\widetilde{\rho}^2}\|z - z'\|_\infty^2, \qquad \|J_W(z')\|_{\infty\to 1} \le \frac{2}{\widetilde{\rho}}, \tag{29}$$

where $W = W_{\rho_{\mathrm{clip}}}(\cdot; \boldsymbol{r})$. The proof is given in Appendix A.12.4.

A.7 OBJECTIVE-REPORT ANALYSIS

Assumption A.18 (Objective-report sampling)**.** During round t, after S_t is selected and at the common pre-update model θ^t, each participating client k draws a report index set

$$\mathcal{B}_{k,t}^{\mathrm{rep}} \subseteq \{1, \dots, n_k\}, \qquad \left|\mathcal{B}_{k,t}^{\mathrm{rep}}\right| = b_{k,t}^{\mathrm{rep}},$$

uniformly without replacement. Its report for objective i is the minibatch average

$$\widehat{f}_{i,k}^{\,t} := \frac{1}{b_{k,t}^{\mathrm{rep}}} \sum_{j \in \mathcal{B}_{k,t}^{\mathrm{rep}}} \ell_i(\theta^t; x_{k,j}), \qquad k \in S_t. \tag{30}$$

Let q_t denote the configured number of report batches and let b be the local minibatch size. The implementation uses

$$b_{k,t}^{\mathrm{rep}} = \begin{cases} n_k, & q_t = 0, \\ \min\{q_t b, n_k\}, & q_t \ge 1. \end{cases}$$

The notation matches the experimental budget as

$$q_t = B_{\mathrm{rep}} \quad \text{for finite } B_{\mathrm{rep}}, \qquad q_t = 0 \quad \Longleftrightarrow \quad B_{\mathrm{rep}} = \mathrm{all}.$$

Thus $q_t = 0$ evaluates the complete client dataset, while $q_t \geq 1$ takes the first q_t batches of a fresh random permutation, which is equivalent to a uniformly sampled subset of $b_{k,t}^{\mathrm{rep}}$ examples. The report loader uses `drop_last=False`.

Define the report error

$$\nu_{i,k}^t := \widehat{f}_{i,k}^t - f_i^{(k)}(\theta^t),$$

and the within-client finite-population dispersion

$$v_{i,k}(\theta) := \frac{1}{n_k} \sum_{j=1}^{n_k} \Big(\ell_i(\theta; x_{k,j}) - f_i^{(k)}(\theta) \Big)^2 .$$

For $n_k > 1$, set

$$\gamma_{k,t} := \frac{n_k - b_{k,t}^{\mathrm{rep}}}{b_{k,t}^{\mathrm{rep}}(n_k - 1)},$$

and set $\gamma_{k,t} = 0$ when $n_k = 1$. For the fixed target horizon T, assume the finite-second-moment bounds

$$\tau_{i,T}^2 := \max_{0 \leq t < T} \mathbb{E}\left[\frac{1}{N} \sum_{k=1}^{N} \gamma_{k,t} v_{i,k}(\theta^t) \right] < \infty, \tag{31}$$

and

$$\omega_{i,T}^2 := \max_{0 \leq t < T} \mathbb{E}\left[\frac{1}{N} \sum_{k=1}^{N} \big(f_i^{(k)}(\theta^t) - f_i(\theta^t) \big)^2 \right] < \infty.$$

For fixed n_k, $\gamma_{k,t}$ decreases as the report sample size $b_{k,t}^{\mathrm{rep}}$ increases. In particular, evaluating the full local client dataset gives $\gamma_{k,t} = 0$ and therefore removes the within-client report-sampling contribution $\tau_{i,T}^2$; client subsampling can still leave a nonzero between-client contribution through $\omega_{i,T}^2$. The conditional zero-mean and variance of $\nu_{i,k}^t$ are derived in Lemma A.19, rather than postulated.

For distinct clients, we use the optional cross-client orthogonality condition

$$\mathbb{E}[\nu_{i,k}^t \nu_{i,\ell}^t \mid \mathcal{F}_t, S_t] = 0, \qquad k \neq \ell,$$

only when deriving the usual $1/M$ reduction for averaged report noise. The main objective-report bound remains valid without this condition by Jensen's inequality. Report randomness is generated before local SGD and is included in $\mathcal{U}_t^{\mathrm{rep}}$ within the local-step filtration. The report and the current client update may share the same client data; no independence between them is used in the report-error proof. The server estimate is

$$\widehat{f}_i^t := \frac{1}{M} \sum_{k \in S_t} \widehat{f}_{i,k}^t, \qquad \widehat{\boldsymbol{F}}^t := (\widehat{f}_1^t, \ldots, \widehat{f}_m^t)^\top.$$

The estimate is used to construct $\boldsymbol{w}^{t+1}$, rather than the weight vector used by the same cohort S_t.

Lemma A.19 (Finite-population report moments)**.** *Under Assumption A.18, for every objective i and participating client k,*

$$\mathbb{E}\left[\widehat{f}_{i,k}^t \middle| \mathcal{F}_t, S_t \right] = f_i^{(k)}(\theta^t),$$

and, for $n_k > 1$,

$$\mathbb{E}\left[(\nu_{i,k}^t)^2 \middle| \mathcal{F}_t, S_t \right] = \gamma_{k,t} v_{i,k}(\theta^t).$$

For $n_k = 1$, the report is exact and both sides of the second identity are zero.

Lemma A.20 (Finite-population objective-report error)**.** *Under Assumptions A.1 and A.18, define*

$$\Xi_{\mathrm{loss},T} := \max_{0 \leq t < T} \mathbb{E}\left[\left\| \widehat{\boldsymbol{F}}^t - \boldsymbol{F}(\theta^t) \right\|_\infty^2 \right].$$

Without any cross-client orthogonality assumption,

$$\Xi_{\mathrm{loss},T} \le \sum_{i=1}^{m} \left[\chi_N(M)\omega_{i,T}^2 + \tau_{i,T}^2\right].$$

If, in addition, the cross-client report condition

$$\mathbb{E}\left[\nu_{i,k}^t \nu_{i,\ell}^t \middle| \mathcal{F}_t, S_t\right] = 0, \qquad k \neq \ell, \tag{CCO-report}$$

holds, the final term improves from $\tau_{i,T}^2$ to $\tau_{i,T}^2/M$. The theorem uses the first, assumption-free bound. Thus participation affects report quality through $\chi_N(M)\omega_{i,T}^2$, while the report budget affects $\tau_{i,T}^2$. Full participation and full-client reports jointly make the displayed upper bound zero.

A.8 One-Round-Stale Weight Error

Lemma A.21 (One-round-stale weight discrepancy)**.** *For every round $1 \le t < T$ under the fixed-horizon condition in Assumption A.4, the discrepancy between the stale FedHV weight $\boldsymbol{w}^t$ used by the algorithm and the ideal current hypervolume weight $\boldsymbol{w}^{\star,t}$ satisfies*

$$\begin{aligned}\left\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\right\|_1 \le{}& C(\rho_{\mathrm{clip}}, \boldsymbol{r}) \left\|\widehat{\boldsymbol{F}}^{\,t-1} - \boldsymbol{F}(\theta^{t-1})\right\|_\infty \\ &+ C(\rho_{\mathrm{clip}}, \boldsymbol{r}) \left\|\boldsymbol{F}(\theta^{t-1}) - \boldsymbol{F}(\theta^t)\right\|_\infty.\end{aligned} \tag{32}$$

Consequently,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\right\|_1^2\right] \le{}& 2C(\rho_{\mathrm{clip}}, \boldsymbol{r})^2 \Xi_{\mathrm{loss},T} \\ &+ 2C(\rho_{\mathrm{clip}}, \boldsymbol{r})^2 \mathbb{E}\left[\left\|\boldsymbol{F}(\theta^t) - \boldsymbol{F}(\theta^{t-1})\right\|_\infty^2\right].\end{aligned} \tag{33}$$

For the initialization $t = 0$, both $\boldsymbol{w}^0$ and $\boldsymbol{w}^{\star,0}$ belong to the simplex Δ_m, and therefore

$$\left\|\boldsymbol{w}^0 - \boldsymbol{w}^{\star,0}\right\|_1 \le 2.$$

A.9 Explicit Convergence Theorem and Rate

Theorem A.22 (Explicit form of Theorem 4.1)**.** *Under Assumptions A.1, A.2, A.3, A.4, A.9, and A.18, with client participation rate $p \in (0, 1]$ and the deterministic target horizon $T \ge 1$ fixed above. The fixed-horizon domain condition in Assumption A.4 ensures that every point used in the proof lies in $\mathcal{D}_\rho$. The main theorem uses the Jensen-valid choice $\kappa_M = 1$, so it does not require a cross-client orthogonality assumption. For the fixed-horizon process,*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla\Phi(\theta^t)\|^2\right] \le{}& \frac{A_1\big(\Phi(\theta^0) - \Phi_{\mathrm{lb}}\big)}{\eta_{\mathrm{eff}} T} + A_2 \eta_{\mathrm{eff}} \Sigma^2 \\ &+ A_3 \eta_{\mathrm{eff}} \mathcal{H}_N(p;\zeta) + A_4 \eta_\ell^2 (K-1)(2K-1) \\ &+ A_5 \Xi_{\mathrm{loss},T} + A_6 \eta_{\mathrm{eff}}^2 \mathcal{V} + \frac{A_7}{T}.\end{aligned} \tag{34}$$

Here

$$\eta_{\mathrm{eff}} := \eta_g \eta_\ell K, \qquad \Sigma^2 := \sum_{i=1}^m \sigma_i^2, \qquad \zeta^2 := \sum_{i=1}^m \zeta_i^2, \qquad Q := \zeta^2 + \Sigma^2,$$

$$G_{\max} := \max_{i\in[m]} G_i, \qquad C_\delta := 4, \qquad \bar{G}^2 := 2\big(G_{\max}^2 + C_\delta Q\big).$$

The quantity

$$\mathcal{V} := 6G_{\max}^2 + 3\frac{\Sigma^2}{K} + 6\mathcal{H}_N(p;\zeta) + \frac{L_w^2 \bar{G}^2}{2}\eta_\ell^2 (K-1)(2K-1) \tag{35}$$

controls the inter-round model displacement that generates the one-round weight-staleness error. The lower-bound term is explicit:

$$\Phi_{\mathrm{lb}} := -\sum_{i=1}^m \log r_i.$$

The fourth term captures the error caused by multiple local steps and vanishes when $K = 1$. The fifth term captures error in the client-reported objective estimates. The sixth term is the additional error caused by using the one-round-stale weight vector instead of the ideal current hypervolume weights. The final A_7/T term is the one-time initialization contribution. The constants in (34) *can be chosen explicitly. Let*

$$\bar{\eta}_{\text{eff}} := \frac{c_2 c_{\min}^2}{c_{\max} L_\Phi}, \qquad E_{\text{wt}} := c_{\max} + 6L_\Phi \bar{\eta}_{\text{eff}},$$

where $c_1 = 1/4$ and $c_2 = 1/48$ are fixed in Assumption A.9. With the sharp global constant $C(\rho_{\text{clip}}, \boldsymbol{r}) = 2/\rho_{\text{clip}}$ from Lemma A.17, one admissible choice is

$$\begin{aligned}
&A_1 := 4c_{\max}, && A_2 := 6c_{\max} L_\Phi, && A_3 := 12 c_{\max} L_\Phi,\\
&A_4 := c_{\max} L_w^2 \bar{G}^2 \big(c_{\max} + L_\Phi \bar{\eta}_{\text{eff}}\big), && A_5 := 8 c_{\max} E_{\text{wt}} G_{\max}^2 C(\rho_{\text{clip}}, \boldsymbol{r})^2,\\
&A_6 := 8 c_{\max} E_{\text{wt}} G_{\max}^4 C(\rho_{\text{clip}}, \boldsymbol{r})^2, && A_7 := 16 c_{\max} E_{\text{wt}} G_{\max}^2.
\end{aligned}$$

Thus A_1, A_2, A_3 depend only on the displayed smoothness and slack constants; A_4 additionally depends on $L_w, \bar{G}$; and A_5, A_6, A_7 additionally depend on $G_{\max}$ and the explicit weight-map Lipschitz constant $C(\rho_{\text{clip}}, \boldsymbol{r})$. None of the A_j depends on $T, N, p, \eta_\ell, \eta_g$, or K; dependence on Σ, ζ occurs only through the displayed quantities $\mathcal{H}_N, \mathcal{V}$, and $\bar{G}$. The bound is a fixed-horizon ergodic stationarity-to-neighborhood guarantee. For a family of target horizons, the report contribution is governed by the corresponding sequence $\Xi_{\text{loss},T}$; no asymptotic conclusion is claimed without controlling that sequence. Partial participation and finite report minibatches are two mechanisms that can keep $\Xi_{\text{loss},T}$ positive. The proof is provided in Appendix A.12.9; it telescopes the ordinary fixed-horizon increments. Consequently, the displayed average also bounds

$$\min_{0 \le t < T} \mathbb{E}\|\nabla \Phi(\theta^t)\|^2.$$

If the fixed-horizon condition in Assumption A.4 cannot be verified for a chosen T, this theorem is not asserted for that horizon.

Remark A.23 (Optional cross-client variance specializations). The main theorem uses only Jensen's inequality. A sharper $1/M$ factor is available only under an additional, formally stated cross-client condition. Namely, if for every local step s and every distinct $k, \ell \in S_t$,

$$\mathbb{E}\left[\left\langle \xi_{k,s}^t, \xi_{\ell,s}^t \right\rangle \middle| \mathcal{G}_{t,s}\right] = 0, \tag{CCO-grad}$$

then the client-averaged stochastic-gradient term may be sharpened by replacing Σ^2 with Σ^2/M in the stochastic terms of Theorem A.22 and in $\mathcal{V}$. Likewise, if the condition (CCO-report) stated in Lemma A.20 holds, the report-noise contribution in the bound for $\Xi_{\text{loss},T}$ improves from $\tau_{i,T}^2$ to $\tau_{i,T}^2/M$. These are optional refinements and are not part of the main guarantee.

Corollary A.24 (Rate and stationarity neighborhood)**.** *Under the conditions of Theorem A.22, fix a deterministic target horizon $T \ge 1$, set $\eta_g = \bar{\eta}_g > 0$, and choose*

$$\eta_\ell = \frac{\eta_0}{\bar{\eta}_g K \sqrt{T}},$$

where

$$0 < \eta_0 \le \min\left\{ \frac{\bar{\eta}_g c_1}{L_w}, \bar{\eta}_{\text{eff}} \right\}.$$

Then $\eta_{\text{eff}} = \eta_0/\sqrt{T}$, and

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla \Phi(\theta^t)\|^2\right] \le{}& \frac{A_1(\Phi(\theta^0) - \Phi_{\text{lb}})}{\eta_0 \sqrt{T}} + \frac{A_2 \eta_0 \Sigma^2}{\sqrt{T}}\\
&+ \frac{A_3 \eta_0 \mathcal{H}_N(p;\zeta)}{\sqrt{T}} + \frac{A_4 \eta_0^2 (K-1)(2K-1)}{\bar{\eta}_g^2 K^2 T}\\
&+ A_5 \Xi_{\text{loss},T} + \frac{A_6 \eta_0^2}{T}\mathcal{V}_T + \frac{A_7}{T},
\end{aligned} \tag{36}$$

where

$$\mathcal{V}_T := 6G_{\max}^2 + 3\frac{\Sigma^2}{K} + 6\mathcal{H}_N(p;\zeta) + \frac{L_w^2\bar{G}^2}{2}\frac{\eta_0^2}{\bar{\eta}_g^2 K^2 T}(K-1)(2K-1).$$

Consequently, the displayed ergodic bound implies

$$\min_{0\le t<T}\mathbb{E}\|\nabla\Phi(\theta^t)\|^2 \le \mathcal{O}(T^{-1/2}) + A_5\Xi_{\mathrm{loss},T}.$$

This statement should be read as a family of horizon-T guarantees, because the chosen η_ℓ depends on the target horizon. If $\Xi_{\mathrm{loss},T} = 0$ for a given horizon, the stochastic optimization terms scale as $\mathcal{O}(T^{-1/2})$; more generally, the report term determines the stationarity neighborhood. The experiments use fixed configured stepsizes rather than this theoretical horizon-dependent schedule.

A.10 Pareto Stationarity and Special Cases

Corollary A.25 (Quantitative approximate Pareto stationarity)**.** *Define the Pareto-stationarity residual*

$$\mathcal{P}(\theta) := \min_{\lambda\in\Delta_m}\left\|\sum_{i=1}^m \lambda_i\nabla f_i(\theta)\right\|. \tag{37}$$

Then, for every θ,

$$\mathcal{P}(\theta) \le \frac{\|\nabla\Phi(\theta)\|}{c(\theta)} \le \frac{\|\nabla\Phi(\theta)\|}{c_{\min}}, \tag{38}$$

where

$$c(\theta) = \sum_{i=1}^m \frac{1}{r_i - f_i(\theta)}.$$

Consequently, on the fixed horizon $t = 0,\dots,T-1$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{P}(\theta^t)^2\right] \le \frac{1}{c_{\min}^2 T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla\Phi(\theta^t)\|^2\right]. \tag{39}$$

Therefore,

$$\min_{0\le t<T}\mathbb{E}\mathcal{P}(\theta^t)^2 \le \frac{1}{c_{\min}^2}\min_{0\le t<T}\mathbb{E}\|\nabla\Phi(\theta^t)\|^2.$$

Thus, Theorem A.22 also provides a quantitative bound on the Pareto-stationarity residual. The proof is given in Appendix A.12.8.

Corollary A.26 (Single-objective specialization)**.** *When $m = 1$, the FedHV weight satisfies*

$$w_1^t = 1$$

at every communication round, independently of the objective-value estimate. Consequently, FedHV coincides with single-objective local SGD applied to f_1.

Moreover,

$$\nabla\Phi(\theta) = \frac{1}{r_1 - f_1(\theta)}\nabla f_1(\theta),$$

and, under Assumption A.4 and nonnegative losses,

$$\|\nabla f_1(\theta)\| \le r_1\|\nabla\Phi(\theta)\|.$$

A direct specialization of the proof of Theorem A.22 to $m = 1$ eliminates the weight-estimation term, since $w_1^t \equiv 1$, and yields a standard local-SGD stationarity bound.

A.11 Reference-Dependent Nash-Welfare Surrogate and Local Proportional-Fairness Interpretation

The log-hypervolume objective also provides a task-balancing interpretation. Recall that

$$s_i(\theta) = r_i - f_i(\theta) > 0$$

is the reference-relative slack of objective i. Minimizing the negative log-hypervolume is equivalent to maximizing

$$\Psi(\theta) := -\Phi(\theta) = \sum_{i=1}^{m} \log s_i(\theta) = \log\left(\prod_{i=1}^{m} s_i(\theta)\right). \tag{40}$$

Equation (40) is algebraically a log-product criterion over the reference-relative task slacks (Nash, 1950; Navon et al., 2022). We therefore regard Ψ as a *reference-dependent Nash-welfare surrogate in slack space*, rather than as a claim that FedHV solves a classical global Nash bargaining problem. In particular, the attainable slack set induced by a neural network is generally nonconvex, so global Nash-bargaining or global proportional-fairness conclusions do not follow from our stationarity analysis.

The marginal utility of objective i under this surrogate is $1/s_i(\theta)$, so an objective with a smaller reference-relative slack receives greater local importance. This motivates a *local first-order* connection to proportional fairness (Kelly, 1997). The interpretation is reference dependent and concerns balance among task objectives; it should not be interpreted as a guarantee of client-level, demographic, or societal fairness. For context, the classical global proportional-fairness definition and the local first-order implication used here are given in Appendix A.12.10.

To connect this interpretation to the convergence analysis, let Θ denote the feasible parameter set and let $T_\Theta(\theta)$ be its tangent cone at θ. In our unconstrained setting, $\Theta = \mathbb{R}^d$. We use

$$\Theta_+ := \{\theta \in \Theta : r_i - f_i(\theta) > 0,\ \forall i \in [m]\}$$

for the domain on which the reference-dependent Nash-welfare surrogate is defined.

Definition A.27 (Local proportional-fairness violation)**.** For $\theta \in \Theta_+$, define

$$\mathrm{PFgap}(\theta) := \sup_{\substack{d \in T_\Theta(\theta) \\ \|d\| \le 1}} \langle \nabla\Psi(\theta), d\rangle. \tag{41}$$

This quantity measures the largest first-order increase in aggregate relative slack over feasible unit-norm directions.

Lemma A.28 (Stationarity controls local proportional-fairness violation)**.** *For every $\theta \in \Theta_+$,*

$$\mathrm{PFgap}(\theta) \le \|\nabla\Psi(\theta)\| = \|\nabla\Phi(\theta)\|. \tag{42}$$

If $\Theta = \mathbb{R}^d$, then equality holds. The proof is given in Appendix A.12.10.

Corollary A.29 (FedHV controls local proportional-fairness violation)**.** *Under the fixed-horizon condition in Assumption A.4, let*

$$t^\star \in \arg\min_{0 \le t < T} \mathbb{E}\left[\|\nabla\Phi(\theta^t)\|^2\right].$$

Then

$$\mathbb{E}\left[\mathrm{PFgap}(\theta^{t^\star})\right] \le \left(\min_{0 \le t < T} \mathbb{E}\left[\|\nabla\Phi(\theta^t)\|^2\right]\right)^{1/2}. \tag{43}$$

Consequently, Theorem A.22 directly yields a quantitative bound on the largest feasible local first-order improvement in aggregate relative slack; it is not a global proportional-fairness guarantee. The proof is given in Appendix A.12.10.

Therefore, near-stationary FedHV iterates have a small local proportional-fairness violation in task-slack space, in the first-order sense quantified by PFgap: no feasible unit-norm direction can substantially increase the reference-dependent logarithmic Nash-welfare surrogate. This is a local stationarity-based interpretation, not a claim of a classical global Nash bargaining solution. The inverse-slack weights provide the corresponding optimization mechanism, since objectives closer to the reference boundary have larger coefficients $1/s_i(\theta)$.

A.12 Proofs and Auxiliary Lemmas

A.12.1 Proof of Lemma A.10

Proof. Let $s_i(\theta) = r_i - f_i(\theta)$ and define $u_i(\theta) = s_i(\theta)^{-1}$ wherever the slacks are positive. In particular, assume $s_i(\theta^\star) > 0$ for every i, so that $u_i(\theta^\star) > 0$. By direct differentiation of $\Phi(\theta) = -\sum_{i=1}^m \log s_i(\theta)$,

$$\begin{aligned} \nabla\Phi(\theta) &= \sum_{i=1}^m \frac{1}{s_i(\theta)} \nabla f_i(\theta) \\ &= \sum_{i=1}^m u_i(\theta) \nabla f_i(\theta). \end{aligned} \tag{44}$$

Evaluate at $\theta^\star$ and suppose $\nabla\Phi(\theta^\star) = \mathbf{0}$. Set

$$\lambda_i = \frac{u_i(\theta^\star)}{\sum_{j=1}^m u_j(\theta^\star)} \quad \text{so that} \quad \lambda_i > 0, \ \sum_{i=1}^m \lambda_i = 1. \tag{45}$$

Then

$$\begin{aligned} \sum_{i=1}^m \lambda_i \nabla f_i(\theta^\star) &= \frac{1}{\sum_{j=1}^m u_j(\theta^\star)} \sum_{i=1}^m u_i(\theta^\star) \nabla f_i(\theta^\star) \\ &= \frac{1}{\sum_{j=1}^m u_j(\theta^\star)} \nabla\Phi(\theta^\star) = \mathbf{0}. \end{aligned} \tag{46}$$

Hence $\mathbf{0}$ lies in the convex hull of $\{\nabla f_i(\theta^\star)\}_{i=1}^m$, i.e., $\theta^\star$ satisfies the first-order Pareto stationarity condition. □

A.12.2 Proof of Lemma A.6

Proof. Because the local objectives are nonnegative, their uniform average f_i is nonnegative. For $L_i > 0$, the descent lemma for an L_i-smooth function, applied at $\theta - (1/L_i)\nabla f_i(\theta)$, gives

$$f_i\left(\theta - \frac{1}{L_i}\nabla f_i(\theta)\right) \le f_i(\theta) - \frac{1}{2L_i}\|\nabla f_i(\theta)\|^2.$$

The left-hand side is nonnegative, so $\|\nabla f_i(\theta)\|^2 \le 2L_i f_i(\theta)$. If $L_i = 0$, smoothness and nonnegativity imply that f_i is constant and the same inequality holds trivially. On $\mathcal{D}_\rho$, the definition of the safe subset and nonnegativity imply $f_i(\theta) \le r_i - \rho$, which proves the stated bound. Finally, Jensen's inequality gives

$$\left\|\sum_i w_i \nabla f_i(\theta)\right\| \le \sum_i w_i G_i \le G_{\max}.$$

□

A.12.3 Proof of Lemma A.8

Proof. Let $\theta, \theta' \in \mathcal{D}$ and assume that $[\theta, \theta'] \subseteq \mathcal{D}_\rho$. Write $u_i(\theta) = 1/(r_i - f_i(\theta))$. Since the entire segment is in $\mathcal{D}_\rho$, the definition of the safe subset gives $u_i(\theta) \le 1/\rho$ and $u_i(\theta') \le 1/\rho$. Therefore,

$$\begin{aligned} \|\nabla\Phi(\theta) - \nabla\Phi(\theta')\| \le &\sum_{i=1}^m u_i(\theta)\|\nabla f_i(\theta) - \nabla f_i(\theta')\| \\ &+ \sum_{i=1}^m |u_i(\theta) - u_i(\theta')|\|\nabla f_i(\theta')\|. \end{aligned} \tag{47}$$

The first term is bounded by $\sum_i L_i\|\theta - \theta'\|/\rho$. Moreover, because $x \mapsto 1/x$ is $1/\rho^2$-Lipschitz on $[\rho, \infty)$,

$$|u_i(\theta) - u_i(\theta')| \le \frac{|f_i(\theta) - f_i(\theta')|}{\rho^2} \le \frac{G_i}{\rho^2}\|\theta - \theta'\|,$$

where the last inequality follows from Lemma A.6. Applying the same lemma to $\|\nabla f_i(\theta')\|$ yields

$$\|\nabla\Phi(\theta) - \nabla\Phi(\theta')\| \le \left(\sum_{i=1}^{m} \frac{L_i}{\rho} + \sum_{i=1}^{m} \frac{G_i^2}{\rho^2}\right) \|\theta - \theta'\|.$$

Thus Φ is L_Φ-smooth without requiring twice differentiability. □

A.12.4 Proof of Lemma A.17

Proof. Write $a_i = a_i(z) := \max\{r_i - z_i, \rho_{\text{clip}}\}$, $u_i := a_i^{-1}$, and $U := \sum_j u_j$. For the smooth map $V_i(a) = u_i/U$, direct differentiation gives

$$\frac{\partial V_i}{\partial a_j} = -\frac{\mathbf{1}\{i=j\}u_i^2}{U} + \frac{u_i u_j^2}{U^2}.$$

Using $\sum_i u_i = U$ and $\sum_i u_i^2 \le U/a_{\min}$,

$$\sum_{i,j} \left|\frac{\partial V_i}{\partial a_j}\right| \le \frac{\sum_i u_i^2}{U} + \frac{(\sum_i u_i)(\sum_j u_j^2)}{U^2} \le \frac{2}{a_{\min}}.$$

Differentiating once more yields

$$\frac{\partial^2 V_i}{\partial a_j \partial a_k} = \frac{2\mathbf{1}\{i=j=k\}u_i^3}{U} - \frac{\mathbf{1}\{i=j\}u_i^2 u_k^2}{U^2} - \frac{\mathbf{1}\{i=k\}u_i^2 u_j^2}{U^2} - \frac{2\mathbf{1}\{j=k\}u_i u_j^3}{U^2} + \frac{2u_i u_j^2 u_k^2}{U^3}.$$

Since $\sum_i u_i^3 \le U/a_{\min}^2$, summing the absolute values of these five groups gives respectively $2, 1, 1, 2, 2$ times $a_{\min}^{-2}$. This proves (27).

The clipped-slack map $z \mapsto a(z)$ is 1-Lipschitz in ℓ_∞. Every point on the segment between $a(z)$ and $a(z')$ has minimum coordinate at least ρ_{clip}. Applying the mean-value inequality to V on this segment and using the first bound above proves (28). If clipping is inactive on the segment $[z, z']$ and all its slacks are at least $\widetilde{\rho}$, then $a(z) = \boldsymbol{r} - z$ is affine there and the same argument gives the local constant $2/\widetilde{\rho}$.

Under the strict inequality $\widetilde{\rho} > \rho_{\text{clip}}$, the map is twice differentiable near that segment. Applying scalar Taylor's theorem to each coordinate along $z' + \lambda(z - z')$, summing absolute values, and using the second derivative bound in (27) gives

$$\frac{1}{2}\frac{8}{\widetilde{\rho}^2}\|z - z'\|_\infty^2,$$

which is (29); its Jacobian bound follows from the first derivative estimate. □

A.12.5 Proof of Lemma A.19

Proof. Fix i, k, t, condition on $\mathcal{F}_t, S_t$, and write

$$y_j := \ell_i(\theta^t; x_{k,j}), \qquad \bar{y} := f_i^{(k)}(\theta^t), \qquad a_j := y_j - \bar{y}.$$

Then $\sum_{j=1}^{n_k} a_j = 0$. Let $I_j = \mathbf{1}\{j \in \mathcal{B}_{k,t}^{\text{rep}}\}$. For a uniform subset of size $b_{k,t}^{\text{rep}}$, when $n_k > 1$,

$$\mathbb{E}[I_j \mid \mathcal{F}_t, S_t] = \frac{b_{k,t}^{\text{rep}}}{n_k}, \qquad \mathbb{E}[I_j I_\ell \mid \mathcal{F}_t, S_t] = \frac{b_{k,t}^{\text{rep}}(b_{k,t}^{\text{rep}} - 1)}{n_k(n_k - 1)} \quad (j \ne \ell).$$

Therefore,

$$\widehat{f}_{i,k}^t - \bar{y} = \frac{1}{b_{k,t}^{\text{rep}}} \sum_{j=1}^{n_k} I_j a_j,$$

which gives

$$\mathbb{E}\left[\widehat{f}_{i,k}^t - \bar{y} \,\middle|\, \mathcal{F}_t, S_t\right] = 0.$$

Using $\sum_{j\neq\ell} a_j a_\ell = -\sum_j a_j^2$, its conditional second moment is

$$\mathbb{E}\left[(\widehat{f}_{i,k}^t - \bar{y})^2 \middle| \mathcal{F}_t, S_t\right] = \frac{1}{(b_{k,t}^{\mathrm{rep}})^2}\left[\frac{b_{k,t}^{\mathrm{rep}}}{n_k} - \frac{b_{k,t}^{\mathrm{rep}}(b_{k,t}^{\mathrm{rep}}-1)}{n_k(n_k-1)}\right]\sum_{j=1}^{n_k} a_j^2$$
$$= \frac{n_k - b_{k,t}^{\mathrm{rep}}}{b_{k,t}^{\mathrm{rep}}(n_k-1)} v_{i,k}(\theta^t). \tag{48}$$

Since $\nu_{i,k}^t = \widehat{f}_{i,k}^t - \bar{y}$, this proves the lemma. If $n_k = 1$, necessarily $b_{k,t}^{\mathrm{rep}} = 1$ and the report is exact. □

A.12.6 Proof of Lemma A.20

Proof. Fix an objective i and condition first on $\mathcal{F}_t$. Write

$$a_{i,k}^t := f_i^{(k)}(\theta^t) - f_i(\theta^t).$$

Then $\sum_{k=1}^N a_{i,k}^t = 0$, and

$$\widehat{f}_i^t - f_i(\theta^t) = \frac{1}{M}\sum_{k\in S_t} a_{i,k}^t + \frac{1}{M}\sum_{k\in S_t}\nu_{i,k}^t.$$

The first term is the mean of a finite population sampled without replacement. Hence,

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{k\in S_t} a_{i,k}^t\right)^2 \middle| \mathcal{F}_t\right] = \frac{N-M}{M(N-1)}\frac{1}{N}\sum_{k=1}^N (a_{i,k}^t)^2.$$

Define

$$R_i^t := \frac{1}{M}\sum_{k\in S_t}\nu_{i,k}^t.$$

By Lemma A.19, the conditional report means are zero. Therefore the cross term between the client-sampling term and R_i^t vanishes after conditioning on $\mathcal{F}_t, S_t$. Under the optional cross-client orthogonality condition,

$$\mathbb{E}\left[(R_i^t)^2 \middle| \mathcal{F}_t, S_t\right] = \frac{1}{M^2}\sum_{k\in S_t}\gamma_{k,t} v_{i,k}(\theta^t).$$

Taking expectations over the uniformly sampled S_t and using the definition of $\tau_{i,T}^2$ gives

$$\mathbb{E}[(R_i^t)^2] \le \frac{\tau_{i,T}^2}{M}.$$

Without cross-client orthogonality, Jensen's inequality instead gives

$$\mathbb{E}[(R_i^t)^2] \le \mathbb{E}\left[\frac{1}{M}\sum_{k\in S_t}\gamma_{k,t} v_{i,k}(\theta^t)\right] \le \tau_{i,T}^2.$$

Combining these bounds with the finite-population client-sampling identity, summing over i, and using $\|\boldsymbol{z}\|_\infty^2 \le \|\boldsymbol{z}\|_2^2$ proves the claim. The local model update does not enter this calculation; hence no independence between the report and that update is used. □

A.12.7 Proof of Lemma A.21

Proof. For $1 \le t < T_\rho$,

$$\boldsymbol{w}^t = W_{\rho_{\mathrm{clip}}}\left(\widehat{\boldsymbol{F}}^{t-1}; \boldsymbol{r}\right), \qquad \boldsymbol{w}^{\star,t} = W_{\rho_{\mathrm{clip}}}\left(\boldsymbol{F}(\theta^t); \boldsymbol{r}\right).$$

Adding and subtracting $W_{\rho_{\text{clip}}}(\boldsymbol{F}(\theta^{t-1}); \boldsymbol{r})$ and applying the triangle inequality gives

$$\begin{aligned}\left\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\right\|_1 \leq & \left\|W_{\rho_{\text{clip}}}\left(\widehat{\boldsymbol{F}}^{\,t-1}; \boldsymbol{r}\right) - W_{\rho_{\text{clip}}}\left(\boldsymbol{F}(\theta^{t-1}); \boldsymbol{r}\right)\right\|_1 \\ & + \left\|W_{\rho_{\text{clip}}}\left(\boldsymbol{F}(\theta^{t-1}); \boldsymbol{r}\right) - W_{\rho_{\text{clip}}}\left(\boldsymbol{F}(\theta^{t}); \boldsymbol{r}\right)\right\|_1. \end{aligned} \tag{49}$$

Applying Lemma A.17 to both terms yields

$$\begin{aligned}\left\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\right\|_1 \leq & C(\rho_{\text{clip}}, \boldsymbol{r}) \left\|\widehat{\boldsymbol{F}}^{\,t-1} - \boldsymbol{F}(\theta^{t-1})\right\|_\infty \\ & + C(\rho_{\text{clip}}, \boldsymbol{r}) \left\|\boldsymbol{F}(\theta^{t-1}) - \boldsymbol{F}(\theta^{t})\right\|_\infty. \end{aligned} \tag{50}$$

Using

$$(a+b)^2 \leq 2a^2 + 2b^2$$

and Lemma A.20 gives

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\right\|_1^2\right] \leq & 2C(\rho_{\text{clip}}, \boldsymbol{r})^2 \Xi_{\text{loss},T} \\ & + 2C(\rho_{\text{clip}}, \boldsymbol{r})^2 \mathbb{E}\left[\left\|\boldsymbol{F}(\theta^t) - \boldsymbol{F}(\theta^{t-1})\right\|_\infty^2\right]. \end{aligned} \tag{51}$$

Finally, for $t = 0$, $\boldsymbol{w}^0, \boldsymbol{w}^{\star,0} \in \Delta_m$, so

$$\left\|\boldsymbol{w}^0 - \boldsymbol{w}^{\star,0}\right\|_1 \leq 2.$$

□

A.12.8 Proof of Corollary A.25

The inverse-slack weights

$$w_i^\star(\theta) = \frac{\left(r_i - f_i(\theta)\right)^{-1}}{c(\theta)}, \qquad c(\theta) = \sum_{j=1}^m \frac{1}{r_j - f_j(\theta)},$$

belong to Δ_m. Hence, they form a feasible choice in the minimization defining $\mathcal{P}(\theta)$, and therefore

$$\begin{aligned}\mathcal{P}(\theta) &\leq \left\|\sum_{i=1}^m w_i^\star(\theta) \nabla f_i(\theta)\right\| \\ &= \frac{1}{c(\theta)} \left\|\nabla \Phi(\theta)\right\|. \end{aligned} \tag{52}$$

By Assumption A.4 and the nonnegativity of the objectives,

$$r_i - f_i(\theta) \leq r_i,$$

and consequently

$$c(\theta) = \sum_{i=1}^m \frac{1}{r_i - f_i(\theta)} \geq \sum_{i=1}^m \frac{1}{r_i} = c_{\min}.$$

Thus,

$$\mathcal{P}(\theta) \leq \frac{\|\nabla \Phi(\theta)\|}{c(\theta)} \leq \frac{\|\nabla \Phi(\theta)\|}{c_{\min}}.$$

Squaring this inequality, taking expectations, and summing over $t = 0, \ldots, T-1$ proves (39). Averaging followed by the minimum over $0 \leq t < T$ gives its second displayed consequence. □

A.12.9 Proof of Theorem A.22

We prove Theorem A.22 by following the standard bias–variance–drift analysis of federated local SGD, adapted to the log–hypervolume objective Φ and hypervolume weights. Because the theorem uses the Jensen-valid choice $\kappa_M = 1$, we instantiate the client-averaged noise bound with $\kappa_M = 1$ throughout this proof. The symbol κ_M is retained only in the separate auxiliary statements that record the optional orthogonality refinement. By Assumption A.4, every point used in the following argument lies in $\mathcal{D}_\rho$ for each $t = 0, \ldots, T-1$. Thus all one-round inequalities below are applied directly to the algorithmic iterates.

Per-round notation and ideal colinearity. Fix a communication round t. Recall the true slack, inverse slack, and normalizing factor

$$s_i^t = r_i - f_i(\theta^t), \qquad u_i^t = (s_i^t)^{-1}, \qquad c^t = \sum_{j=1}^m u_j^t.$$

The ideal current hypervolume weights are

$$w_i^{\star,t} = \frac{u_i^t}{c^t}, \qquad \boldsymbol{w}^{\star,t} \in \Delta_m,$$

and the corresponding global scalarized objective is

$$L_{\boldsymbol{w}^{\star,t}}(\theta) = \sum_{i=1}^m w_i^{\star,t} f_i(\theta).$$

In contrast, the actual weight vector used by the algorithm in round t is the one-round-stale vector $\boldsymbol{w}^t$, which is $\mathcal{F}_t$-measurable and fixed before the participating client set S_t is sampled.

By direct differentiation of

$$\Phi(\theta) = \sum_{i=1}^m -\log\big(r_i - f_i(\theta)\big),$$

we have

$$\nabla\Phi(\theta) = \sum_{i=1}^m u_i(\theta)\nabla f_i(\theta).$$

Hence, at θ^t,

$$\nabla\Phi(\theta^t) = c^t \nabla L_{\boldsymbol{w}^{\star,t}}(\theta^t). \tag{53}$$

Thus, the log-hypervolume gradient is colinear with the gradient of the scalarized objective formed using the *ideal current* hypervolume weights $\boldsymbol{w}^{\star,t}$, not necessarily with the stale weights $\boldsymbol{w}^t$ used by the algorithm.

Bounds on the scaling c^t. Under the safe–reference assumption (Assumption A.4),

$$s_i^t \;=\; r_i - f_i(\theta^t) \;\geq\; \rho > 0 \quad \Rightarrow \quad u_i^t \;\leq\; \frac{1}{\rho}.$$

Assuming nonnegative losses $f_i(\theta) \geq 0$ along the trajectory, we also have $s_i^t \leq r_i$ and hence

$$u_i^t \;=\; \frac{1}{s_i^t} \;\geq\; \frac{1}{r_i}.$$

Summing over i,

$$c_{\min} \;\leq\; c^t \;\leq\; c_{\max}, \qquad c_{\min} \;:=\; \sum_{i=1}^m \frac{1}{r_i} \;\geq\; \frac{m}{\max_i r_i}, \qquad c_{\max} \;:=\; \frac{m}{\rho}. \tag{54}$$

Combining (53) and (54) gives

$$\left\|\nabla L_{\boldsymbol{w}^{\star,t}}(\theta^t)\right\|^2 = \frac{1}{(c^t)^2}\left\|\nabla\Phi(\theta^t)\right\|^2 \geq \frac{1}{c_{\max}^2}\left\|\nabla\Phi(\theta^t)\right\|^2, \tag{55}$$

and similarly

$$\left\|\nabla L_{\boldsymbol{w}^{\star,t}}(\theta^t)\right\|^2 \leq \frac{1}{c_{\min}^2}\left\|\nabla\Phi(\theta^t)\right\|^2.$$

Local objectives and stochastic gradients. For each client k, define the actual stale-weight local objective used in round t by

$$L_{\boldsymbol{w}^t}^{(k)}(\theta) = \sum_{i=1}^{m} w_i^t f_i^{(k)}(\theta).$$

The corresponding global scalarized objective is

$$L_{\boldsymbol{w}^t}(\theta) = \frac{1}{N}\sum_{k=1}^{N} L_{\boldsymbol{w}^t}^{(k)}(\theta) = \sum_{i=1}^{m} w_i^t f_i(\theta).$$

All local trajectories in the analysis below are the actual trajectories generated with the stale weight vector $\boldsymbol{w}^t$. The ideal weight vector $\boldsymbol{w}^{\star,t}$ is used only to compare the actual direction with the current log-hypervolume direction through the error e_t.

Because $\boldsymbol{w}^t \in \Delta_m$, Assumption A.2 implies that every $L_{\boldsymbol{w}^t}^{(k)}$ is L_w-smooth, where we may use the uniform constant

$$L_w := \sum_{i=1}^{m} L_i.$$

For fixed round t, write $\theta_k^s := \theta_{k,s}^t$ for the actual local iterate and let

$$g_{i,k}^s := g_i^{(k)}\big(\theta_k^s; B_{k,s}^t\big),$$

and define the weighted local stochastic gradient

$$g_k^s := \sum_{i=1}^{m} w_i^t g_{i,k}^s.$$

Each participating client $k \in S_t$ performs

$$\theta_k^{s+1} = \theta_k^s - \eta_\ell g_k^s, \qquad \theta_k^0 = \theta^t, \qquad s = 0, \dots, K-1.$$

Since $\boldsymbol{w}^t$ is $\mathcal{F}_t$-measurable, the component-wise oracle conditions in Assumption A.2 and the definition of $\mathcal{G}_{t,s}$ give

$$\mathbb{E}\left[g_k^s \mid \mathcal{G}_{t,s}\right] = \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s).$$

Moreover, using $w_i^t \geq 0$ and $\sum_i w_i^t = 1$, we may use the uniform bound

$$\mathbb{E}\left[\left\|g_k^s - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s)\right\|^2 \mid \mathcal{G}_{t,s}\right] \leq \Sigma^2, \qquad \Sigma^2 := \sum_{i=1}^{m} \sigma_i^2.$$

The local model update is

$$\Delta\theta_k^t = \theta_k^K - \theta^t = -\eta_\ell \sum_{s=0}^{K-1} g_k^s,$$

and the server forms the averaged update

$$\Delta^t := \frac{1}{|S_t|}\sum_{k\in S_t} \Delta\theta_k^t, \qquad \theta^{t+1} = \theta^t + \eta_g \Delta^t.$$

Bias–variance–drift decomposition. Condition on the round history $\mathcal{F}_t$. By construction, θ^t and $\boldsymbol{w}^t$ are then fixed, while S_t is sampled uniformly without replacement from the N clients.

Using the tower rule and the conditional unbiasedness of the minibatch gradients,

$$\mathbb{E}\left[\Delta^t \mid \mathcal{F}_t\right] = -\eta_\ell \sum_{s=0}^{K-1} \mathbb{E}\left[\frac{1}{|S_t|}\sum_{k\in S_t} \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s)\,\middle|\, \mathcal{F}_t\right]. \tag{56}$$

Add and subtract $\nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^t)$ inside the sum to obtain

$$\mathbb{E}\left[\Delta^t \mid \mathcal{F}_t\right] = -\eta_\ell K \mathbb{E}\left[\frac{1}{|S_t|}\sum_{k\in S_t} \nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^t)\middle| \mathcal{F}_t\right] + b_t, \tag{57}$$

where

$$b_t := -\eta_\ell \sum_{s=0}^{K-1} \mathbb{E}\left[\frac{1}{|S_t|}\sum_{k\in S_t}\left(\nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^s_k) - \nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^t)\right)\middle| \mathcal{F}_t\right]. \tag{58}$$

Because $\boldsymbol{w}^t$ and θ^t are fixed conditional on $\mathcal{F}_t$, uniform sampling without replacement gives

$$\mathbb{E}\left[\frac{1}{|S_t|}\sum_{k\in S_t} \nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^t)\middle| \mathcal{F}_t\right] = \frac{1}{N}\sum_{k=1}^{N} \nabla L^{(k)}_{\boldsymbol{w}^t}(\theta^t) = \nabla L_{\boldsymbol{w}^t}(\theta^t). \tag{59}$$

Lemma A.30 (Virtual-trajectory deviation)**.** *Under Assumptions A.1, A.2, A.3, A.4, and A.9, fix a communication round $t \in \{0,\dots,T-1\}$ and condition on $\mathcal{F}_t$. Define the virtual trajectory*

$$v_{t,0} := \theta^t, \qquad v_{t,s+1} := v_{t,s} - \eta_\ell \nabla L_{\boldsymbol{w}^t}(v_{t,s}), \qquad s = 0,\dots,K-1. \tag{60}$$

The virtual trajectory is $\mathcal{F}_t$-measurable and uses the same frozen weight vector as the actual local trajectories. Define

$$\zeta^2 := \sum_{i=1}^{m} \zeta_i^2, \qquad Q := \zeta^2 + \Sigma^2, \qquad C_\delta := 4, \qquad \bar{G}^2 := 2\big(G_{\max}^2 + C_\delta Q\big).$$

By Assumption A.4, for every $k \in S_t$ and $0 \le s \le K$,

$$\theta^t_{k,s} \in \mathcal{D}_\rho \subseteq \mathcal{D}, \qquad v_{t,s} \in \mathcal{D}_\rho \subseteq \mathcal{D},$$

and the connecting segments required by the proof are contained in $\mathcal{D}_\rho$. Consequently, the heterogeneity condition is available at every virtual iterate, and the actual local trajectories remain inside the analysis domain. The local stochastic gradients are those generated by the fresh with-replacement minibatch oracle in Assumption A.2; this is the oracle used at every actual local iterate in the statement.

If $\eta_\ell L_w K \le 1/4$, then, for every $0 \le s \le K$,

$$D_{t,s} := \mathbb{E}\left[\frac{1}{M}\sum_{k\in S_t} \left\|\theta^t_{k,s} - v_{t,s}\right\|^2 \middle| \mathcal{F}_t\right] \le C_\delta \eta_\ell^2 s^2 Q. \tag{61}$$

Proof. For brevity write

$$h_k(\theta) := \nabla L^{(k)}_{\boldsymbol{w}^t}(\theta), \qquad h(\theta) := \nabla L_{\boldsymbol{w}^t}(\theta), \qquad \delta_{k,s} := \theta^t_{k,s} - v_{t,s}.$$

Also set

$$a_{k,s} := h_k(v_{t,s}) - h(v_{t,s}), \qquad q_{k,s} := h_k(\theta^t_{k,s}) - h_k(v_{t,s}).$$

Then

$$\delta_{k,s+1} = \delta_{k,s} - \eta_\ell\big(q_{k,s} + a_{k,s} + \xi^t_{k,s}\big), \qquad \delta_{k,0} = 0.$$

By the fresh-minibatch oracle and the filtration $\mathcal{G}_{t,s}$,

$$\mathbb{E}[\xi^t_{k,s} \mid \mathcal{G}_{t,s}] = 0, \qquad \mathbb{E}[\|\xi^t_{k,s}\|^2 \mid \mathcal{G}_{t,s}] \le \Sigma^2.$$

Thus the local stochastic errors form a martingale-difference sequence with respect to the within-round filtration. The argument below uses only this conditional first-moment property and the conditional second-moment bound, together with Cauchy–Schwarz; it never factors products of noises from different local steps and does not assume independence across local steps. The filtration already contains the objective-report randomness and all earlier minibatches.

Using the recursion, Cauchy–Schwarz, and $\|x+y+z\|^2 \le 3(\|x\|^2 + \|y\|^2 + \|z\|^2)$, we obtain

$$D_{t,s} \le 3\eta_\ell^2 s \sum_{\tau=0}^{s-1} \left(L_w^2 D_{t,\tau} + \zeta^2 + \Sigma^2\right).$$

Here we used L_w-smoothness to bound $\|q_{k,\tau}\| \le L_w\|\delta_{k,\tau}\|$, the uniform sampling identity, and the fact that $v_{t,\tau} \in \mathcal{D}$ on the fixed-horizon trajectory. More explicitly, Jensen's inequality on the simplex and Assumption A.3 give

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\|a_{k,\tau}\|^2 &= \frac{1}{N}\sum_{k=1}^{N}\left\|\sum_{i=1}^{m} w_i^t\left(\nabla f_i^{(k)}(v_{t,\tau}) - \nabla f_i(v_{t,\tau})\right)\right\|^2 \\ &\le \sum_{i=1}^{m} w_i^t \frac{1}{N}\sum_{k=1}^{N}\left\|\nabla f_i^{(k)}(v_{t,\tau}) - \nabla f_i(v_{t,\tau})\right\|^2 \\ &\le \sum_{i=1}^{m} w_i^t \zeta_i^2 \le \zeta^2.\end{aligned}$$

Therefore, uniform client sampling gives

$$\mathbb{E}\left[\frac{1}{M}\sum_{k\in S_t}\|a_{k,\tau}\|^2\,\middle|\,\mathcal{F}_t\right] = \frac{1}{N}\sum_{k=1}^{N}\|a_{k,\tau}\|^2 \le \zeta^2,$$

and the tower property applied to the martingale-difference oracle gives the corresponding Σ^2 noise bound.

We prove (61) by induction. If $D_{t,\tau} \le 4\eta_\ell^2\tau^2 Q$ for $\tau < s$, then

$$D_{t,s} \le 4\eta_\ell^4 L_w^2 s^4 Q + 3\eta_\ell^2 s^2 Q \le 4\eta_\ell^2 s^2 Q,$$

where the last inequality follows from $\eta_\ell L_w s \le 1/4$. The case $s = 0$ is immediate. □

Lemma A.31 (Local-drift bias bound)**.** *Under Assumptions A.1, A.2, A.4, A.3, and A.9, and the virtual-trajectory bound in Lemma A.30, the conditional-mean bias b_t defined in* (58) *for every round $t = 0, \dots, T-1$ satisfies*

$$\|b_t\| \le \frac{L_w\eta_\ell^2\bar{G}}{2}K(K-1), \tag{62}$$

where

$$\bar{G} := \sqrt{\bar{G}^2}.$$

Consequently,

$$\|b_t\|^2 \le \frac{L_w^2\eta_\ell^4\bar{G}^2}{4}K^2(K-1)^2. \tag{63}$$

Proof. Let

$$d_{k,s} := \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_{k,s}^t) - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t), \qquad D_s^{\mathrm{drift}} := \frac{1}{M}\sum_{k\in S_t} d_{k,s}.$$

Using the virtual trajectory and L_w-smoothness,

$$\|d_{k,s}\| \le L_w\|\theta_{k,s}^t - v_{t,s}\| + L_w\|v_{t,s} - \theta^t\|.$$

Since $\|\nabla L_{\boldsymbol{w}^t}(v_{t,u})\| \le G_{\max}$,

$$\|v_{t,s} - \theta^t\| \le \eta_\ell s G_{\max}.$$

Jensen's inequality and (61) therefore give

$$\mathbb{E}\left[\,\|D_s^{\mathrm{drift}}\|\,\middle|\,\mathcal{F}_t\right] \le L_w\eta_\ell s\left(\sqrt{C_\delta Q} + G_{\max}\right) \le L_w\eta_\ell s\bar{G}.$$

Substituting this estimate into the definition of b_t and using $\sum_{s=0}^{K-1} s = K(K-1)/2$ proves (62). Squaring gives (63). □

Lemma A.32 (Second moment of the averaged update)**.** *Assume Assumptions A.1, A.2, A.4, A.3, and A.9. For every round $t = 0, \dots, T-1$, condition on the round history $\mathcal{F}_t$. The weight vector $\boldsymbol{w}^t$ and the global model θ^t are then fixed, and S_t is sampled uniformly without replacement with*

$$|S_t| = M = pN.$$

Each selected client $k \in S_t$ performs K local SGD steps

$$\theta_k^{s+1} = \theta_k^s - \eta_\ell g_k^s, \qquad \theta_k^0 = \theta^t.$$

Define

$$\Delta\theta_k^t := \theta_k^K - \theta^t, \qquad \Delta^t := \frac{1}{M}\sum_{k\in S_t} \Delta\theta_k^t.$$

Then, conditionally on $\mathcal{F}_t$,

$$\begin{aligned}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right] \le{}& C_0\eta_\ell^2 K^2 \left\|\nabla L_{\boldsymbol{w}^t}(\theta^t)\right\|^2 + C_1\eta_\ell^2 K \kappa_M \Sigma^2 \\ &+ C_2\eta_\ell^2 K^2 \mathcal{H}_N(p;\zeta) \\ &+ C_B\eta_\ell^4 K^2 (K-1)(2K-1),\end{aligned} \tag{64}$$

where

$$\Sigma^2 := \sum_{i=1}^m \sigma_i^2, \qquad C_B := \frac{L_w^2 \bar{G}^2}{2},$$

and $C_0, C_1, C_2 > 0$ are numerical constants independent of $t, T, N, p, \eta_\ell, \eta_g$, and K. As before, one may take

$$C_0 = 6, \qquad C_1 = 3, \qquad C_2 = 6.$$

Proof. Throughout the proof, we condition on the history of the round $\mathcal{F}_t$. Hence, θ^t and the stale weight vector $\boldsymbol{w}^t$ are fixed, while the remaining randomness is due to the current client sample S_t and the current-round minibatches.

Recall

$$\Delta^t = -\eta_\ell \frac{1}{M}\sum_{k\in S_t}\sum_{s=0}^{K-1} g_k^s.$$

Decompose each stochastic gradient as

$$g_k^s = \underbrace{\nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t)}_{=:a_k} + \underbrace{\left(\nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s) - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t)\right)}_{=:d_k^s} + \underbrace{\left(g_k^s - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s)\right)}_{=:\xi_{k,s}^t}.$$

By Lemma A.7,

$$\mathbb{E}\left[\xi_{k,s}^t \mid \mathcal{G}_{t,s}\right] = 0, \qquad \mathbb{E}\left[\|\xi_{k,s}^t\|^2 \mid \mathcal{G}_{t,s}\right] \le \Sigma^2.$$

Define

$$A := \frac{1}{M}\sum_{k\in S_t}\sum_{s=0}^{K-1} a_k = K\frac{1}{M}\sum_{k\in S_t} a_k,$$

$$B := \frac{1}{M}\sum_{k\in S_t}\sum_{s=0}^{K-1} d_k^s, \qquad C := \frac{1}{M}\sum_{k\in S_t}\sum_{s=0}^{K-1} \xi_{k,s}^t.$$

Thus

$$\Delta^t = -\eta_\ell(A + B + C).$$

(i) Bounding the main client-sampling term A. Because θ^t and $\boldsymbol{w}^t$ are $\mathcal{F}_t$-measurable, the vectors

$$a_k = \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t), \qquad k = 1, \dots, N,$$

form a fixed finite population conditional on $\mathcal{F}_t$. Define its population mean by

$$\mu_t := \frac{1}{N} \sum_{k=1}^{N} a_k = \nabla L_{\boldsymbol{w}^t}(\theta^t).$$

Therefore,

$$\frac{1}{M} \sum_{k \in S_t} a_k = \mu_t + \frac{1}{M} \sum_{k \in S_t} (a_k - \mu_t).$$

Using $\|x + y\|^2 \le 2\|x\|^2 + 2\|y\|^2$,

$$\begin{aligned} \mathbb{E}\left[\|A\|^2 \mid \mathcal{F}_t\right] &\le 2K^2 \|\mu_t\|^2 \\ &\quad + 2K^2 \mathbb{E}\left[\left\| \frac{1}{M} \sum_{k \in S_t} (a_k - \mu_t) \right\|^2 \middle| \mathcal{F}_t \right]. \end{aligned} \tag{65}$$

To evaluate the second term, let

$$I_k := \mathbf{1}\{k \in S_t\}.$$

Since S_t is sampled uniformly without replacement and is independent of $\mathcal{F}_t$,

$$\mathbb{E}[I_k \mid \mathcal{F}_t] = \frac{M}{N},$$

and, for $k \neq \ell$,

$$\mathbb{E}[I_k I_\ell \mid \mathcal{F}_t] = \frac{M(M-1)}{N(N-1)}.$$

Since

$$\sum_{k=1}^{N} (a_k - \mu_t) = 0,$$

the standard finite-population variance identity gives

$$\begin{aligned} &\mathbb{E}\left[\left\| \frac{1}{M} \sum_{k \in S_t} (a_k - \mu_t) \right\|^2 \middle| \mathcal{F}_t \right] \\ &\quad = \frac{N - M}{M(N-1)} \frac{1}{N} \sum_{k=1}^{N} \|a_k - \mu_t\|^2. \end{aligned} \tag{66}$$

Moreover,

$$a_k - \mu_t = \sum_{i=1}^{m} w_i^t \left(\nabla f_i^{(k)}(\theta^t) - \nabla f_i(\theta^t) \right).$$

Because $\boldsymbol{w}^t \in \Delta_m$ and the squared norm is convex,

$$\|a_k - \mu_t\|^2 \le \sum_{i=1}^{m} w_i^t \left\| \nabla f_i^{(k)}(\theta^t) - \nabla f_i(\theta^t) \right\|^2.$$

Averaging over clients and using the global-iterate clause of the trajectory-wise condition in Assumption A.3,

$$\frac{1}{N} \sum_{k=1}^{N} \|a_k - \mu_t\|^2 \le \sum_{i=1}^{m} w_i^t \zeta_i^2 \le \sum_{i=1}^{m} \zeta_i^2.$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[\|A\|^2 \mid \mathcal{F}_t\right] &\le 2K^2 \left\| \nabla L_{\boldsymbol{w}^t}(\theta^t) \right\|^2 \\ &\quad + 2K^2 \mathcal{H}_N(p; \zeta). \end{aligned} \tag{67}$$

(ii) Stochastic-gradient noise term C. Define

$$H_s := \sum_{k \in S_t} \xi_{k,s}^t, \qquad C = \frac{1}{M} \sum_{s=0}^{K-1} H_s.$$

For a fixed local step s, the definition of κ_M gives

$$\mathbb{E}\left[\|H_s\|^2 \middle| \mathcal{G}_{t,s} \right] \le M^2 \kappa_M \Sigma^2.$$

For $s < u$, H_s is $\mathcal{G}_{t,u}$-measurable, while $\mathbb{E}[H_u \mid \mathcal{G}_{t,u}] = 0$. Therefore all cross terms between different local steps vanish by the tower property:

$$\mathbb{E}\langle H_s, H_u \rangle = \mathbb{E}\left[\langle H_s, \mathbb{E}[H_u \mid \mathcal{G}_{t,u}] \rangle \right] = 0.$$

Consequently,

$$\mathbb{E}\left[\|C\|^2 \middle| \mathcal{F}_t \right] \le \frac{1}{M^2} \sum_{s=0}^{K-1} \mathbb{E}\left[\|H_s\|^2 \middle| \mathcal{F}_t \right] \le \kappa_M K \Sigma^2.$$

(iii) Client-drift term B. For

$$d_k^s := \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s) - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t), \qquad D_s := \frac{1}{M} \sum_{k \in S_t} d_k^s,$$

use the virtual trajectory from Lemma A.30 to write

$$d_k^s = \left[\nabla L_{\boldsymbol{w}^t}^{(k)}(\theta_k^s) - \nabla L_{\boldsymbol{w}^t}^{(k)}(v_{t,s}) \right] + \left[\nabla L_{\boldsymbol{w}^t}^{(k)}(v_{t,s}) - \nabla L_{\boldsymbol{w}^t}^{(k)}(\theta^t) \right].$$

The L_w-smoothness of the local scalarized objectives implies

$$\|d_k^s\| \le L_w \|\theta_k^s - v_{t,s}\| + L_w \|v_{t,s} - \theta^t\|.$$

Moreover,

$$\|v_{t,s} - \theta^t\| \le \eta_\ell s G_{\max},$$

because $\|\nabla L_{\boldsymbol{w}^t}(v_{t,u})\| \le G_{\max}$. Thus, by Jensen's inequality and (61),

$$\begin{aligned} \mathbb{E}\left[\|D_s\|^2 \middle| \mathcal{F}_t \right] &\le \mathbb{E}\left[\frac{1}{M} \sum_{k \in S_t} \|d_k^s\|^2 \middle| \mathcal{F}_t \right] \\ &\le 2L_w^2 \left(D_{t,s} + \eta_\ell^2 s^2 G_{\max}^2 \right) \\ &\le L_w^2 \eta_\ell^2 s^2 \bar{G}^2. \end{aligned} \tag{68}$$

Here $\bar{G}^2 = 2(G_{\max}^2 + C_\delta Q)$ is the constant introduced in Lemma A.30.

Since $B = \sum_{s=0}^{K-1} D_s$, Cauchy–Schwarz gives

$$\|B\|^2 \le K \sum_{s=0}^{K-1} \|D_s\|^2.$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[\|B\|^2 \middle| \mathcal{F}_t \right] &\le K L_w^2 \eta_\ell^2 \bar{G}^2 \sum_{s=0}^{K-1} s^2 \\ &= \frac{L_w^2 \eta_\ell^2 \bar{G}^2}{6} K^2 (K-1)(2K-1), \end{aligned} \tag{69}$$

where

$$\sum_{s=0}^{K-1} s^2 = \frac{K(K-1)(2K-1)}{6}.$$

Therefore,

$$3\eta_\ell^2 \mathbb{E}\left[\|B\|^2 \middle| \mathcal{F}_t \right] \le \frac{L_w^2 \bar{G}^2}{2} \eta_\ell^4 K^2 (K-1)(2K-1). \tag{70}$$

Combining the three terms. Using

$$\Delta^t = -\eta_\ell (A + B + C)$$

and

$$\|A + B + C\|^2 \le 3\|A\|^2 + 3\|B\|^2 + 3\|C\|^2,$$

we obtain

$$\begin{aligned}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right] \le\ & 3\eta_\ell^2 \mathbb{E}\left[\|A\|^2 \mid \mathcal{F}_t\right] \\ & + 3\eta_\ell^2 \mathbb{E}\left[\|B\|^2 \mid \mathcal{F}_t\right] \\ & + 3\eta_\ell^2 \mathbb{E}\left[\|C\|^2 \mid \mathcal{F}_t\right]. \end{aligned} \tag{71}$$

Substituting the bounds for A, B, and C gives

$$\begin{aligned}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right] \le\ & 6\eta_\ell^2 K^2 \left\|\nabla L_{\boldsymbol{w}^t}(\theta^t)\right\|^2 + 3\eta_\ell^2 K \kappa_M \Sigma^2 \\ & + 6\eta_\ell^2 K^2 \mathcal{H}_N(p;\zeta) \\ & + \frac{L_w^2 \bar{G}^2}{2} \eta_\ell^4 K^2 (K-1)(2K-1). \end{aligned} \tag{72}$$

Therefore, (64) follows with

$$C_0 = 6, \qquad C_1 = 3, \qquad C_2 = 6, \qquad C_B = \frac{L_w^2 \bar{G}^2}{2}.$$

The final term is retained explicitly because it depends on η_ℓ and K. It vanishes when $K = 1$.

Stale-weight gradient error. Define the discrepancy between the gradient of the scalarized objective actually used in round t and the scalarized objective associated with the ideal current hypervolume weights by

$$e_t := \nabla L_{\boldsymbol{w}^t}(\theta^t) - \nabla L_{\boldsymbol{w}^{\star,t}}(\theta^t).$$

Equivalently,

$$e_t = \sum_{i=1}^m \left(w_i^t - w_i^{\star,t}\right) \nabla f_i(\theta^t).$$

Let

$$G_{\max} := \max_{i \in [m]} G_i.$$

Then, by Lemma A.6,

$$\|e_t\| \le G_{\max} \|\boldsymbol{w}^t - \boldsymbol{w}^{\star,t}\|_1. \tag{73}$$

Using Lemma A.21, for every $t \ge 1$,

$$\begin{aligned}\mathbb{E}\|e_t\|^2 \le\ & 2G_{\max}^2 C(\rho_{\text{clip}}, \boldsymbol{r})^2 \Xi_{\text{loss},T} \\ & + 2G_{\max}^2 C(\rho_{\text{clip}}, \boldsymbol{r})^2 \mathbb{E}\left[\left\|\boldsymbol{F}(\theta^t) - \boldsymbol{F}(\theta^{t-1})\right\|_\infty^2\right]. \end{aligned} \tag{74}$$

We next control the second term through the inter-round model displacement. Since $\boldsymbol{w}^t \in \Delta_m$,

$$\left\|\nabla L_{\boldsymbol{w}^t}(\theta^t)\right\| \le \sum_{i=1}^m w_i^t \|\nabla f_i(\theta^t)\| \le G_{\max}.$$

Therefore, Lemma A.32 implies

$$\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right] \le \eta_\ell^2 K^2 \mathcal{V}, \tag{75}$$

where

$$\mathcal{V} := C_0 G_{\max}^2 + C_1 \frac{\Sigma^2}{K} + C_2 \mathcal{H}_N(p;\zeta) + C_B \eta_\ell^2 (K-1)(2K-1).$$

Since

$$\theta^{t+1} - \theta^t = \eta_g \Delta^t,$$

we obtain

$$\mathbb{E}\left[\|\theta^{t+1}-\theta^t\|^2\right] \le \eta_{\text{eff}}^2 \mathcal{V}, \qquad \eta_{\text{eff}} = \eta_g \eta_\ell K. \tag{76}$$

By Lemma A.6, each global objective is G_i-Lipschitz on the region containing consecutive global iterates. Hence

$$\left\|\boldsymbol{F}(\theta^t) - \boldsymbol{F}(\theta^{t-1})\right\|_\infty \le G_{\max}\|\theta^t - \theta^{t-1}\|.$$

Combining this relation with (76) and (74) gives, for $t \ge 1$,

$$\mathbb{E}\|e_t\|^2 \le 2G_{\max}^2 C(\rho_{\text{clip}}, \boldsymbol{r})^2 \Xi_{\text{loss},T} + 2G_{\max}^4 C(\rho_{\text{clip}}, \boldsymbol{r})^2 \eta_{\text{eff}}^2 \mathcal{V}. \tag{77}$$

For the initialization round, both $\boldsymbol{w}^0$ and $\boldsymbol{w}^{\star,0}$ belong to Δ_m. Therefore,

$$\|\boldsymbol{w}^0 - \boldsymbol{w}^{\star,0}\|_1 \le 2, \qquad \|e_0\| \le 2G_{\max}.$$

The resulting initialization contribution occurs only once and will therefore produce an $O(T^{-1})$ term after telescoping.

Relating the actual weighted gradient to the log-hypervolume gradient. By the definition of e_t and the ideal colinearity relation (53),

$$\nabla L_{\boldsymbol{w}^t}(\theta^t) = \frac{1}{c^t}\nabla\Phi(\theta^t) + e_t. \tag{78}$$

Consequently,

$$\begin{aligned}\left\|\nabla L_{\boldsymbol{w}^t}(\theta^t)\right\|^2 &\le 2\left\|\nabla L_{\boldsymbol{w}^{\star,t}}(\theta^t)\right\|^2 + 2\|e_t\|^2 \\ &\le \frac{2}{c_{\min}^2}\left\|\nabla\Phi(\theta^t)\right\|^2 + 2\|e_t\|^2.\end{aligned} \tag{79}$$

Substituting (79) into Lemma A.32 yields

$$\begin{aligned}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right] \le{}& \frac{2C_0\eta_\ell^2 K^2}{c_{\min}^2}\left\|\nabla\Phi(\theta^t)\right\|^2 + 2C_0\eta_\ell^2 K^2\|e_t\|^2 \\ &+ C_1\eta_\ell^2 K\kappa_M\Sigma^2 + C_2\eta_\ell^2 K^2\mathcal{H}_N(p;\zeta) \\ &+ C_B\eta_\ell^4 K^2(K-1)(2K-1).\end{aligned} \tag{80}$$

One-round descent with stale estimated weights. By Lemma A.8 and

$$\theta^{t+1} = \theta^t + \eta_g\Delta^t,$$

conditioning on $\mathcal{F}_t$ gives

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1}) \mid \mathcal{F}_t\right] \le{}& \Phi(\theta^t) + \eta_g\left\langle\nabla\Phi(\theta^t), \mathbb{E}\left[\Delta^t \mid \mathcal{F}_t\right]\right\rangle \\ &+ \frac{L_\Phi\eta_g^2}{2}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right].\end{aligned} \tag{81}$$

Using (59),

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1}) \mid \mathcal{F}_t\right] \le{}& \Phi(\theta^t) - \eta_{\text{eff}}\left\langle\nabla\Phi(\theta^t), \nabla L_{\boldsymbol{w}^t}(\theta^t)\right\rangle \\ &+ \eta_g\left\langle\nabla\Phi(\theta^t), b_t\right\rangle + \frac{L_\Phi\eta_g^2}{2}\mathbb{E}\left[\|\Delta^t\|^2 \mid \mathcal{F}_t\right].\end{aligned} \tag{82}$$

From (78),

$$\left\langle\nabla\Phi(\theta^t), \nabla L_{\boldsymbol{w}^t}(\theta^t)\right\rangle = \frac{1}{c^t}\left\|\nabla\Phi(\theta^t)\right\|^2 + \left\langle\nabla\Phi(\theta^t), e_t\right\rangle. \tag{83}$$

Using $c^t \le c_{\max}$ and Young's inequality,

$$\left|\left\langle\nabla\Phi(\theta^t), e_t\right\rangle\right| \le \frac{1}{4c_{\max}}\left\|\nabla\Phi(\theta^t)\right\|^2 + c_{\max}\|e_t\|^2.$$

Therefore,

$$-\eta_{\text{eff}} \left\langle \nabla\Phi(\theta^t), \nabla L_{\boldsymbol{w}^t}(\theta^t) \right\rangle \leq -\frac{3\eta_{\text{eff}}}{4c_{\max}} \left\|\nabla\Phi(\theta^t)\right\|^2 + c_{\max}\eta_{\text{eff}}\|e_t\|^2. \tag{84}$$

For the local-drift bias, Young's inequality similarly gives

$$\eta_g \left|\left\langle \nabla\Phi(\theta^t), b_t \right\rangle\right| \leq \frac{\eta_{\text{eff}}}{4c_{\max}} \left\|\nabla\Phi(\theta^t)\right\|^2 + \eta_g \frac{c_{\max}}{\eta_\ell K}\|b_t\|^2. \tag{85}$$

Finally, substituting (80), (84), and (85) into (82), and choosing the constant in Assumption A.9 as specified there, the smoothness-generated gradient term is bounded by

$$\frac{L_\Phi C_0 \eta_{\text{eff}}^2}{c_{\min}^2} \leq \frac{\eta_{\text{eff}}}{8c_{\max}}.$$

The two Young inequalities already leave $-\eta_{\text{eff}}\|\nabla\Phi(\theta^t)\|^2/(2c_{\max})$; therefore the following one-round inequality holds with the displayed $-\eta_{\text{eff}}/(4c_{\max})$ coefficient:

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1}) \mid \mathcal{F}_t\right] \leq \Phi(\theta^t) &- \frac{\eta_{\text{eff}}}{4c_{\max}} \left\|\nabla\Phi(\theta^t)\right\|^2 \\ &+ C_e\eta_{\text{eff}}\|e_t\|^2 + \eta_g \frac{c_{\max}}{\eta_\ell K}\|b_t\|^2 \\ &+ \frac{L_\Phi C_1}{2}\eta_g^2\eta_\ell^2 K\Sigma^2 \\ &+ \frac{L_\Phi C_2}{2}\eta_g^2\eta_\ell^2 K^2 \mathcal{H}_N(p;\zeta) \\ &+ \frac{L_\Phi C_B}{2}\eta_g^2\eta_\ell^4 K^2(K-1)(2K-1),\end{aligned} \tag{86}$$

where

$$C_e := c_{\max} + L_\Phi C_0 \bar{\eta}_{\text{eff}} = c_{\max} + 6L_\Phi\bar{\eta}_{\text{eff}}, \quad C_\Sigma := \frac{L_\Phi C_1}{2},$$
$$C_H := \frac{L_\Phi C_2}{2}, \qquad C_B' := \frac{L_\Phi C_B}{2}.$$

Telescoping and stationarity rate. For convenience, define

$$C_{\text{est}} := 2G_{\max}^2 C(\rho_{\text{clip}}, \boldsymbol{r})^2, \qquad C_{\text{st}} := 2G_{\max}^4 C(\rho_{\text{clip}}, \boldsymbol{r})^2.$$

By (77), for every $t \geq 1$,

$$\mathbb{E}\|e_t\|^2 \leq C_{\text{est}}\Xi_{\text{loss},T} + C_{\text{st}}\eta_{\text{eff}}^2\mathcal{V}.$$

For $t = 0$, we have

$$\|e_0\|^2 \leq 4G_{\max}^2.$$

Hence, for every $t \geq 0$, the unified bound

$$\mathbb{E}\|e_t\|^2 \leq C_{\text{est}}\Xi_{\text{loss},T} + C_{\text{st}}\eta_{\text{eff}}^2\mathcal{V} + 4G_{\max}^2\mathbf{1}_{\{t=0\}} \tag{87}$$

holds.

Taking total expectation in (86) and using the tower property

$$\mathbb{E}\left[\Phi(\theta^{t+1})\right] = \mathbb{E}\left[\mathbb{E}\left[\Phi(\theta^{t+1}) \mid \mathcal{F}_t\right]\right],$$

Using the explicit definitions

$$C_\Sigma := \frac{L_\Phi C_1}{2}, \qquad C_H := \frac{L_\Phi C_2}{2}, \qquad C_B' := \frac{L_\Phi C_B}{2},$$

and the value of C_e specified in (86), we obtain

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1})\right] \le \mathbb{E}\left[\Phi(\theta^t)\right] &- \alpha\mathbb{E}\left[\left\|\nabla\Phi(\theta^t)\right\|^2\right] \\ &+ C_\Sigma\eta_g^2\eta_\ell^2 K\Sigma^2 + C_H\eta_g^2\eta_\ell^2 K^2\mathcal{H}_N(p;\zeta) \\ &+ \eta_g\frac{c_{\max}}{\eta_\ell K}\mathbb{E}\|b_t\|^2 \\ &+ C'_B\eta_g^2\eta_\ell^4 K^2(K-1)(2K-1) \\ &+ C_e\eta_{\text{eff}}\mathbb{E}\|e_t\|^2, \end{aligned} \tag{88}$$

where

$$\alpha = \frac{\eta_{\text{eff}}}{4c_{\max}}, \qquad \eta_{\text{eff}} := \eta_g\eta_\ell K.$$

Substituting (87) and using the explicit choice of C_e above gives

$$C_{\text{loss}} := C_e C_{\text{est}}, \qquad C_{\text{stale}} := C_e C_{\text{st}}, \qquad C_{\text{init}} := 4C_e G_{\max}^2.$$

Thus

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1})\right] \le \mathbb{E}\left[\Phi(\theta^t)\right] &- \alpha\mathbb{E}\left[\left\|\nabla\Phi(\theta^t)\right\|^2\right] \\ &+ C_\Sigma\eta_g^2\eta_\ell^2 K\Sigma^2 + C_H\eta_g^2\eta_\ell^2 K^2\mathcal{H}_N(p;\zeta) \\ &+ \eta_g\frac{c_{\max}}{\eta_\ell K}\mathbb{E}\|b_t\|^2 \\ &+ C'_B\eta_g^2\eta_\ell^4 K^2(K-1)(2K-1) \\ &+ C_{\text{loss}}\eta_{\text{eff}}\Xi_{\text{loss},T} + C_{\text{stale}}\eta_{\text{eff}}^3\mathcal{V} \\ &+ C_{\text{init}}\eta_{\text{eff}}\mathbf{1}_{\{t=0\}}. \end{aligned} \tag{89}$$

Using Lemma A.31, the conditional-mean local-drift contribution satisfies

$$\begin{aligned}\eta_g\frac{c_{\max}}{\eta_\ell K}\mathbb{E}\|b_t\|^2 &\le \frac{c_{\max}L_w^2\bar{G}^2}{4}\eta_g\eta_\ell^3 K(K-1)^2 \\ &= \widetilde{C}_b\eta_{\text{eff}}\eta_\ell^2(K-1)^2, \end{aligned} \tag{90}$$

where

$$\widetilde{C}_b := \frac{c_{\max}L_w^2\bar{G}^2}{4}.$$

Therefore, the one-round inequality can be written as

$$\begin{aligned}\mathbb{E}\left[\Phi(\theta^{t+1})\right] \le \mathbb{E}\left[\Phi(\theta^t)\right] &- \alpha\mathbb{E}\left[\left\|\nabla\Phi(\theta^t)\right\|^2\right] \\ &+ C_\Sigma\eta_g^2\eta_\ell^2 K\Sigma^2 + C_H\eta_g^2\eta_\ell^2 K^2\mathcal{H}_N(p;\zeta) \\ &+ \widetilde{C}_b\eta_{\text{eff}}\eta_\ell^2(K-1)^2 \\ &+ C'_B\eta_g^2\eta_\ell^4 K^2(K-1)(2K-1) \\ &+ C_{\text{loss}}\eta_{\text{eff}}\Xi_{\text{loss},T} + C_{\text{stale}}\eta_{\text{eff}}^3\mathcal{V} \\ &+ C_{\text{init}}\eta_{\text{eff}}\mathbf{1}_{\{t=0\}}. \end{aligned} \tag{91}$$

Telescoping and final stationarity bound. The objective is bounded below along the optimization trajectory. Indeed, Assumption A.4 and the nonnegativity of the losses imply

$$0 < s_i(\theta) = r_i - f_i(\theta) \le r_i.$$

Consequently,

$$\Phi(\theta) = -\sum_{i=1}^m \log s_i(\theta) \ge -\sum_{i=1}^m \log r_i =: \Phi_{\text{lb}}. \tag{92}$$

Taking total expectations in the unconditional one-round inequality (91), summing over the deterministic range $t = 0, \dots, T-1$, and using the lower bound $\Phi(\theta) \geq \Phi_{\mathrm{lb}}$ gives

$$\begin{aligned}\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla\Phi(\theta^t)\right\|^2\right] \leq \Phi(\theta^0) - \Phi_{\mathrm{lb}} \\ + TC_\Sigma \eta_g^2 \eta_\ell^2 K \Sigma^2 \\ + TC_H \eta_g^2 \eta_\ell^2 K^2 \mathcal{H}_N(p;\zeta) \\ + T\widetilde{C}_b \eta_{\mathrm{eff}} \eta_\ell^2 (K-1)^2 \\ + TC'_B \eta_g^2 \eta_\ell^4 K^2 (K-1)(2K-1) \\ + TC_{\mathrm{loss}} \eta_{\mathrm{eff}} \Xi_{\mathrm{loss},T} \\ + TC_{\mathrm{stale}} \eta_{\mathrm{eff}}^3 \mathcal{V} \\ + C_{\mathrm{init}} \eta_{\mathrm{eff}}. \end{aligned} \tag{93}$$

Dividing this inequality by αT gives the fixed-horizon average bound.

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla\Phi(\theta^t)\right\|^2\right] \leq \frac{4c_{\max}(\Phi(\theta^0) - \Phi_{\mathrm{lb}})}{\eta_{\mathrm{eff}} T} + 4c_{\max} C_\Sigma \frac{\eta_{\mathrm{eff}}}{K} \Sigma^2 \\ + 4c_{\max} C_H \eta_{\mathrm{eff}} \mathcal{H}_N(p;\zeta) \\ + 4c_{\max} \widetilde{C}_b \eta_\ell^2 (K-1)^2 \\ + 4c_{\max} C'_B \eta_{\mathrm{eff}} \eta_\ell^2 (K-1)(2K-1) \\ + 4c_{\max} C_{\mathrm{loss}} \Xi_{\mathrm{loss},T} \\ + 4c_{\max} C_{\mathrm{stale}} \eta_{\mathrm{eff}}^2 \mathcal{V} \\ + \frac{4c_{\max} C_{\mathrm{init}}}{T}. \end{aligned} \tag{94}$$

For the stochastic-gradient term, $K \geq 1$ gives $\eta_{\mathrm{eff}}/K \leq \eta_{\mathrm{eff}}$. By Assumption A.9,

$$\eta_{\mathrm{eff}} \leq \bar{\eta}_{\mathrm{eff}}, \qquad (K-1)^2 \leq (K-1)(2K-1).$$

The right-hand side of (94) is therefore bounded by the theorem's right-hand side after substituting the explicit constants listed in the theorem.

This proves the claimed fixed-horizon bound with the explicit coefficients $A_1, \dots, A_7$ stated in Theorem A.22. □

A.12.10 Proportional-Fairness Background and Local First-Order Proofs

Classical proportional fairness in slack space. Let

$$\mathcal{S} := \{\boldsymbol{s}(\theta) : \theta \in \Theta\} \subset \mathbb{R}^m_{++}$$

denote the achievable slack set.

Definition A.33 (Proportional fairness in slack space)**.** A point $\boldsymbol{s}^\star \in \mathcal{S}$ is proportionally fair if, for every $\boldsymbol{s} \in \mathcal{S}$,

$$\sum_{i=1}^m \frac{s_i - s_i^\star}{s_i^\star} \leq 0. \tag{95}$$

If $\mathcal{S}$ is convex and compact, any maximizer of

$$\sum_{i=1}^m \log s_i$$

over $\mathcal{S}$ is proportionally fair in the sense of Definition A.33 (Kelly, 1997). This classical statement is included only for background. The achievable slack set induced by a neural network is generally

nonconvex, and our convergence result establishes stationarity rather than global optimality. Accordingly, we do not claim that FedHV obtains a classical global Nash bargaining solution or a globally proportionally fair slack vector; the paper uses only the local first-order implication below.

Lemma A.34 (First-order proportional fairness)**.** *Let $\theta^\star \in \Theta$ satisfy $s_i(\theta^\star) > 0$ for every i, and suppose that $\theta^\star$ satisfies the first-order necessary condition for maximizing*

$$\Psi(\theta) = \sum_{i=1}^{m} \log s_i(\theta)$$

over Θ; that is,

$$\langle \nabla \Psi(\theta^\star), d \rangle \le 0, \qquad \forall d \in T_\Theta(\theta^\star). \tag{96}$$

Then, for every $d \in T_\Theta(\theta^\star)$,

$$\sum_{i=1}^{m} \frac{\langle \nabla s_i(\theta^\star), d \rangle}{s_i(\theta^\star)} \le 0. \tag{97}$$

Equivalently,

$$\sum_{i=1}^{m} \frac{\langle \nabla f_i(\theta^\star), d \rangle}{r_i - f_i(\theta^\star)} \ge 0. \tag{98}$$

Proof. Differentiating the reference-dependent Nash-welfare surrogate gives

$$\nabla \Psi(\theta) = \sum_{i=1}^{m} \frac{1}{s_i(\theta)} \nabla s_i(\theta).$$

Therefore,

$$\langle \nabla \Psi(\theta^\star), d \rangle = \sum_{i=1}^{m} \frac{\langle \nabla s_i(\theta^\star), d \rangle}{s_i(\theta^\star)}.$$

The first inequality follows from (96). Since

$$\nabla s_i(\theta) = -\nabla f_i(\theta),$$

the equivalent loss-space expression follows immediately. □

Proof of Lemma A.28. By Definition A.27,

$$\begin{aligned} \mathrm{PFgap}(\theta) &= \sup_{\substack{d \in T_\Theta(\theta) \\ \|d\| \le 1}} \langle \nabla \Psi(\theta), d \rangle \\ &\le \sup_{\|d\| \le 1} \langle \nabla \Psi(\theta), d \rangle = \|\nabla \Psi(\theta)\|. \end{aligned} \tag{99}$$

The last equality is the dual characterization of the Euclidean norm. Since $\Psi = -\Phi$,

$$\|\nabla \Psi(\theta)\| = \|\nabla \Phi(\theta)\|.$$

If $\Theta = \mathbb{R}^d$, then $T_\Theta(\theta) = \mathbb{R}^d$. When $\nabla \Psi(\theta) \neq 0$, choosing

$$d^\star = \frac{\nabla \Psi(\theta)}{\|\nabla \Psi(\theta)\|}$$

attains the upper bound. When $\nabla \Psi(\theta) = 0$, both sides are zero. Hence equality holds in the unconstrained case. □

Proof of Corollary A.29. Lemma A.28 implies, for every t,

$$\mathrm{PFgap}(\theta^t)^2 \le \|\nabla \Phi(\theta^t)\|^2.$$

For the index $t^\star$ defined in Corollary A.29, the Cauchy–Schwarz inequality therefore gives

$$\begin{aligned} \mathbb{E}\left[\mathrm{PFgap}(\theta^{t^\star})\right] &\le \left(\mathbb{E}\left[\mathrm{PFgap}(\theta^{t^\star})^2\right]\right)^{1/2} \\ &\le \left(\mathbb{E}\left[\|\nabla \Phi(\theta^{t^\star})\|^2\right]\right)^{1/2} \\ &= \left(\min_{0 \le t < T} \mathbb{E}\left[\|\nabla \Phi(\theta^t)\|^2\right]\right)^{1/2}. \end{aligned} \tag{100}$$

This proves (43). □

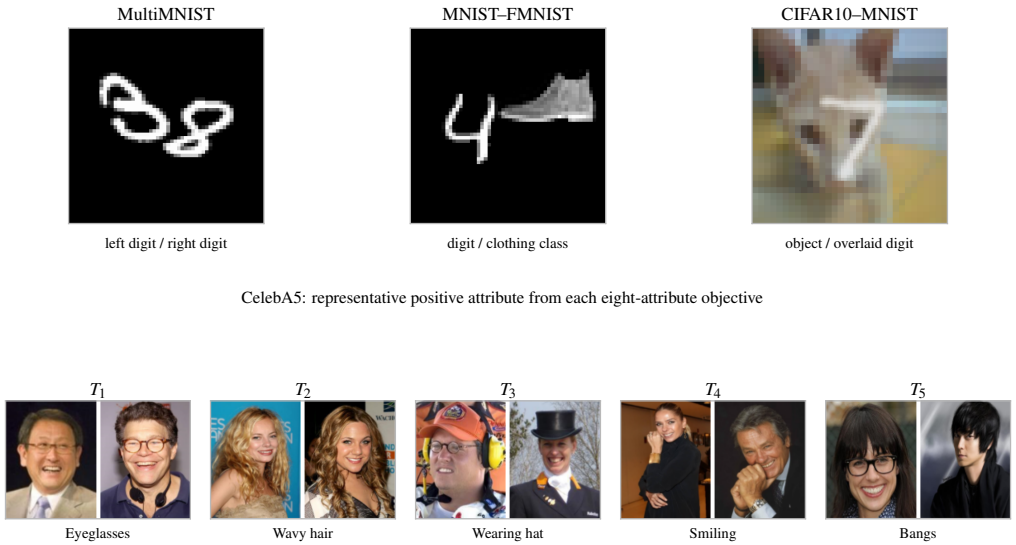


Figure 4: Dataset and benchmark examples. The first row illustrates the two shifted digits in MULTIMNIST, the digit/clothing pair in MNIST–FMNIST, and the overlaid digit/object pair in CIFAR10–MNIST. The second row shows two positive CELEBA examples for one representative attribute from each objective group: Eyeglasses (T_1), Wavy Hair (T_2), Wearing Hat (T_3), Smiling (T_4), and Bangs (T_5). Source datasets are described by LeCun et al. (1998); Xiao et al. (2017); Krizhevsky (2009); Liu et al. (2015).

B EXPERIMENTAL DETAILS AND REPRODUCIBILITY

B.1 BENCHMARK CONSTRUCTION

MULTIMNIST. Each example contains two spatially shifted MNIST digits in one image; overlapping pixels are combined by the maximum intensity. The two tasks predict the left and right digit labels (LeCun et al., 1998).

MNIST–FMNIST. Each example combines one MNIST digit and one Fashion-MNIST item at offset locations in a common grayscale image. The two tasks are digit classification and Fashion-MNIST category classification (LeCun et al., 1998; Xiao et al., 2017).

CIFAR10–MNIST. A semi-transparent MNIST digit is superimposed on a CIFAR-10 image. The two tasks are CIFAR-10 object classification and MNIST digit classification. This benchmark creates a pronounced difference in task difficulty and is therefore useful for examining worst-task behavior (LeCun et al., 1998; Krizhevsky, 2009).

CELEBA5. CelebA contains 40 binary facial attributes. Following the FedCMOO benchmark construction, we partition them into five fixed objectives of eight attributes each. The exact grouping is given in Table 2 (Liu et al., 2015; Askin et al., 2025).

Figure 4 visually summarizes the four benchmark constructions. For CELEBA5, each displayed cue is a representative positive attribute from one eight-attribute objective; the visualization therefore complements, rather than replaces, the complete grouping in Table 2.

Table 2: The five multi-label objectives used in CELEBA5.

Objective	Attributes
T_1	5-o'Clock Shadow, Blond Hair, Bags Under Eyes, Eyeglasses, Mustache, Wavy Hair, Oval Face, Rosy Cheeks
T_2	Arched Eyebrows, Black Hair, Big Nose, Blurry, Goatee, Straight Hair, Pale Skin, Wearing Earrings
T_3	Attractive, Brown Hair, Bushy Eyebrows, Double Chin, Mouth Slightly Open, Receding Hairline, Wearing Hat, Young
T_4	Bald, Big Lips, Chubby, Heavy Makeup, High Cheekbones, Male, Sideburns, Smiling
T_5	Bangs, Gray Hair, Narrow Eyes, No Beard, Pointy Nose, Wearing Lipstick, Wearing Necklace, Wearing Necktie

B.2 OBJECTIVE AND METRIC DEFINITIONS

The reduction convention is part of the definition of FedHV because rescaling an objective changes its slack and hence its adaptive weight. For each two-objective benchmark and client k, task i uses mean multiclass cross-entropy,

$$f_i^{(k)}(\theta) = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathrm{CE}(z_i(x_{k,j}; \theta), y_{i,k,j}) \,, \tag{101}$$

where i denotes the left/right digit, digit/clothing, or CIFAR-10/MNIST head, respectively. Task accuracy is the fraction of examples whose top-1 prediction from that head is correct.

For CELEBA5, let G_i be the eight binary attributes assigned to objective i in Table 2. The client objective is the mean binary cross-entropy over both examples and the eight attributes,

$$f_i^{(k)}(\theta) = \frac{1}{8n_k} \sum_{j=1}^{n_k} \sum_{a \in G_i} \mathrm{BCE}(z_a(x_{k,j}; \theta), y_{a,k,j}) \,. \tag{102}$$

The corresponding task accuracy averages the thresholded binary attribute accuracy over the same $8n_k$ predictions. Objective reports, training references, MeanLoss, and evaluation HV all use these mean reductions; no objective is summed over its classes or attributes.

B.3 MODEL ARCHITECTURES

The model architecture is shared by all compared methods within a benchmark. MULTIMNIST and MNIST–FMNIST use a LeNet-style shared encoder with 3×3 convolutions $1 \to 10 \to 15$, max pooling, a $540 \to 50$ fully connected layer, and dropout 0.5. Each task head uses $50 \to 50 \to 10$ fully connected layers with dropout 0.5 and log-softmax output. CIFAR10–MNIST uses two 3×3 64-channel convolutions with max pooling, followed by $4096 \to 384 \to 192$ fully connected layers with dropout 0.5; each task decoder is $192 \to 100 \to 10$. CELEBA5 uses a ResNet-18-style shared backbone with BasicBlock depths $[2, 2, 2, 2]$ and GroupNorm with two groups, producing a 512-dimensional representation; each of the five decoders is a linear $512 \to 8$ sigmoid head. We use the implementation's standard PyTorch initialization and no pretrained weights.

B.4 FEDERATED PARTITIONING

Table 3 lists the six main benchmark settings. The partition seed is fixed to 10 across methods within each setting.

Table 3: Federated partitioning and default participation in the six main settings. The all-label strategy forms Dirichlet allocation strata from all available task-label coordinates, whereas the first-label strategy forms strata using only the first label coordinate. Smaller α yields more concentrated client label distributions (Hsu et al., 2019).

Setting	N	M	Samples/client	α	Partition rule
MULTIMNIST	30	20	600	0.15	all-label Dirichlet
MNIST–FMNIST	30	10	500	0.30	all-label Dirichlet
CELEBA5-A	30	10	500	0.15	all-label Dirichlet
CELEBA5-B	30	10	500	0.30	first-label Dirichlet
CIFAR10–MNIST-A	30	10	500	0.10	all-label Dirichlet
CIFAR10–MNIST-B	30	10	500	0.30	all-label Dirichlet

The two CELEBA5 settings jointly vary partition strategy and concentration: setting A uses all-label allocation with $\alpha = 0.15$, whereas setting B uses first-label allocation with $\alpha = 0.30$. The two CIFAR10–MNIST settings retain the same all-label strategy at $\alpha \in \{0.10, 0.30\}$.

B.5 TRAINING AND EVALUATION PROTOCOL

All methods use SGD and the same round budget and local-training structure within a setting. Table 4 summarizes the shared quantities. No local learning-rate scheduler is used.

Table 4: Shared training and evaluation settings. K is the number of local SGD steps per selected client.

Setting	Adaptive rounds	Batch size	K	Momentum	Test period
MULTIMNIST $\alpha = 0.15$	200	32	10	0	3
MNIST–FMNIST $\alpha = 0.30$	500	128	10	0	3
CELEBA5 $\alpha = 0.15$	300	64	5	0.9	5
CELEBA5 $\alpha = 0.30$	300	64	5	0.9	5
CIFAR10–MNIST $\alpha = 0.10$	500	128	10	0	3
CIFAR10–MNIST $\alpha = 0.30$	500	128	10	0	3

Final tables use the final scheduled test evaluation. Curves report the mean across seeds with a ± 1 standard-deviation band. The common evaluation hypervolume reference is $\boldsymbol{r}^{\mathrm{eval}} = 3\,\mathbf{1}_m$ for all methods and is distinct from FedHV's training reference.

B.6 BASELINE IMPLEMENTATIONS AND LEARNING RATES

We evaluate FSMGDA and FedCMOO using their defining algorithm-specific procedures under the common experimental protocol described above. FedCMOO uses its randomized-SVD estimator with additional upload dimension 1, followed by 1000 projected server-side weight-update iterations with step size 10^{-3} and zero momentum. FSMGDA performs objective-specific local updates and computes the server-side combination weights using the MinNorm/MGDA procedure.

All methods use the same model architecture, client partition, participation schedule, number of communication rounds, and local-update budget within each benchmark, as specified in Tables 3 and 4. Thus, the methods differ principally in how they estimate and update the objective weights and in the resulting computation and communication required by those procedures.

Uniform uses fixed weights $w_i = 1/m$ throughout training and otherwise follows the same single-trajectory local optimization and model-aggregation structure as FedHV. It does not perform reference calibration or communicate objective reports for adaptive weighting.

We use method-specific optimization rates rather than imposing a single learning-rate pair across algorithms with different update mechanisms. Table 5 reports the server and local learning rates, η_g/η_ℓ, used for every method and benchmark. Each reported pair is held fixed across the three training seeds.

Table 5: Server/local learning rates η_g/η_ℓ used in the reported runs. Each pair is fixed across training seeds.

Setting	Uniform	FSMGDA	FedCMOO	FedHV
MULTIMNIST $\alpha = 0.15$	$1.0/.3$	$2.0/.1$	$1.2/.5$	$1.2/.3$
MNIST–FMNIST $\alpha = 0.30$	$1.2/.4$	$2.0/.1$	$1.2/.5$	$1.6/.3$
CELEBA5 $\alpha = 0.15$	$.8/.3$	$1.6/.3$	$1.6/.3$	$1.0/.2$
CELEBA5 $\alpha = 0.30$	$.8/.3$	$1.6/.3$	$1.6/.3$	$1.0/.2$
CIFAR10–MNIST $\alpha = 0.10$	$1.2/.4$	$2.0/.1$	$1.6/.3$	$1.6/.3$
CIFAR10–MNIST $\alpha = 0.30$	$1.2/.4$	$1.6/.3$	$1.6/.3$	$1.6/.3$

B.7 FEDHV HYPERPARAMETERS

Unless an ablation changes it, FedHV uses full-client objective reports ($B_{\text{rep}} = \text{all}$), no update normalization, and numerical estimated-slack floor $\rho_{\text{clip}} = 10^{-6}$. The weight vector is held fixed over all K local steps. The default M and K values are those in Tables 3 and 4. The objective-report budget ablation replaces the full-client report by $B_{\text{rep}} \in \{1, 2, 5\}$ minibatches without changing the local training minibatches.

B.8 REFERENCE CALIBRATION

The fixed FedHV training reference is constructed before adaptive training. For MULTIMNIST and MNIST–FMNIST, a single 20-round uniform-weight calibration run is used. For the CIFAR10–MNIST family, one 20-round calibration run with the all-label $\alpha = 0.30$ partition is used and the resulting reference is reused for both $\alpha = 0.10$ and $\alpha = 0.30$; hence the reference is not retuned to the lower-α setting. In each case we take the componentwise maximum reported calibration loss and add margin $\delta_r = .15$.

For CELEBA5, the robust reference used in the reported runs is constructed from three independent 20-round calibration runs with seeds $\{0, 42, 2026\}$ under the first-label $\alpha = 0.30$ calibration configuration. We take the componentwise maximum over all three runs and add $\delta_r = .30$. This single reference is then reused for both CELEBA5 benchmark settings and all adaptive-training seeds.

Table 6: Frozen training references used by FedHV. These are optimization references $\boldsymbol{r}^{\text{train}}$ and are distinct from the common evaluation reference $\boldsymbol{r}^{\text{eval}} = 3\,\mathbf{1}_m$.

Benchmark family	$\boldsymbol{r}^{\text{train}}$
MULTIMNIST	$(2.458940,\ 2.453542)$
MNIST–FMNIST	$(2.454883,\ 2.464302)$
CIFAR10–MNIST	$(2.452658,\ 2.466771)$
CELEBA5	$(1.297312,\ 1.023714,\ 1.060341,\ 1.108676,\ 1.085493)$

Only the resulting reference is reused: adaptive models are restarted from their normal initialization rather than continued from calibration, and all reported adaptive-round curves exclude calibration. Reference construction is therefore a one-time preprocessing cost that baselines do not require. It costs 20 calibration rounds for each two-task benchmark family; the robust CELEBA5 reference used here costs three 20-round calibration runs (60 sequential-round equivalents), shared across both CelebA5 settings and all adaptive seeds. This setup cost is distinct from our within-round communication claim: every adaptive FedHV round still uses one downlink and one uplink with no additional synchronization stage.

The calibration rule is empirical and does not by itself certify the true safe-reference condition in Assumption A.4. The numerical floor ρ_{clip} stabilizes estimated slacks but likewise does not establish that theoretical condition. Table 7 reports the closest diagnostic available from the retained logs: the minimum raw slack formed from the aggregated server objective report. All recorded standard-run report slacks remain positive and clipping is inactive. These values are not true full-client training losses at every local proof point, so they cannot certify Assumption A.4. The current runs also use

only $\delta_r \in \{0.15, 0.30\}$ as described above; they do not constitute a controlled reference-margin sensitivity sweep. Theorem A.15 provides a separate conditional route from a verified initial margin to trajectory safety, but its bounded-update and budget inequalities were not evaluated for these runs.

Table 7: Reference diagnostics over all adaptive rounds and three training seeds. "Minimum report slack" is $\min_{t,i}(r_i - \widehat{f}_i^{\,t})$ for the server-aggregated report; it is a practical diagnostic rather than the true-slack condition in Assumption A.4.

Setting	Minimum report slack	Floor activation
MULTIMNIST	0.1463	0/1200
MNIST–FMNIST	0.1413	0/3000
CELEBA5-A	0.2981	0/4500
CELEBA5-B	0.3000	0/4500
CIFAR10–MNIST-A	0.0282	0/3000
CIFAR10–MNIST-B	0.1412	0/3000

B.9 RANDOM SEEDS AND REPRODUCIBILITY

The client partition is generated once with partition seed 10 and is held fixed across methods within a setting. Adaptive training is repeated with run seeds 0, 42, and 2026. We report the arithmetic mean and sample standard deviation across these three runs. GPU and framework operations are not assumed fully deterministic, so small sub-standard-deviation differences are not interpreted as statistically significant. The three ablation sweeps were launched as independent experiment suites; consequently, nominally identical default variants (e.g., $K = 10$, $M = 10$, or $B_{\mathrm{rep}} = \mathrm{all}$) can differ by a few tenths of a percentage point across sweeps. We therefore interpret comparisons within each controlled sweep and do not treat such small cross-sweep differences as effects. All compared methods use the same evaluation schedule and test data within each setting.

C Complete Experimental Results

C.1 Loss and Hypervolume Results

Table 8: Final loss-based results for Uniform and the three adaptive FedMOO methods, mean ± standard deviation over three seeds. Lower MeanLoss is better; higher HV is better. HV uses the common evaluation reference $\boldsymbol{r}^{\mathrm{eval}} = 3\,\mathbf{1}_m$, not FedHV's training reference. Bold marks the best result and underline the second-best result within each setting.

Setting	Method	MeanLoss ↓	HV ↑
MultiMNIST	Uniform	$\underline{0.328 \pm 0.008}$	$\underline{7.14 \pm 0.04}$
	FSMGDA	0.359 ± 0.032	6.97 ± 0.17
	FedCMOO	0.354 ± 0.009	7.00 ± 0.05
	FedHV	$\mathbf{0.312 \pm 0.019}$	$\mathbf{7.23 \pm 0.10}$
MNIST–FMNIST	Uniform	$\underline{0.379 \pm 0.006}$	$\underline{6.82 \pm 0.03}$
	FSMGDA	0.427 ± 0.020	6.57 ± 0.10
	FedCMOO	$\mathbf{0.372 \pm 0.007}$	$\mathbf{6.87 \pm 0.04}$
	FedHV	0.382 ± 0.014	6.81 ± 0.07
CelebA5-A	Uniform	$\mathbf{0.267 \pm 0.003}$	$\mathbf{152.32 \pm 0.77}$
	FSMGDA	0.295 ± 0.003	144.94 ± 0.85
	FedCMOO	0.302 ± 0.005	142.96 ± 1.32
	FedHV	$\underline{0.269 \pm 0.004}$	$\underline{151.78 \pm 1.05}$
CelebA5-B	Uniform	$\mathbf{0.271 \pm 0.002}$	$\mathbf{151.40 \pm 0.48}$
	FSMGDA	0.318 ± 0.004	138.74 ± 0.94
	FedCMOO	0.342 ± 0.038	132.92 ± 9.31
	FedHV	$\underline{0.273 \pm 0.001}$	$\underline{150.92 \pm 0.25}$
CIFAR10–MNIST-A	Uniform	$\underline{0.821 \pm 0.013}$	$\underline{4.37 \pm 0.07}$
	FSMGDA	$\mathbf{0.812 \pm 0.007}$	$\mathbf{4.44 \pm 0.05}$
	FedCMOO	0.857 ± 0.027	4.14 ± 0.17
	FedHV	0.897 ± 0.019	3.98 ± 0.11
CIFAR10–MNIST-B	Uniform	$\underline{0.766 \pm 0.013}$	$\underline{4.67 \pm 0.06}$
	FSMGDA	0.798 ± 0.017	4.48 ± 0.09
	FedCMOO	$\mathbf{0.753 \pm 0.008}$	$\mathbf{4.72 \pm 0.03}$
	FedHV	0.826 ± 0.020	4.37 ± 0.10

The CIFAR rows make the metric trade-off explicit: FedHV improves mean and worst-task accuracy (Table 1) without also optimizing the common-reference loss HV.

C.2 Per-Task Final Metrics

Table 9: Final task-level test performance for the four two-objective settings, mean ± standard deviation over three seeds. Accuracies are percentages. Bold is best and underline is second best in each metric among FSMGDA, FedCMOO, and FedHV within a setting.

Setting	Method	Task 1 Acc.	Task 2 Acc.	Task 1 Loss	Task 2 Loss
MultiMNIST $\alpha = 0.15$	FSMGDA	$\underline{91.41 \pm 0.31}$	86.15 ± 2.53	$\underline{0.292 \pm 0.012}$	0.425 ± 0.056
	FedCMOO	90.99 ± 0.41	$\underline{88.79 \pm 0.57}$	0.339 ± 0.025	$\underline{0.369 \pm 0.008}$
	FedHV	$\mathbf{92.50 \pm 0.68}$	$\mathbf{88.94 \pm 0.61}$	$\mathbf{0.257 \pm 0.018}$	$\mathbf{0.366 \pm 0.021}$
MNIST–FMNIST $\alpha = 0.30$	FSMGDA	94.11 ± 0.81	75.58 ± 1.05	0.204 ± 0.021	0.650 ± 0.020
	FedCMOO	$\underline{94.81 \pm 0.43}$	$\mathbf{78.58 \pm 0.66}$	$\underline{0.186 \pm 0.019}$	$\mathbf{0.558 \pm 0.017}$
	FedHV	$\mathbf{95.67 \pm 0.31}$	$\underline{77.03 \pm 0.70}$	$\mathbf{0.161 \pm 0.018}$	$\underline{0.602 \pm 0.012}$
CIFAR10–MNIST $\alpha = 0.10$	FSMGDA	$\underline{51.96 \pm 0.55}$	$\underline{93.56 \pm 0.25}$	$\mathbf{1.405 \pm 0.022}$	$\underline{0.218 \pm 0.009}$
	FedCMOO	47.98 ± 1.15	$\mathbf{95.10 \pm 0.34}$	$\underline{1.529 \pm 0.064}$	$\mathbf{0.184 \pm 0.011}$
	FedHV	$\mathbf{53.03 \pm 0.35}$	93.34 ± 0.08	1.561 ± 0.041	0.232 ± 0.003
CIFAR10–MNIST $\alpha = 0.30$	FSMGDA	51.85 ± 1.00	$\underline{94.36 \pm 0.32}$	$\underline{1.408 \pm 0.028}$	$\underline{0.188 \pm 0.007}$
	FedCMOO	$\underline{54.23 \pm 0.88}$	$\mathbf{95.53 \pm 0.16}$	$\mathbf{1.328 \pm 0.008}$	$\mathbf{0.177 \pm 0.011}$
	FedHV	$\mathbf{57.55 \pm 0.20}$	93.38 ± 0.13	1.423 ± 0.031	0.228 ± 0.009

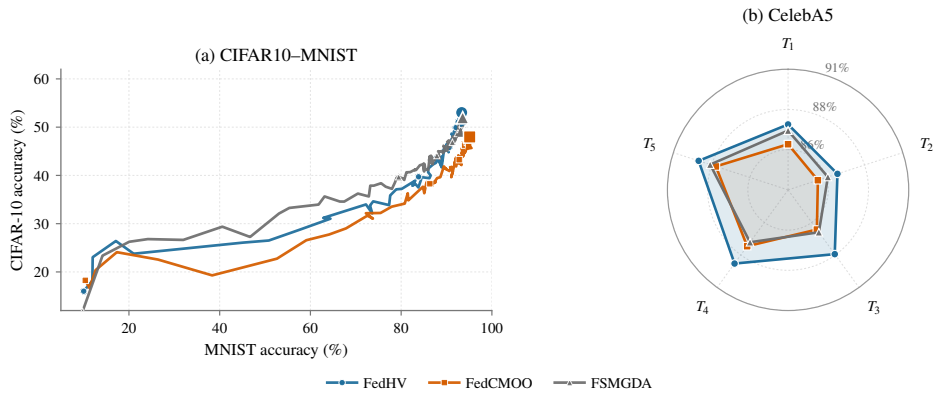


Figure 5: Alternate task-level views using the same visual encoding as Figure 2. **(a)** Round-by-round training trajectories for CIFAR10–MNIST-A (all-label Dirichlet, $\alpha = 0.10$), with MNIST accuracy on the horizontal axis and CIFAR-10 accuracy on the vertical axis. Each path follows scheduled evaluation checkpoints from early to late training; the terminal cross shows ± 1 standard deviation on both objectives. **(b)** Mean five-objective radar profile for CELEBA5-A (all-label Dirichlet, $\alpha = 0.15$). The radial scale is truncated to 83–91% for legibility; task-wise uncertainty is reported in Table 9. The main paper reports the corresponding B settings.

C.3 AGGREGATE LEARNING CURVES

Each curve below is the mean over three seeds; shaded bands show ± 1 standard deviation. The following figures show the optimization trajectories of FedHV, FSMGDA, and FedCMOO across communication rounds.

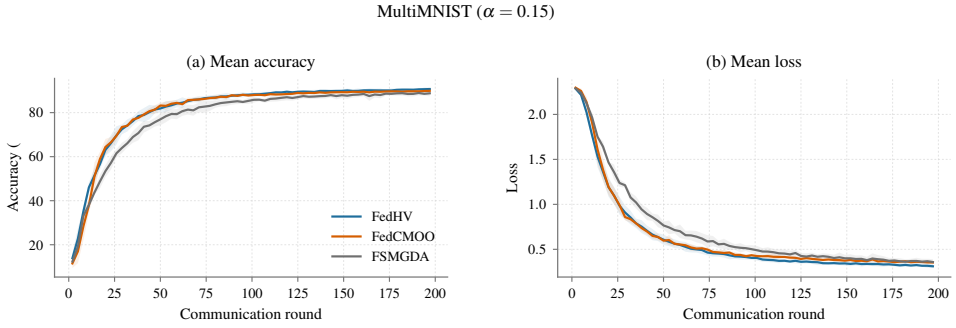


Figure 6: Aggregate test accuracy and loss for MULTIMNIST $\alpha = 0.15$ across adaptive communication rounds.

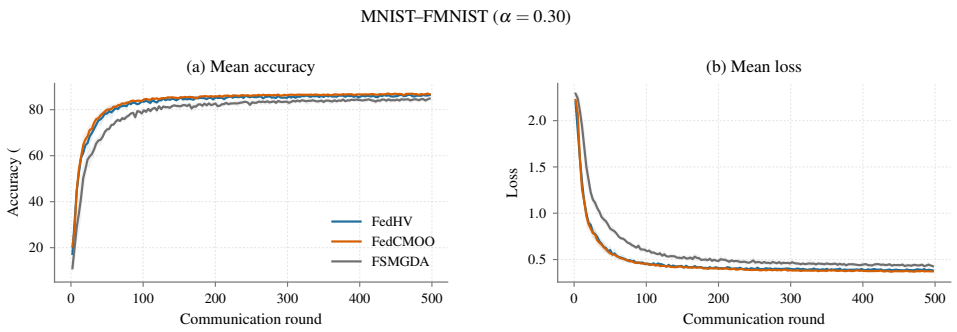


Figure 7: Aggregate test accuracy and loss for MNIST–FMNIST $\alpha = 0.30$ across adaptive communication rounds.

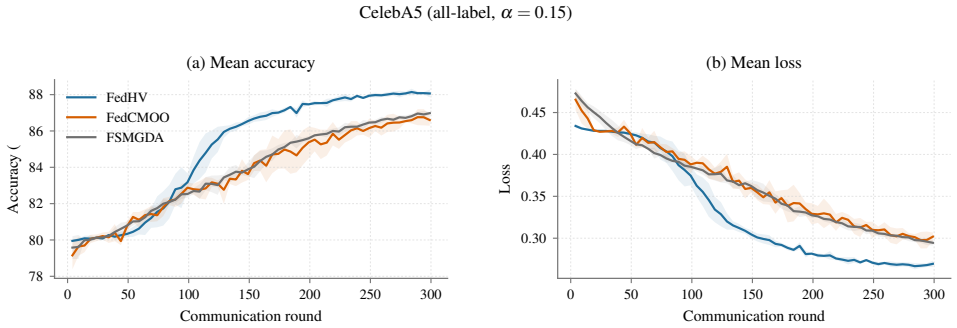


Figure 8: Aggregate test accuracy and loss for CELEBA5 $\alpha = 0.15$ across adaptive communication rounds.

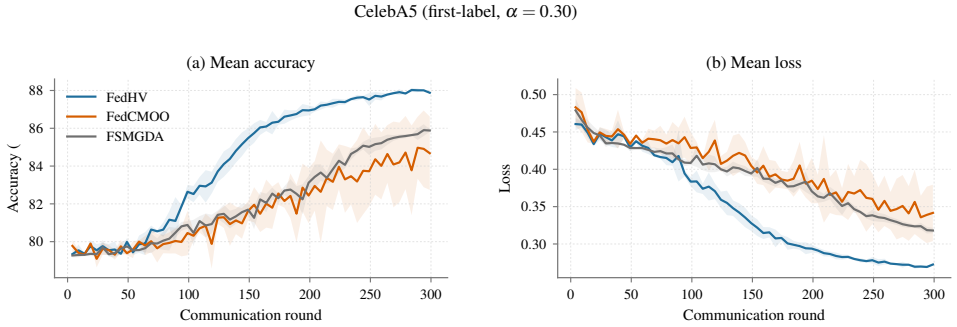


Figure 9: Aggregate test accuracy and loss for CELEBA5 $\alpha = 0.30$ across adaptive communication rounds.

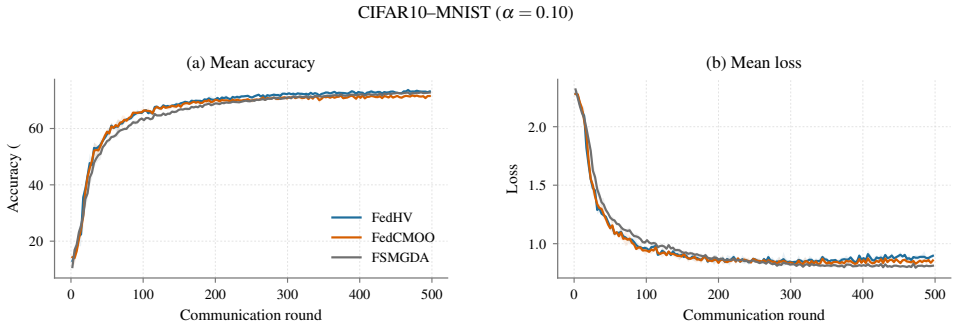


Figure 10: Aggregate test accuracy and loss for CIFAR10–MNIST $\alpha = 0.10$ across adaptive communication rounds.

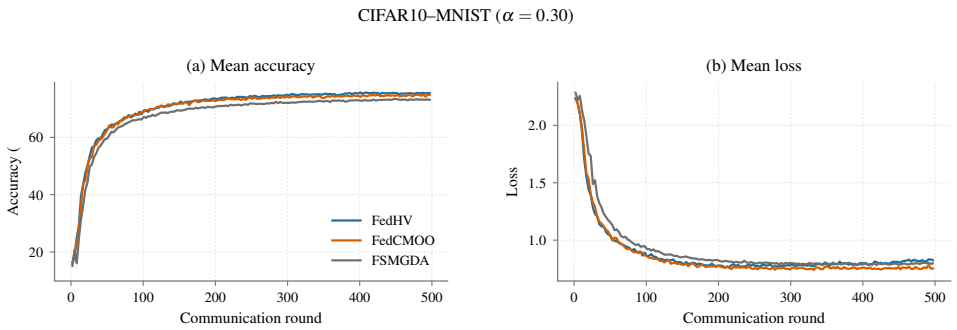


Figure 11: Aggregate test accuracy and loss for CIFAR10–MNIST $\alpha = 0.30$ across adaptive communication rounds.

C.4 Per-Task Learning Curves

For the two-objective benchmarks, the following figures separate the accuracy and loss trajectory of each task. This avoids hiding a difficult objective behind the mean.

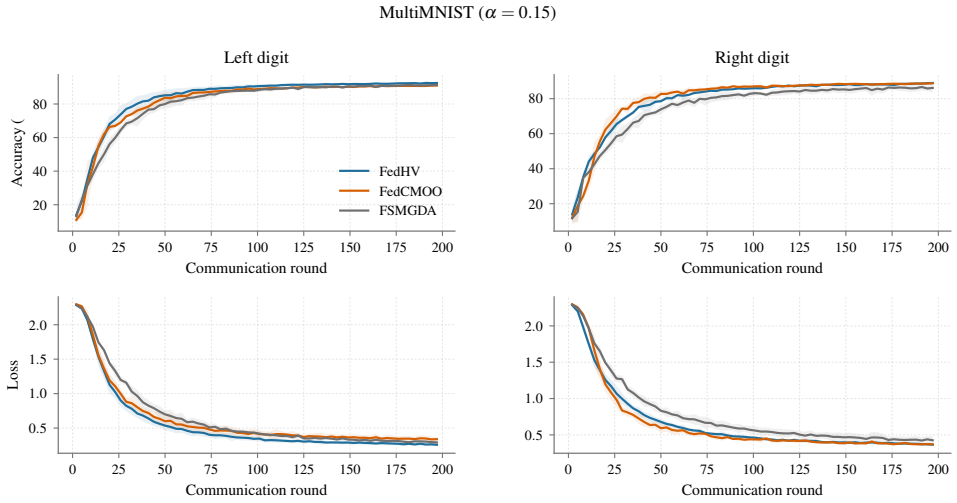


Figure 12: Per-task test accuracy and loss for MultiMNIST $\alpha = 0.15$, mean $\pm$ standard deviation over three seeds.

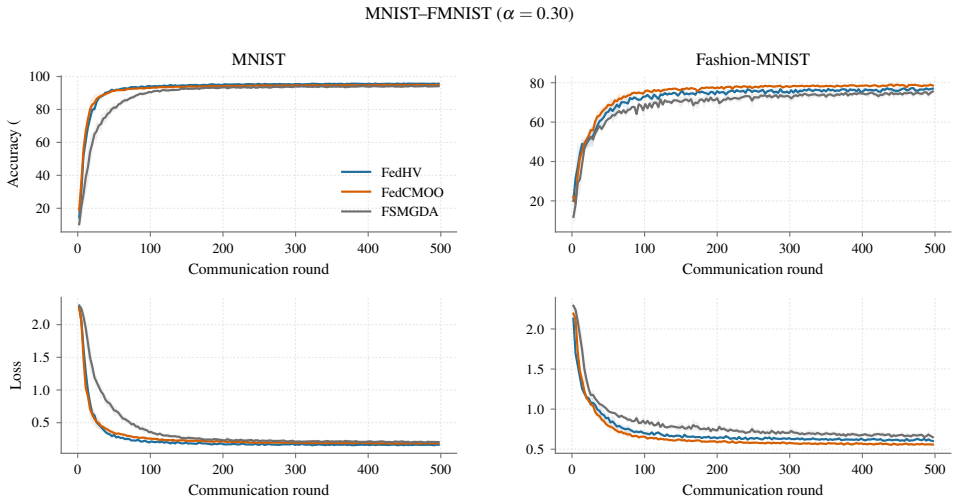


Figure 13: Per-task test accuracy and loss for MNIST–FMNIST $\alpha = 0.30$, mean $\pm$ standard deviation over three seeds.

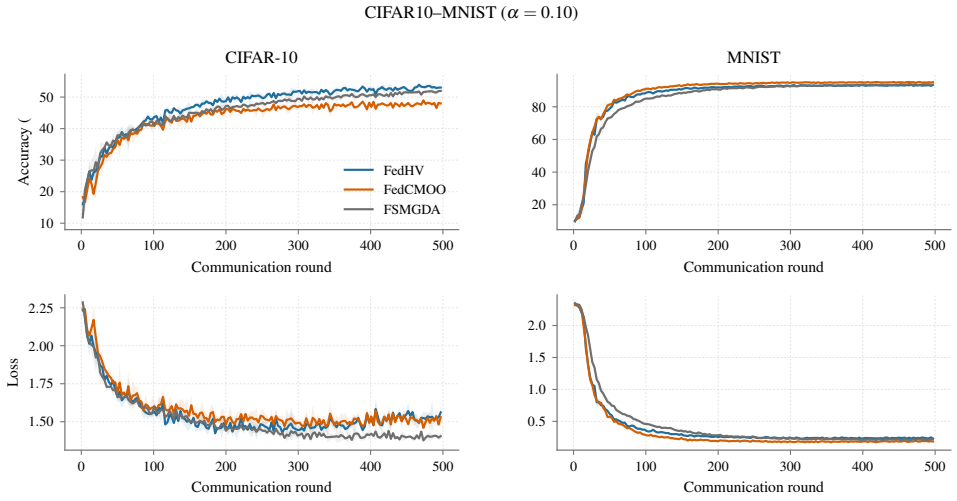


Figure 14: Per-task test accuracy and loss for CIFAR10–MNIST $\alpha = 0.10$, mean $\pm$ standard deviation over three seeds.

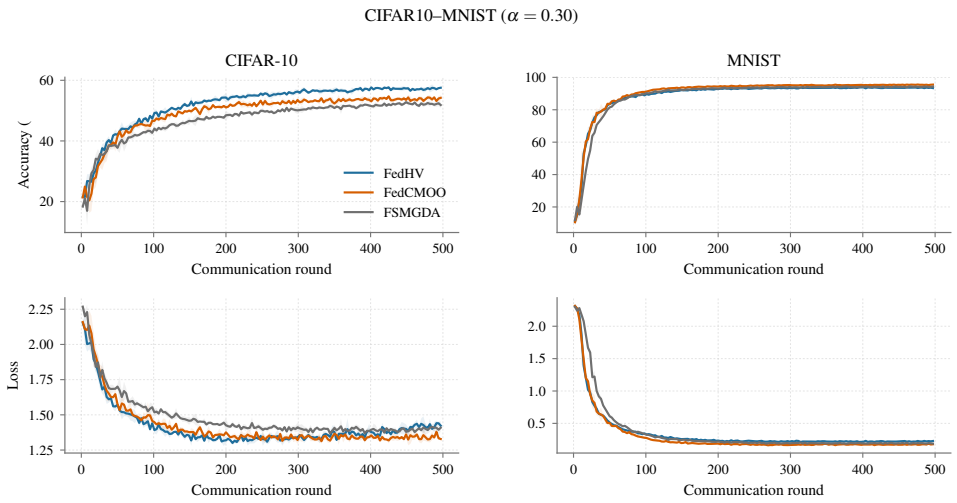


Figure 15: Per-task test accuracy and loss for CIFAR10–MNIST $\alpha = 0.30$, mean $\pm$ standard deviation over three seeds.

C.5 FedHV Ablations

Table 10: FedHV ablations on CIFAR10–MNIST $\alpha = 0.30$, mean $\pm$ standard deviation over three seeds. Bold marks the best value and underline the second-best value within each ablation sweep; loss uses four decimals to resolve rounding ties.

Ablation	Variant	MeanAcc ↑	MeanLoss ↓
Report budget	$B_{\text{rep}} = 1$	75.52 ± 0.19	0.8314 ± 0.0078
	$B_{\text{rep}} = 2$	$\underline{75.54 \pm 0.45}$	$\underline{0.8279 \pm 0.0151}$
	$B_{\text{rep}} = 5$	$\mathbf{75.66 \pm 0.18}$	$\mathbf{0.8274 \pm 0.0148}$
	$B_{\text{rep}} = \text{all}$	75.43 ± 0.43	0.8332 ± 0.0203
Participation	$M = 5/30$	73.96 ± 0.93	0.8639 ± 0.0346
	$M = 10/30$	75.41 ± 0.24	0.8396 ± 0.0101
	$M = 20/30$	$\underline{76.13 \pm 0.13}$	$\underline{0.8166 \pm 0.0050}$
	$M = 30/30$	$\mathbf{76.27 \pm 0.27}$	$\mathbf{0.8097 \pm 0.0112}$
Local steps	$K = 1$	61.98 ± 1.59	1.0522 ± 0.0373
	$K = 5$	74.82 ± 0.26	$\mathbf{0.7455 \pm 0.0055}$
	$K = 10$	$\mathbf{75.61 \pm 0.36}$	$\underline{0.8330 \pm 0.0175}$
	$K = 20$	$\underline{75.18 \pm 0.50}$	0.9778 ± 0.0252

The report-budget ablation demonstrates that FedHV remains effective with inexpensive objective estimation: one to five reporting minibatches closely track full-client reporting, supporting sparse objective reports in the tested regime. Increasing participation improves both accuracy and loss as broader client coverage reduces the sampling channels represented by $\chi_N(M)$ in the analysis. The local-step sweep exhibits the expected progress–drift trade-off: moderate K supplies enough local optimization to avoid undertraining, whereas larger K strengthens client drift after accuracy has saturated. Together, these results motivate the default reporting, participation, and local-computation budgets used in the main study.

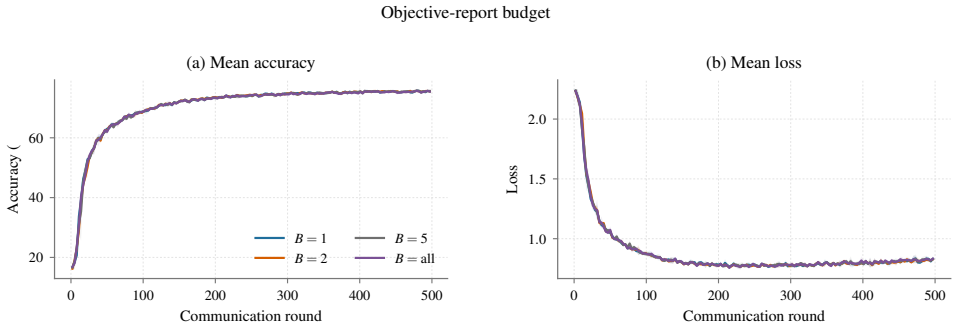


Figure 16: Objective-report-budget learning curves.

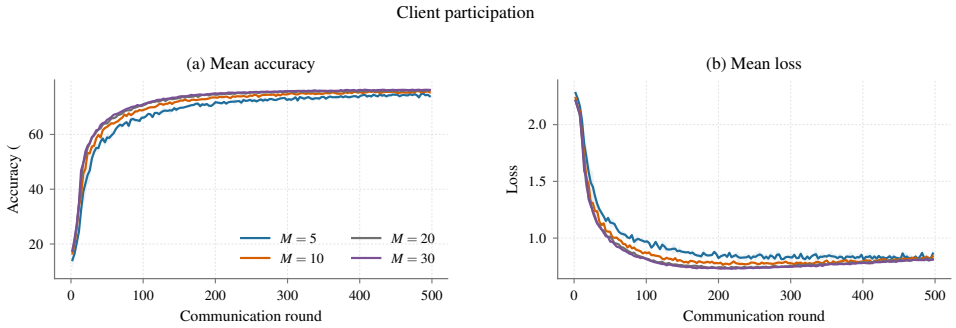


Figure 17: Client-participation learning curves.

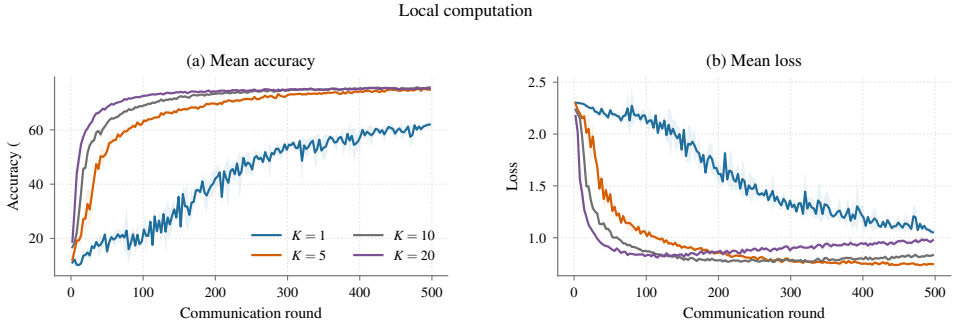


Figure 18: Local-step learning curves.

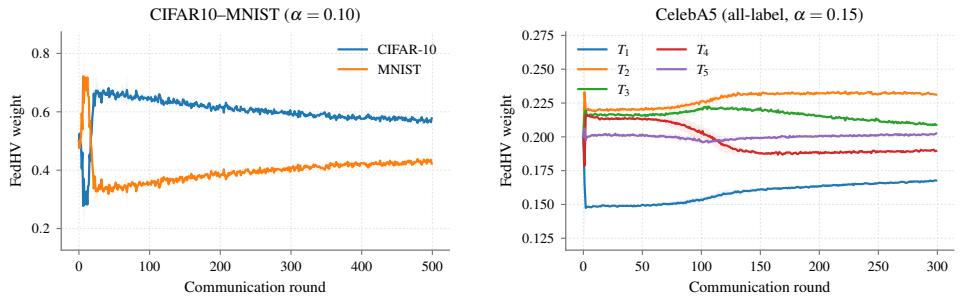


Figure 19: FedHV task weights over adaptive rounds, mean ± standard deviation across three seeds. On CIFAR10–MNIST $\alpha = 0.10$, FedHV assigns greater late-stage weight to the harder CIFAR-10 task; on CELEBA5, mass remains distributed over all five objectives rather than collapsing to one task.

C.6 FEDHV WEIGHT TRAJECTORIES

C.7 ADDITIONAL DIAGNOSTICS

The estimated-slack floor is inactive in the standard benchmark runs and in the B_{rep} and participation ablations. The $K = 1$ configuration is the exception: its diagnostic floor-activation rate is approximately 0.2%, with minimum stabilized slack 10^{-6}, and its task weights can become nearly degenerate. This is a diagnostic of the *estimated* slack stabilization and does not establish failure of the true safe-reference condition in Assumption A.4.

Table 11: Mean accuracy averaged over the final ten scheduled test evaluations, then aggregated over three seeds. This diagnostic reduces sensitivity to one final checkpoint; bold is best and underline is second best among FSMGDA, FedCMOO, and FedHV within each setting.

Setting	FSMGDA	FedCMOO	FedHV
MULTIMNIST $\alpha = 0.15$	88.60 ± 1.07	$\underline{89.66 \pm 0.08}$	$\mathbf{90.38 \pm 0.60}$
MNIST–FMNIST $\alpha = 0.30$	84.39 ± 1.01	$\mathbf{86.75 \pm 0.21}$	$\underline{86.13 \pm 0.17}$
CELEBA5 $\alpha = 0.15$	$\underline{86.75 \pm 0.11}$	86.50 ± 0.32	$\mathbf{88.05 \pm 0.11}$
CELEBA5 $\alpha = 0.30$	$\underline{85.55 \pm 0.18}$	84.43 ± 1.92	$\mathbf{87.87 \pm 0.06}$
CIFAR10–MNIST $\alpha = 0.10$	$\underline{72.61 \pm 0.30}$	71.44 ± 0.19	$\mathbf{73.13 \pm 0.22}$
CIFAR10–MNIST $\alpha = 0.30$	73.19 ± 0.35	$\underline{74.67 \pm 0.16}$	$\mathbf{75.41 \pm 0.29}$

Averaging the last ten scheduled evaluations preserves the main qualitative MeanAcc story: FedHV remains strongest in five of six settings against FSMGDA and FedCMOO, while FedCMOO remains ahead on MNIST–FMNIST. On CIFAR10–MNIST, the accuracy and cross-entropy trajectories provide complementary views of the learned trade-off, since the two metrics emphasize prediction quality and surrogate optimization, respectively.

D COMPLEXITY COMPARISON

Let d denote the model dimension, m the number of objectives, $M = |S_t|$ the number of participating clients, and K the number of local optimization steps.

FedHV overhead. FedHV uses one server-to-client downlink and one client-to-server uplink per communication round. Specifically, the server sends $(\boldsymbol{\theta}^t, \boldsymbol{w}^t)$ to each selected client in a single downlink, and client k returns $(\Delta\boldsymbol{\theta}_k^t, \widehat{\boldsymbol{F}}_k^t)$ in a single uplink. The objective estimates collected in round t are used only to prepare $\boldsymbol{w}^{t+1}$, so no intermediate loss-reporting round followed by a second weight-broadcast round is required.

A standard model exchange communicates d model scalars in each direction, whereas FedHV additionally communicates m task weights in the downlink and m objective values in the uplink. Hence

Table 12: Dominant per-round complexity of the compared methods. Here, d is the model dimension, m is the number of objectives, and M is the number of participating clients. Empirical transmission values reflect the CIFAR10–MNIST benchmark.

Method	**Client-side procedure**	**Per-client communication**	**Additional server-side procedure**
FSMGDA	Separate objective-specific local updates	$\Theta(md)$ (Empirical: ~ 205.5 MB)	Task-update geometry and multi-gradient optimization
Uniform	One uniformly weighted local trajectory	$\Theta(d)$ (Empirical: ~ 138.1 MB)	None beyond model aggregation
FedCMOO	Task-gradient estimation and weighted local training	$\Theta(d)$ (Empirical: ~ 205.5 MB)	Approximate Gram matrix construction and iterative weight updates
FedHV	Objective estimation and one weighted local trajectory	$\Theta(d+m) = \Theta(d)$ (Empirical: ~ 138.1 MB)	$O(Mm)$ objective aggregation and closed-form inverse-slack weighting

the total per-participating-client payload is $2d + 2m$ scalars per round, compared with $2d$ for the model exchange alone. Equivalently, the communication complexity is $\Theta(d + m) = \Theta(d)$ when $m \ll d$.

Including reference construction changes the end-to-end accounting. Let T_{cal} be the number of sequential calibration-round equivalents and M_{cal} its participating-client count. The total FedHV payload over calibration and adaptive training is $2d\,T_{\text{cal}}M_{\text{cal}} + (2d + 2m)TM$ scalars, before protocol-specific baseline auxiliaries. Here $T_{\text{cal}} = 20$ for each two-task benchmark family and $T_{\text{cal}} = 60$ for the robust CELEBA5 reference. Thus the calibration cost is included in the end-to-end budget even though it is excluded from adaptive-round learning curves.

Comparison with FSMGDA. FSMGDA generates and communicates objective-specific local updates, leading to communication that scales with md (Yang et al., 2023). FedHV instead maintains one local model trajectory and communicates only an m-dimensional objective vector in addition to the ordinary model update.

Comparison with FedCMOO. The communication-efficient version of FedCMOO uses randomized low-rank estimation to obtain model-order communication complexity (Askin et al., 2025). However, it still requires task-gradient information, an approximate task-gradient Gram matrix, and iterative server-side weight optimization. While FedCMOO successfully reduces communication from $\mathcal{O}(md)$ to $\mathcal{O}(d)$ via randomized-SVD, our empirical profiling reveals that uploading the compressed task-geometry matrix effectively doubles the actual uplink payload (~ 136.50 MB compared to the ~ 69.06 MB base model update). FedHV avoids gradient-geometry estimation entirely, requiring only 80 bytes of auxiliary scalar reports on the uplink, thereby achieving the theoretical $\mathcal{O}(d)$ bound with a practical multiplier of 1.

Empirical Efficiency & Synchronization. Protocol profiling further reveals that FedCMOO requires two distinct server-client synchronization stages per round to negotiate update geometry and weight projections. FedHV resolves weights entirely from the previous round's one-dimensional reports, operating seamlessly within a single standard synchronization stage. In high-latency federated environments, avoiding this secondary barrier is critical.

Table 13 and Figure 20 summarize the resulting end-to-end efficiency. Despite including the upfront penalty of its 20-round calibration phase, FedHV has lower measured wall-clock time and exhibits a tighter variance (± 5.69 seconds) compared to the erratic runtime fluctuations of FSMGDA (± 33.31 seconds). This confirms that calculating weights from scalar objective slacks mitigates the computational volatility inherent to projecting multi-dimensional task gradients.

Table 13: **Empirical Efficiency Analysis (CIFAR10–MNIST).** Metrics are averaged over three independent seeds for 300 adaptive communication rounds. Total values for FedHV strictly incorporate the one-time 20-round uniform-weight calibration phase ($\sim$ 117s, $\sim$ 2.76 GB) to ensure a fair end-to-end comparison.

Method	**Total Wall-Clock Time (s)**	**Downlink / Uplink per Client (MB)**
FSMGDA	2824.73 ± 33.31	69.06 / 136.50
FedCMOO	2521.24 ± 8.98	69.06 / 136.50
FedHV (Ours)	$\mathbf{1899.25 \pm 5.69}$	**69.06 / 69.06**

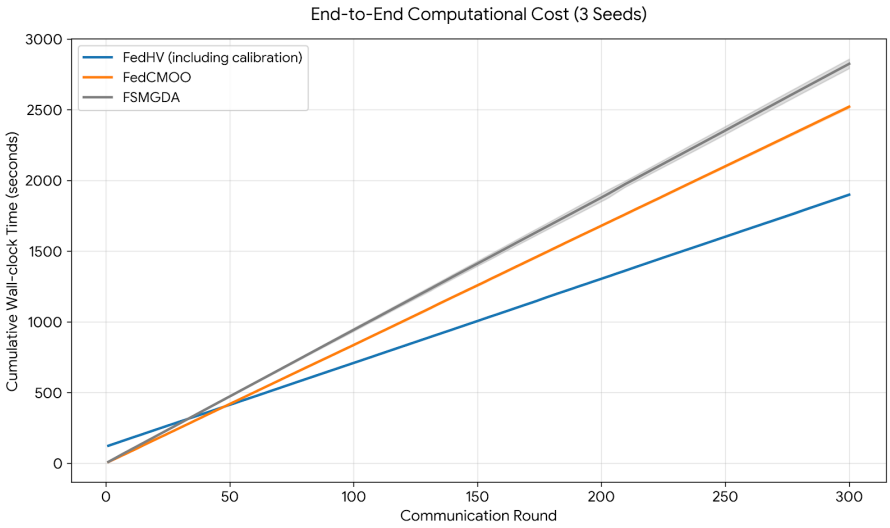


Figure 20: **End-to-End Computational Cost.** Cumulative wall-clock time over 300 communication rounds on CIFAR10–MNIST ($\alpha = 0.30$). The FedHV trajectory strictly includes the runtime of its 20-round uniform-weight calibration phase as an initial offset. FedHV avoids iterative server-side gradient optimization, resulting in a $\sim 24.7\%$ speedup over FedCMOO and $\sim 32.9\%$ speedup over FSMGDA with significantly lower runtime variance.

E COMPLETE FEDHV ALGORITHM

Algorithm 1 gives the full round-level pseudocode. The reference vector is calibrated and frozen before this adaptive phase; see Appendix B.8.

Algorithm 1 FedHV training procedure with one-round-stale inverse-slack weights

Input: rounds T; local steps K; client fraction p; report budget B_{rep}; stepsizes η_ℓ, η_g; fixed reference $\boldsymbol{r}$; estimated-slack floor ρ_{clip}.
Server initialization: $\boldsymbol{\theta}^0 \leftarrow \textsc{InitializeModel}()$ and $\boldsymbol{w}^0 \leftarrow \frac{1}{m}\mathbf{1}_m$.
for $t = 0, 1, \ldots, T-1$ **do**
Server selection: sample $S_t \subseteq [N]$ with $|S_t| = pN$ and broadcast $(\boldsymbol{\theta}^t, \boldsymbol{w}^t)$.
for all $k \in S_t$ **in parallel**
Client report: estimate $\widehat{\boldsymbol{F}}_k^t$ from B_{rep} minibatches at the pre-update model $\boldsymbol{\theta}^t$.
Client initialization: $\boldsymbol{\theta}_{k,0}^t \leftarrow \boldsymbol{\theta}^t$.
for $s = 0, 1, \ldots, K-1$ **do**
Draw a fresh minibatch $B_{k,s}^t$
$\boldsymbol{\theta}_{k,s+1}^t \leftarrow \boldsymbol{\theta}_{k,s}^t - \eta_\ell \nabla_\theta \left[\sum_{i=1}^{m} \mathrm{sg}\big(w_i^t\big)\ \ell_i(\boldsymbol{\theta}_{k,s}^t; B_{k,s}^t) \right]$
end for
$\Delta\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}_{k,K}^t - \boldsymbol{\theta}^t$
Client return: $(\Delta\boldsymbol{\theta}_k^t, \widehat{\boldsymbol{F}}_k^t)$.
end for
Server model aggregation: $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \eta_g |S_t|^{-1} \sum_{k \in S_t} \Delta\boldsymbol{\theta}_k^t$
Server report aggregation: $\widehat{\boldsymbol{F}}^t \leftarrow |S_t|^{-1} \sum_{k \in S_t} \widehat{\boldsymbol{F}}_k^t$
Server weight update: $\widehat{s}_i^t \leftarrow \max\{r_i - \widehat{f}_i^t, \rho_{\mathrm{clip}}\}$ for every $i \in [m]$.
$w_i^{t+1} \leftarrow \frac{(\widehat{s}_i^t)^{-1}}{\sum_{j=1}^{m} (\widehat{s}_j^t)^{-1}}$ for every $i \in [m]$.
end for
Output: global model $\boldsymbol{\theta}^T$.